### TASK DETAILS

<table>
<thead>
<tr>
<th>Task Name</th>
<th>Submit Revised Final Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due Date</td>
<td>24/05/2017</td>
</tr>
<tr>
<td>Submitted Date</td>
<td></td>
</tr>
</tbody>
</table>

### REPORT DETAILS

<table>
<thead>
<tr>
<th>Programme</th>
<th>Project Reference Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Editorial Reports Only</td>
<td>RP-PG-0707-10101</td>
</tr>
<tr>
<td><strong>Title of Report</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Lead Author</strong></td>
<td></td>
</tr>
</tbody>
</table>
### RESPONSE TO FR COMMENTS

<table>
<thead>
<tr>
<th>Staff Comments to Author</th>
<th>Author Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Editor Comments</td>
<td>Response to Editor and Reviewer comments are submitted in the &quot;Uploads&quot; section.</td>
</tr>
<tr>
<td>Editors’ Comments</td>
<td>Revised final report along with additional information has now been uploaded. Please see response to Editors comments in provided information.</td>
</tr>
<tr>
<td>Upload Type</td>
<td>File Name</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Revised Final Report</td>
<td>NIHR Final Report Evaluation of Biomarkers RESUBMITTION2 V240517.docx</td>
</tr>
</tbody>
</table>
A Multi-Centre Programme into Methods for the Evaluation of Biomarkers Suitable for Use in Patients with Kidney and Liver Diseases and a Randomised Controlled Trial for Their Use

Version 24 May 2017

Peter J Selby1,2,3*, Rosamonde E Banks1,3, Walter Gregory4, Jenny Hewison5, William Rosenberg6, Douglas G Altman7, Jonathan J Deeks8, Chris McCabe9, Julie Parkes10, Catharine Sturgeon11, Douglas Thompson2, Maureen Twiddy5, Janine Bestall5, Joan Bedlington12, Tilly Hale12, Jacqueline Dinnes8, Marc Jones4, Andrew Lewington2, Michael P Messenger2, Vicky Napp4, Alice Sitch8, Sudeep Tanwar6, Naveen S Vasudev1,2, Paul Baxter13, Sue Bell4, David A Cairns1, Nicola Calder2, Neil Corrigan4, Francesco Del Galdo14, Peter Heudtlass4, Nick Hornigold1 Claire Hulme15, Michelle Hutchinson1, Carys Lippiatt16, Tobias Livingstone2, Roberta Longo5, Matthew Potton4, Stephanie Roberts1, Sheryl Sin1, Sebastian Trainor1, Matthew Welberry Smith1,2, James Neuberger17, Douglas Thorburn18, Paul Richardson19, John Christie20, Neil Sheerin21, William McKane22, Paul Gibbs23, Anusha Edwards24, Naeem Soomro21, Adebamji Adeyoju25, Grant D Stewart26,27, David Hrouda28

1Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, UK.
2Leeds Teaching Hospitals NHS Trust, Leeds, UK.
3Clinical and Biomedical Proteomics Group, Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, UK.
4Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, UK.
5Leeds Institute of Health Sciences, University of Leeds, Leeds, UK.
6Institute for Liver and Digestive Health, Division of Medicine, University College London, London, UK.
7Centre for Statistics in Medicine, University of Oxford, Oxford, UK.
Institute of Applied Health Research, University of Birmingham, Birmingham, UK.

Department of Emergency Medicine, University of Alberta Hospital, Edmonton, Alberta, Canada.

University of Southampton, Southampton, UK.

Royal Infirmary of Edinburgh, Edinburgh, UK.

LIVERNORTH Liver Patient Support

Leeds Institute of Cardiovascular & Metabolic Medicine, University of Leeds, Leeds, UK.

Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, UK.

Academic Unit of Health Economics, Leeds Institute of Health Sciences, University of Leeds, Leeds, UK.

Department of Specialist Laboratory Medicine, Block 46, SJUH, Leeds Teaching Hospitals NHS Trust, Leeds, UK.

University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK.

Royal Free London NHS Foundation Trust, London, UK.

The Royal Liverpool and Broadgreen University Hospitals NHS Trust, Liverpool, UK.

Royal Devon and Exeter NHS Foundation Trust, Exeter, UK.

Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK.

Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK.

Portsmouth Hospitals NHS Trust, Hampshire, UK.

North Bristol NHS Trust, Bristol, UK.

Stockport NHS Foundation Trust, Stockport, UK.

NHS Lothian, Edinburgh, UK.

Academic Urology Group, University of Cambridge

Charing Cross Hospital, Imperial College Healthcare NHS Trust, London, UK.

*Corresponding author

Competing interests:

- Prof. Peter Selby reports International Patent Application No PCT/GB2014/050768; Biomarker; University of Leeds issued.

- Prof. Banks reports grants from various non-commercial funding sources, during the conduct of the study; in addition, Dr. Banks has a patent International Patent Application No PCT/GB2014/050768; Biomarker; University of Leeds issued.
• Prof. Jenny Hewison is a panel member of CTU Board
• Dr. Grant Stewart has a patent 1408091.5 pending to MRC HGU.
• James Neuberger is a consultant to Astellas and has received speaker support funding from Astellas and Novartis
• Prof. Claire Hulme has received grants from NIHR, during the conduct of the study and panel member of HTA Commissioning Board
• Dr. Andrew Lewington has received personal fees from Fresenius Medical Care, Baxter, AM Pharma, Bioporto and GE Health Care, outside the submitted work; In addition, Dr. Lewington has a patent ACY-1 Biomarker pending.
• Dr. Naveen Vasudev has received personal fees from Bristol Myers Squib and Glaxo Smith Kline, outside the submitted work
• Prof. William Rosenberg William Rosenberg was an inventor of the ELF test when he was an employee of the University of Southampton. His rights were transferred to Siemens Healthcare Diagnostics by the University of Southampton. He does not receive any payment in relation to sales of the test by the manufacturer Siemens Helathcare Diagnostics. He has received grant support and speaker fees from Siemens Healthcare. He is a director of iQur Limited, a company that provides ELF testing. In the context of the NIHR funded study, all ELF testing provided by iQur was performed on a not-for-profit, cost-recovery basis. William Rosenberg is married to Julie Parkes who is a co-investigator on the programme.
• Dr. Parkes reports receiving support from speaker bureau Siemens Healthcare Diagnostic, outside the submitted work; and Julie Parkes is married to William Rosenberg (Co-applicant).
• Prof. Jon Deeks is a panel member of HTA Commissioning Board

Keywords
Biomarkers; liver disease; kidney disease; prostate specific antigen; monitoring trials; simulation of biomarker studies; ELF test; ELUCIDATE; renal cancer; renal transplantation; diagnosis of cirrhosis.

Word count for main body of report 117,749 (not including summaries, tables, figures or acknowledgements)
Abstract

Background

Protein biomarkers with associations with the activity and outcome of diseases are being identified by modern proteomic technologies. They may be simple, accessible, cheap and safe tests which can inform diagnosis, prognosis, treatment selection, monitoring of disease activity and therapy and may substitute for complex, invasive and expensive tests. However, their potential is not yet being realised.

Design and Methods

Our NIHR Applied Programme included three workstreams to create a framework for research: WS1) METHODOLOGY. To define current practice and explore methodology innovations for biomarkers to monitor disease. WS2) CLINICAL TRANSLATION. To create a framework of research practice, high quality samples and related clinical data in order to evaluate the validity and clinical utility of protein biomarkers. WS3) ELUCIDATE RCT. An exemplar randomised trial (RCT) of an established test, ADVIA Centaur® Enhanced Liver Fibrosis (ELF) test (comprised of a panel of three markers - serum hyaluronic acid (HA), amino-terminal propeptide of type III procollagen (PIIINP) and tissue inhibitor of metalloproteinase 1 (TIMP-1)), for liver cirrhosis to determine its impact on diagnostic timing and the management of cirrhosis, and on the process of care and improve outcomes.

Results

The methodology workstream evaluated the quality of recommendations for using Prostate Specific Antigen (PSA) to monitor patients, systematically reviewed RCT of monitoring strategies, reviewed the monitoring biomarker literature and how monitoring can impact on outcomes. We conducted simulation studies to evaluate monitoring and improve merits of healthcare. The literature on biomarkers monitoring diseases is modest and robust conclusions are infrequent. We recommended improvements in research practice. Patients strongly endorsed the need for robust and conclusive research in this area.
The clinical translation workstream focused on analytic and clinical validity. Cohorts were established for renal cancer (RCC) and renal transplantation (RT) with samples and patient data from multiple Centres as a rapid access resource to evaluate the validity of biomarkers. Candidate biomarkers for RCC and RT were identified from the literature, their quality evaluated and selected biomarkers were prioritised. The duration of follow up was a limitation but they identified biomarkers which may be taken forward for clinical utility.

The ELUCIDATE RCT registered 1303 patients and randomised 878 patients out of a target of 1000. The trial started late and recruited, initially, slowly but ultimately recruited with good statistical power to answer the key questions. ELF monitoring altered the patient process of care and may show benefits from the early introduction of interventions with further follow up. ELUCIDATE was an “exemplar” trial which has demonstrated the challenges of evaluating biomarker strategies in “end-to-end” RCTs and will inform future study designs.

Conclusions

The limitations in the programme were principally that during the collection and curation of the cohorts of patients with RCC and RT, the pace of discovery of the new biomarkers in commercial and non-commercial research was slower than anticipated, so conclusive evaluations using the cohorts are few but access to the cohorts will be sustained for future new biomarkers. The ELUCIDATE trial was slow to start and recruit – with a late surge of recruitment so final conclusions about the impact on key outcomes await further follow-up.

The findings in the three workstreams were used to synthesise a strategy and framework for future biomarker evaluations incorporating innovations in study design, health economics and health informatics.

Study registrations

- Enhanced Liver fibrosis (ELF) test to Uncover Cirrhosis as an Indication for Diagnosis and Action for Treatable Events ISRCTN74815110
- Evaluation of Biomarkers for Prognosis and Monitoring in Patients with Renal Cell Carcinoma UKCRN ID 9954
- Evaluation of biomarkers for post-renal transplant complications UKCRN ID 11930
Funding details

This research was funded by the National Institute for Health Research Programme Grants for Applied Research Programme which provided comprehensive funding and two programme extensions.

Word count: 610
Table of contents

Abstract .......................................................................................................................... 5
Background ..................................................................................................................... 5
Design and Methods ....................................................................................................... 5
Results ............................................................................................................................. 5
Conclusions ..................................................................................................................... 6
Study registrations .......................................................................................................... 6
Funding details ................................................................................................................. 7
Table of contents ........................................................................................................... 8
List of figures .................................................................................................................. 25
List of tables ................................................................................................................... 31
List of boxes .................................................................................................................... 42
Glossary .......................................................................................................................... 44
Scientific summary .......................................................................................................... 48
Plain English summary ................................................................................................... 55
Chapter 1 - Introducing New Biomarkers for Renal and Hepatic Diseases into Healthcare Systems ......................................................................................................................... 56
  Background .................................................................................................................... 57
  Selected Diseases .......................................................................................................... 58
    Chronic Liver Disease (CLD) ...................................................................................... 58
    Acute Renal Transplant ............................................................................................. 59
    Renal Cell Carcinoma (RCC) ..................................................................................... 59
  Protein Biomarkers ....................................................................................................... 59
  Monitoring Studies ....................................................................................................... 60
  Choice of the ELF test for the ELUCIDATE RCT ....................................................... 61
Chapter 2 - Introduction to the Methodological Workstream (WS1) ............................. 65
Chapter 3 - How is evidence being used to make recommendations about monitoring: the example of prostate specific antigen (PSA) ................................................................................. 70
Chapter 4 - Has the randomised controlled trial design been successfully used to evaluate strategies for monitoring disease progression or recurrence? An assessment of experience to date

Introduction .................................................................................................................. 93

Methods ....................................................................................................................... 94

Literature search .......................................................................................................... 94

Inclusion criteria ........................................................................................................... 94

Data extraction and analysis ....................................................................................... 95

Results ......................................................................................................................... 95

General description of included trials ......................................................................... 97

Description of monitoring strategies .......................................................................... 99

What was the aim of the monitoring evaluation and what change in patient care was implemented? ............................................................................................................. 102

Primary outcomes and results .................................................................................... 105

Discussion .................................................................................................................. 108

Chapter 5 - A review of monitoring-related methodology literature ......................... 114

Introduction ............................................................................................................... 115

Methods ...................................................................................................................... 115
Chapter 6 - How can monitoring impact on patient outcomes? 

Methods

Monitoring versus screening or diagnosis

Potential benefits and harms from monitoring

When is randomised evidence needed?
Chapter 7 - Simulating monitoring data and evaluating monitoring strategies ..........157

Introduction .................................................................................................................. 158

Aims and objectives ........................................................................................................159

Methods .......................................................................................................................... 159

Simulation of true disease progression ........................................................................ 161

Fibrosis progression—random slope .............................................................................. 161
Fibrosis stage at entry—random intercept ..................................................................... 162
ELF score link to fibrosis stage ..................................................................................... 162
ELF progression between fibrosis stages ....................................................................... 162
ELF progression—random slope .................................................................................. 162
ELF value at entry—random intercept ......................................................................... 163
Random slope and random intercept model in terms of ELF ..................................... 163
Simulation of observed values ..................................................................................... 165

Error ............................................................................................................................... 165
Entry criteria .................................................................................................................. 165
Data sources .................................................................................................................. 165

Fibrosis progression rate ............................................................................................. 165
Fibrosis stage at entry ................................................................................................... 167
Measurement error ........................................................................................................ 167
ELF score link to fibrosis stage ..................................................................................... 167
Implementation of a monitoring strategy ..................................................................... 168

Monitoring data required ............................................................................................ 168
Monitoring strategies .................................................................................................... 168
Simple decision rule (strategy A) .................................................................................. 168
Retesting (strategy B) ..................................................................................................... 169
Frequency of monitoring (strategy C) ......................................................................... 169
Alternative decision rules ............................................................................................. 169
Evaluation of a monitoring strategy ............................................................................. 170

Comparison of observed results with the underlying disease state ............................. 170
Measuring the performance of a monitoring strategy ................................................... 172
Evaluation of strategies ................................................................................................. 172
Results ............................................................................................................................ 173

Reference monitoring strategy (strategy A) ................................................................. 173
Chapter 8 - Methodological considerations in the optimization of monitoring biomarkers to meet value-based market access hurdles .................213

Background ..................................................................................214

Cost effectiveness in personalized medicine technologies ............216

A framework for characterizing Personalised Medicine Technologies ....222

Extending Phelps and Mushlin for monitoring tests ....................224

Some observations on the estimation of the value of information for monitoring tests ..........................................................234

Chapter 9 - An overview and patient perspective on biomarker guided research ......238

Introduction..................................................................................239
Stage 1: Understanding outputs from the methodological workstream and agreeing key areas for discussion with patient and public representatives. ................................................................. 240

Monitoring tests ......................................................................................................................... 242
A case study ................................................................................................................................. 244
Trial evidence ............................................................................................................................... 245
The methodology literature ......................................................................................................... 245
Patient outcomes and monitoring ............................................................................................... 246
The first modelling study ............................................................................................................. 247
The second modelling study ......................................................................................................... 248
The third modelling study ........................................................................................................... 250
Implications for the design of clinical utility studies ................................................................. 253

Stage 2: Obtaining patient and public perspectives ..................................................................... 256

Establishing a shared understanding ......................................................................................... 256
Initial presentation to PPI representatives .................................................................................. 256
Liver North Workshop ................................................................................................................ 257
Patient perspectives and the “E” in ACCE ................................................................................ 259

Chapter 10 - Biomarker Pipelines – Ensuring Clinical Translation Using Renal Cancer and Kidney Transplantation as Exemplars ................................................................. 261

Main Aims of the Clinical Translation Workstream (WS2) .......................................................... 262
Critical Elements in the Biomarker Discovery and Translation Process ..................................... 263
General Concepts ........................................................................................................................ 263
Specific Lab-based/Clinical Biochemistry Aspects ....................................................................... 267
Establishing the Pipeline in Renal Cancer and Kidney Transplantation .................................... 270
Renal Cancer ............................................................................................................................... 270
Clinical Context ......................................................................................................................... 270
Determining prognosis in RCC ................................................................................................... 273
Kidney Transplantation ............................................................................................................. 274
Clinical Context ......................................................................................................................... 275
Delayed Graft Function (DGF) ................................................................................................... 277
Overview of the Work Undertaken in the Clinical Translation Workstream (WS2) to Develop and Use These Pipelines ................................................................. 280

Sample Banks and Clinical Data ................................................................. 280

Biomarker Prioritisation, Assay Validation and Evaluation of Clinical Utility ......... 283

Additional Deliverables ................................................................................. 284

Chapter 11 - Establishment of multicentre prospective observational cohorts with sample banks for biomarker validation ................................................. 285

Disease areas ............................................................................................... 286

Renal Cancer ............................................................................................... 286

Objectives and Endpoints ............................................................................. 286
Eligibility criteria .......................................................................................... 287
Study Design .................................................................................................. 287
Statistical considerations ............................................................................. 292

Renal Transplant .......................................................................................... 293

Objectives and Endpoints ............................................................................. 293
Eligibility Criteria .......................................................................................... 293
Study Design .................................................................................................. 293
Statistical Considerations ............................................................................. 296

Liver Disease ............................................................................................... 297

Study Management ..................................................................................... 297

Investigator authorisation ............................................................................ 297
Recruitment .................................................................................................... 298
Patient Consent ............................................................................................. 298
Patient registration ........................................................................................ 300
Data collection and storage ......................................................................... 300
Sample collection and storage ..................................................................... 300
Monitoring and quality assurance ................................................................. 301

Study Management Results and Discussion of Issues ................................. 301

Consent ........................................................................................................ 302
Sample Processing ........................................................................................ 303
Clinical Data .................................................................................................. 311

Recruitment to time and target .................................................................... 313
Study Governance Aspects ................................................................. 318
Ethical considerations ........................................................................ 318
Establishment of a Research Tissue Bank ......................................... 318
Local and National NHS approvals .................................................. 319
Governance Results and Discussion of Issues .................................. 320
NIHR CRN Portfolio Adoption ............................................................ 320
NHS Permissions .............................................................................. 321
Human Tissue Act (HTA) Relevant Material ....................................... 321
NHS Service Support Costs ............................................................... 321
Local and National NHS Approvals .................................................. 322
Summary of final resource ................................................................ 328
Liver Disease cohort summary statistics ................................ .......... 328
Renal Cell Carcinoma and healthy control cohort summary statistics .... 328
Renal Transplant cohort summary statistics ....................................... 333
Learning and recommendations for future bioresources ................. 335
Concluding remarks ......................................................................... 338
Chapter 12 - Review and prioritisation of circulating biomarkers in renal cancer and renal transplantation ......................................... 339
Renal Cancer ..................................................................................... 340
Literature Search Strategy ................................................................. 340
Routinely Measured Analytes ............................................................ 342
  C-Reactive Protein (CRP) ......................................................... 345
  Serum amyloid A (SAA) .......................................................... 347
  Ferritin...................................................................................... 348
  Erythropoietin (EPO) .............................................................. 348
  Vascular endothelial growth factor (VEGF) ................................ 348
  Carbonic anhydrase IX (CAIX) ................................................. 350
  Matrix metalloproteinase-7 (MMP-7) ......................................... 351
  Osteopontin (OPN) ................................................................. 351
  Immunosuppressive acidic protein (IAP) ..................................... 352
  Tumour M2 pyruvate kinase (TuM2-PK) and Thymidine kinase (TK1) 353
  Soluble II-2 receptor (sIL-2R) .................................................. 353
  Basic fibroblast growth factor (bFGF) ......................................... 353
Chapter 13 - Exploring Technical Aspects of Biomarker Assays – Verification, Validation and Pre-Analytical Variables

Conclusions and Prioritisation ........................................................................................................ 354

Renal Transplantation .................................................................................................................. 355

Literature Search Strategy ..................................................................................................... 355

Review of serum and plasma biomarkers of DGF following renal transplantation – diagnosis and prognostic utility for long-term outcome ........................................ 357

Creatinine .................................................................................................................................. 357
Cystatin C .................................................................................................................................. 357
Neutrophil Gelatinase Associated Lipocalin (NGAL) ................................................................. 358
Aminoacylase-1 ......................................................................................................................... 360
Other Promising biomarkers ....................................................................................................... 360

Complement .............................................................................................................................. 360
C-Terminal Agrin Fragment (CAF) ............................................................................................. 361
Fms-Like Tyrosine Kinase (Flt-1) ............................................................................................... 361
IgA Antibodies to β2-Glycoprotein I ......................................................................................... 361
Pregnancy – associated plasma protein (PAPP-A) .................................................................. 361
Interleukin-16 (IL-16) ................................................................................................................ 362
Interleukin-18 (IL-18) ................................................................................................................ 362
Leptin .......................................................................................................................................... 362
Resistin ......................................................................................................................................... 363
ST6Gal ......................................................................................................................................... 363
Stem Cell Factor (SCF) ................................................................................................................. 363
Hydroxyeicosatetraenoic Acids ................................................................................................. 364
Malondialdehyde ......................................................................................................................... 364
Neutrophil-Lymphocyte Ratio .................................................................................................... 365
Regulatory T Cells ..................................................................................................................... 365

Summary .................................................................................................................................... 366

Appendix A - Analytical Variables and Criteria for the Development of Verification Protocols

Appraisal of Assay Performance .............................................................................................. 369

General Concepts ..................................................................................................................... 369

Guidelines ................................................................................................................................... 370

Pre-Analytical Errors and Variation ........................................................................................ 372

Development of Verification Protocols for Commercially Available Immunoassays .......... 376

Key Technical Validation (Verification) Elements and Criteria ............................................ 376

Standard Assessments ............................................................................................................... 377

Additional Assessments ............................................................................................................ 379

Specific Biomarker Technical (Assay and Pre-Analytical) Studies Undertaken Within the Programme...................................................................................................................... 379

16
### Osteopontin

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>380</td>
</tr>
<tr>
<td>Methodology</td>
<td>380</td>
</tr>
<tr>
<td>Results</td>
<td>381</td>
</tr>
<tr>
<td>Serum versus EDTA plasma</td>
<td>381</td>
</tr>
<tr>
<td>Imprecision</td>
<td>382</td>
</tr>
<tr>
<td>Parallelism</td>
<td>382</td>
</tr>
<tr>
<td>Recovery, Interference and Hook Effect</td>
<td>382</td>
</tr>
<tr>
<td>LLOQ</td>
<td>383</td>
</tr>
<tr>
<td>Sample Stability</td>
<td>383</td>
</tr>
<tr>
<td>Discussion</td>
<td>384</td>
</tr>
</tbody>
</table>

### Neutrophil Gelatinase-Associated Lipocalin (NGAL)

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>387</td>
</tr>
<tr>
<td>Methodology</td>
<td>387</td>
</tr>
<tr>
<td>Results</td>
<td>388</td>
</tr>
<tr>
<td>Standardisation</td>
<td>388</td>
</tr>
<tr>
<td>Imprecision</td>
<td>388</td>
</tr>
<tr>
<td>Parallelism</td>
<td>388</td>
</tr>
<tr>
<td>Recovery and Specificity</td>
<td>389</td>
</tr>
<tr>
<td>Measurement of CAIX and effects of metal ions</td>
<td>390</td>
</tr>
<tr>
<td>Discussion</td>
<td>393</td>
</tr>
</tbody>
</table>

### Clinical Validation

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>395</td>
</tr>
<tr>
<td>Methodology</td>
<td>396</td>
</tr>
<tr>
<td>Results</td>
<td>397</td>
</tr>
<tr>
<td>Imprecision</td>
<td>397</td>
</tr>
<tr>
<td>Parallelism</td>
<td>397</td>
</tr>
<tr>
<td>Recovery</td>
<td>397</td>
</tr>
<tr>
<td>Selectivity</td>
<td>401</td>
</tr>
<tr>
<td>Haemoglobin Interference</td>
<td>401</td>
</tr>
<tr>
<td>Hook Effect</td>
<td>401</td>
</tr>
<tr>
<td>Inter-assay NGAL comparison</td>
<td>401</td>
</tr>
<tr>
<td>Discussion</td>
<td>404</td>
</tr>
</tbody>
</table>

### Vascular Endothelial Growth Factor (VEGF) - Relative value of serum or plasma and QC aspects

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>407</td>
</tr>
<tr>
<td>Methodology</td>
<td>407</td>
</tr>
<tr>
<td>Results</td>
<td>407</td>
</tr>
<tr>
<td>Discussion</td>
<td>415</td>
</tr>
</tbody>
</table>

## Chapter 14 - Circulating Prognostic Biomarkers in Renal Cancer – Clinical Validation

### Study of Promising Candidates

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>423</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>424</td>
</tr>
</tbody>
</table>
Methods .......................................................................................................................... 425
Patient population ........................................................................................................... 425
VEGF, OPN and CAIX measurement .............................................................................. 425
Clinico-pathological variables ....................................................................................... 426
Statistical methods ........................................................................................................ 426
Sample size ..................................................................................................................... 429
Results .............................................................................................................................. 429
Patient and tumour characteristics ................................................................................. 430
Biomarker associations / correlations ............................................................................ 430
Univariate analysis of biomarkers and time-to-event endpoints ..................................... 434
Metastasis-Free Survival ............................................................................................... 434
Overall Survival ........................................................................................................... 437
Multivariable Analysis of MFS ...................................................................................... 440
Discussion ......................................................................................................................... 441

Chapter 15 - Conclusions of the Clinical Translational Workstream ........................................... 447

Chapter 16 - Introduction to the ELUCIDATE Trial (including scientific background and explanation of rationale) ..................................................................................................................... 452

Chapter 17 - Verification of ADVIA Centaur© ELF Test analytical performance ............ 456

Introduction to analytical performance evaluation, precision and bias ......................... 457
Precision and Imprecision ............................................................................................... 459
Trueness and Bias .......................................................................................................... 460
Analytical performance goals incorporating biological variation .................................... 461
Single Site Evaluation of the ADVIA Centaur© ELF Test .............................................. 461
Study Aims - Single site evaluation ............................................................................... 462
Study Methods - Single site evaluation ......................................................................... 462
Sample collection .......................................................................................................... 462
Sample Preparation ........................................................................................................ 463
Measurement of Pooled Serum Samples and Reference Controls ............................... 463
Data Analysis and Verification ...................................................................................... 463
Results - Single site evaluation ................................................................................... 464
Assessment of imprecision ............................................................................................ 464
Assessment of bias ......................................................................................................... 465
Determination of analytical performance goals (APG) ................................................... 465
Discussion - Single site evaluation ............................................................................... 472
Multi-Site Evaluation of the ADVIA Centaur© ELF Test

Study Aim – Multi-site evaluation

Study Design Multi-site evaluation

Test materials and distribution:

Laboratory procedures

Results - Multi-site evaluation

Discussion - Multi-site evaluation

Study conclusions

Chapter 18 - Design and Set up of the ELUCIDATE Trial

Study design

Ethical approval and research governance

Participants

Inclusion criteria for registration
Exclusion criteria for registration
Screening and consent procedure
Registration and baseline ELF test
Eligibility for randomisation
Randomisation
Quality assurance of ELF test

Treatment group allocation

Screening for cirrhosis with standard clinical monitoring
Screening for cirrhosis with standard clinical monitoring plus ELF test

Data collection and management

Baseline registration assessment
Baseline randomisation assessment

Follow-up

Follow-up and management of patients diagnosed with cirrhosis

Endpoints

Primary endpoint (according to protocol v7.0, 30 January 2013)
Secondary Endpoints (according to protocol v7.0, 30 January 2013)
Health-related quality of life
Long term follow-up
Statistical analysis
Chapter 19 - Value of Modelling Development, Modification and Extension of the ELUCIDATE Trial

Background ........................................................................................................ 495
Modelling methodology .................................................................................. 499
Conclusions ....................................................................................................... 503

Chapter 20 - Recruitment and Delivery of the ELUCIDATE Trial

Trial Organisation .............................................................................................. 506
Registration and Randomisation ..................................................................... 507
Clinical Queries ................................................................................................ 507
Project and Trial Management ....................................................................... 507
Trial Steering Committee ................................................................................ 507
Leeds NIHR Biomarker Bank .......................................................................... 508
ELF testing and the provision of ELF and ELF/NIHR Biomarker Bank kits ...... 508
Accrual .............................................................................................................. 508

Chapter 21 - Preliminary Analysis of the ELUCIDATE Trial

Baseline Characteristics .................................................................................... 513
Protocol Violators ............................................................................................. 517
ELF test below 8.4 at follow-up ...................................................................... 517
More than 12 weeks between registration and randomisation ....................... 518
Visit Compliance ............................................................................................... 520
Compliance with pre-cirrhotic follow-up visits ............................................. 520
Timing of pre-cirrhotic follow-up visits ......................................................... 520
Compliance with post-cirrhotic follow-up visits ............................................. 522
Timing of post-cirrhotic follow-up visits ......................................................... 522
ELF Test Compliance ....................................................................................... 523
ELF Test compliance in the ELF arm ............................................................... 523
ELF Test compliance in the standard of care arm ........................................524
Compliance with cirrhosis management .......................................................525
Compliance with timing of Alpha-Fetoprotein (AFP) tests, scans (USS, CT or MRI) and Endoscopies (OGD) was assessed for patients who were diagnosed with cirrhosis. ........525
Alpha-Fetoprotein (AFP) tests .......................................................................525
  AFP compliance .........................................................................................525
  AFP timings ...............................................................................................525
Scans (USS, CT or MRI) .................................................................................526
  Scan compliance .........................................................................................527
  Scan timings ...............................................................................................527
Endoscopies (OGD) .......................................................................................528
  OGD compliance ........................................................................................529
Withdrawals ..................................................................................................529
Disease Progression to Cirrhosis .................................................................530
Process of Care Outcomes ............................................................................533
  Frequency of Biopsies ...............................................................................533
  Frequency of Alpha-Fetoprotein (AFP) tests ............................................534
  Frequency of Ultrasound Scans (USS) ......................................................537
  Frequency of Oesophago-Gastro-Duodenoscopies (OGD) ......................540
  Frequency of Beta-Blocker/Band litigation treatment ................................542
  Frequency of treatment to normalise LFTs ...............................................542
Health Economic consequences ..................................................................543
  Descriptive analysis ..................................................................................545
    Cost associated with process of care outcomes ......................................545
    Severe complications and deaths ............................................................546
Discussion .....................................................................................................546

Chapter 22 - Workstream 3 – Next Steps and Preliminary Conclusions ..........551
  Summary and Discussion of the Results of ELUCIDATE to date ...............552
Chapter 23

Chapter 23 - Patient and public perspectives

Part one: Patients Experiences in the ELUCIDATE trial: A Qualitative study about patient experiences of taking part in a trial to test biological fluid biomarkers for Liver Disease.
Interviews ........................................................................................................... 571
Analysis ........................................................................................................... 571
Results ............................................................................................................. 572
Patient Characteristics .................................................................................... 572
Key themes ....................................................................................................... 574
Patients’ experience and understanding of the ELUCIDATE trial .................. 574
Patients’ experience and understanding of the Enhanced Liver Fibrosis (ELF) test 577
Support and information received as part of the ELUCIDATE trial ................. 578
Patients’ perspective of being part in the ELUCIDATE study: discussion points 579
Limitations ....................................................................................................... 580
Conclusion ....................................................................................................... 580
Part two: drawing threads together .............................................................. 581
Guidelines to improve research quality ......................................................... 584
Gaps in the evidence ....................................................................................... 585
Choosing wisely ............................................................................................... 586
Personalised medicine, innovation and the funding of research ................. 587

Chapter 24 - Programme Conclusions and the Framework for Biomarker Evaluation ........................................................................................................... 589
General conclusions ........................................................................................ 590
The NIHR Diagnostic Evidence Cooperative at Leeds .................................... 595
IVD evaluation methodology research ............................................................ 596
Qualify Analytical Validity ............................................................................. 598
Qualify Clinical Validity ................................................................................... 598
Evaluation of Clinical Utility and Cost Effectiveness .................................. 599
The Interactions between Workstreams ......................................................... 599
Future Methodology Research ...................................................................... 600
Limitations ...................................................................................................... 601
Final Comments .............................................................................................. 602
Acknowledgements ......................................................................................... 603
Author contributions ...................................................................................... 604
Data sharing .................................................................................................... 611
References ........................................................................................................ 612
Appendices

Appendix 1 – Appendices to chapter 11 .................................................................663
Appendix 2 - Appendices to chapter 17 ..................................................................720
Appendix 3 – summary of changes to the original ELUCIDATE protocol ............727
Appendix 4 - Original Sample size calculation for the ELUCIDATE Trial ............731
# List of figures

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1</td>
<td>Figure 1</td>
<td>The Programme</td>
<td>64</td>
</tr>
<tr>
<td>Chapter 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Figure 2</td>
<td>Rigour of guideline development</td>
<td>75</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Figure 3</td>
<td>Flow diagram of the trial selection process</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>Figure 4</td>
<td>Number of trials published over time</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>Figure 5</td>
<td>Adequate description of monitoring strategy elements</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Figure 6</td>
<td>Citation of evidence to support features of monitoring strategy</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Figure 7</td>
<td>Aim of monitoring evaluation (n=68 experimental arms)</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>Figure 8</td>
<td>Trial validity</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>Figure 9</td>
<td>Comparison of observed versus predicted effects</td>
<td>106</td>
</tr>
<tr>
<td>Chapter 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 6</td>
<td>Figure 10</td>
<td>Monitoring care pathway</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>Figure 11</td>
<td>Summary of potential harms and benefits from a new monitoring strategy</td>
<td>143</td>
</tr>
<tr>
<td>Chapter 7</td>
<td>Figure 12</td>
<td>Fibrosis units linked to ELF score (left), ELF progression through time (centre), and starting stage ELF adjusted ELF progression through time (right)</td>
<td>164</td>
</tr>
<tr>
<td></td>
<td>Figure 13</td>
<td>Observed ELF measures</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>Figure 14</td>
<td>Results of implementing a monitoring strategy</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>Figure 15</td>
<td>Results of various monitoring strategies</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>Figure 16</td>
<td>Results of strategies using data with adjusted fibrosis progression estimate</td>
<td>203</td>
</tr>
<tr>
<td>Chapter 8</td>
<td>Figure 17</td>
<td>Relationship between cut-point and test performance</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td>Figure 18</td>
<td>Relationship between cut-point and Test Performance</td>
<td>217</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>Figure 19</td>
<td>The cost effectiveness bookshelf</td>
<td>219</td>
<td></td>
</tr>
<tr>
<td>Figure 20</td>
<td>Characterizing personalised medicine technologies</td>
<td>223</td>
<td></td>
</tr>
<tr>
<td>Figure 21</td>
<td>Decision tree for a 3 period monitoring test</td>
<td>227</td>
<td></td>
</tr>
<tr>
<td>Figure 22</td>
<td>Data for ROC curve at 2\textsuperscript{nd} administration for test positive individuals</td>
<td>232</td>
<td></td>
</tr>
<tr>
<td>Figure 23</td>
<td>Data for ROC curve at 2\textsuperscript{nd} administration for test negative individuals</td>
<td>233</td>
<td></td>
</tr>
<tr>
<td>Chapter 9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 10</td>
<td>Figure 24</td>
<td>Linear depiction of the phases in the biomarker pathway</td>
<td>262</td>
</tr>
<tr>
<td>Figure 25</td>
<td>Cyclic framework for the evaluation of in vitro medical tests</td>
<td>263</td>
<td></td>
</tr>
<tr>
<td>Figure 26</td>
<td>An overview of WS2 Key activities and deliverables are shown together with the inter-relationships with WS1 and WS3</td>
<td>278</td>
<td></td>
</tr>
<tr>
<td>Figure 27</td>
<td>Schematic of the integrated approach adopted within the Programme to sample banking and clinical data collection</td>
<td>280</td>
<td></td>
</tr>
<tr>
<td>Figure 28</td>
<td>The phased approach to biomarker validation with evidence-based progression and utilisation of sample banks depending on the stage of the biomarker</td>
<td>281</td>
<td></td>
</tr>
<tr>
<td>Chapter 11</td>
<td>Figure 29</td>
<td>ELF sample and Liver Biobank sample kit</td>
<td>303</td>
</tr>
<tr>
<td>Figure 30</td>
<td>Renal Transplant and Renal Cell Carcinoma sample kit</td>
<td>304</td>
<td></td>
</tr>
<tr>
<td>Figure 31</td>
<td>Monthly and cumulative accrual of patients with suspected RCC</td>
<td>312</td>
<td></td>
</tr>
<tr>
<td>Figure 32</td>
<td>RCC targets and accrual by centre</td>
<td>313</td>
<td></td>
</tr>
<tr>
<td>Figure 33</td>
<td>Renal Transplant: monthly and cumulative recruitment from the waiting list and transplantation</td>
<td>314</td>
<td></td>
</tr>
<tr>
<td>Figure 34</td>
<td>Renal Transplant targets and accrual by centre</td>
<td>315</td>
<td></td>
</tr>
<tr>
<td>Figure 35</td>
<td>Time taken for R&amp;D approvals plotted over time</td>
<td>324</td>
<td></td>
</tr>
<tr>
<td>Figure 36</td>
<td>SSI Submission as a percentage of total time for local review and approval</td>
<td>325</td>
<td></td>
</tr>
<tr>
<td>Figure 37</td>
<td>Combined timescales for overall study set-up time plotted over time</td>
<td>325</td>
<td></td>
</tr>
<tr>
<td>Chapter 12</td>
<td>Figure 38</td>
<td>The systematic literature review process adopted for circulating serum or plasma biomarkers of prognosis in clear cell RCC</td>
<td>339</td>
</tr>
<tr>
<td>Chapter 13</td>
<td>Figure 39</td>
<td>The systematic literature review process adopted for circulating biomarkers of DGF in renal transplantation</td>
<td>354</td>
</tr>
<tr>
<td>Figure 40</td>
<td>The effects of immediate versus delayed processing on plasma OPN concentrations (n=9)</td>
<td>381</td>
<td></td>
</tr>
<tr>
<td>Figure 41</td>
<td>Results from the assessment of parallelism of the CAIX assays</td>
<td>387</td>
<td></td>
</tr>
<tr>
<td>Figure 42</td>
<td>The relationship between concentrations of CAIX in EDTA plasma or serum</td>
<td>389</td>
<td></td>
</tr>
<tr>
<td>Figure 43</td>
<td>The effect of crossing over antibodies between assays to determine which antibody-antigen interaction accounts for the metal-ion dependent effects seen in the measurement of CAIX</td>
<td>390</td>
<td></td>
</tr>
<tr>
<td>Figure 44</td>
<td>Results from the assessment of parallelism for each of the five assays, comparing dilution-adjusted NGAL concentrations (log scale) against serial double dilutions of each of three samples represented by different colours</td>
<td>395</td>
<td></td>
</tr>
<tr>
<td>Figure 45</td>
<td>Results from the hook effect analysis</td>
<td>398</td>
<td></td>
</tr>
<tr>
<td>Figure 46</td>
<td>Modified Bland-Altman Plots for various comparisons of the NGAL assays</td>
<td>399</td>
<td></td>
</tr>
<tr>
<td>Figure 47</td>
<td>Frequency distribution of plasma VEGF concentrations as a percentage of the serum VEGF for 426 patients with RCC</td>
<td>405</td>
<td></td>
</tr>
<tr>
<td>Figure 48</td>
<td>Frequency distribution of the difference between serum VEGF and the average of the matched serum and plasma VEGF results for each of 430 RCC patients expressed as a % of that average</td>
<td>407</td>
<td></td>
</tr>
<tr>
<td>Figure 49</td>
<td>Greiner blood collection tubes showing the differences and overlap in colour closures and labelling (top) and the printed indication of additives (bottom)</td>
<td>408</td>
<td></td>
</tr>
<tr>
<td>Figure 50</td>
<td>Relationship between serum VEGF and plasma or serum minus plasma VEGF concentrations (n=380)</td>
<td>409</td>
<td></td>
</tr>
<tr>
<td>Figure 51</td>
<td>Comparison of VEGF concentrations of matched plasma and serum samples stored frozen and thawed immediately prior to analysis, with paired aliquots which had been subjected to an additional freeze-thaw cycle</td>
<td>411</td>
<td></td>
</tr>
<tr>
<td>Chapter 14</td>
<td>Figure 52</td>
<td>Box and whisker plots for pre-operative circulating CRP, OPN, CAIX and VEGF according to TNM stage (I-IV)</td>
<td>428</td>
</tr>
<tr>
<td></td>
<td>Figure 53</td>
<td>Box and whisker plots for pre-operative circulating CRP, OPN, CAIX and VEGF according to Leibovich risk classification (low, intermediate or high)</td>
<td>429</td>
</tr>
<tr>
<td></td>
<td>Figure 54</td>
<td>Kaplan-Meier survival curves showing MFS by TNM stage, Fuhrman grade and Leibovich score</td>
<td>430</td>
</tr>
<tr>
<td></td>
<td>Figure 55</td>
<td>Kaplan-Meier survival curves showing MFS for dichotomised markers, serum sodium and platelet count (only markers found significant in univariate analysis shown)</td>
<td>432</td>
</tr>
<tr>
<td></td>
<td>Figure 56</td>
<td>Kaplan-Meier survival curves showing MFS for dichotomised markers, serum sodium and platelet count (only markers found significant in univariate analysis shown) with the Leibovich score included as a further predictor variable</td>
<td>435</td>
</tr>
<tr>
<td>Chapter 15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 17</td>
<td>Figure 57</td>
<td>Feather diagram depicting biological, pre-analytical and analytical factors contributing towards measurement uncertainty ($U_M$)</td>
<td>451</td>
</tr>
<tr>
<td></td>
<td>Figure 58</td>
<td>Schematic illustration of the relationships between precision, bias, trueness accuracy and uncertainty</td>
<td>452</td>
</tr>
<tr>
<td></td>
<td>Figure 59</td>
<td>Levey–Jennings charts of ELF Scores across five days</td>
<td>460</td>
</tr>
<tr>
<td></td>
<td>Figure 60</td>
<td>Difference plot showing the bias of three manufacturers reference QC materials</td>
<td>464</td>
</tr>
<tr>
<td>Chapter</td>
<td>Figure 61</td>
<td>The ELUCIDATE flow chart: Flow chart showing ELUCIDATE recruitment, randomisation and follow-up procedures</td>
<td>473</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
<td>--------------------------------------------------------------------------------------------------</td>
<td>-----</td>
</tr>
<tr>
<td>Chapter 19</td>
<td>Figure 62</td>
<td>A comparison of the model prediction of the identification of patients as cirrhotic, with the observed patterns in the ELUCIDATE trial</td>
<td>489</td>
</tr>
<tr>
<td></td>
<td>Figure 63</td>
<td>Estimated pattern of development of cirrhosis without a trial extension – showing follow-up times assuming intervention finishes at end of October 2014</td>
<td>490</td>
</tr>
<tr>
<td></td>
<td>Figure 64</td>
<td>Expected incidence of severe complications as related to presentation ELF score</td>
<td>493</td>
</tr>
<tr>
<td></td>
<td>Figure 65</td>
<td>Simulation of the development of severe liver complications in the trial by randomised arms</td>
<td>497</td>
</tr>
<tr>
<td></td>
<td>Figure 66</td>
<td>Simulated trial outcomes for each trial arm</td>
<td>498</td>
</tr>
<tr>
<td>Chapter 20</td>
<td>Figure 67</td>
<td>Monthly and cumulative accrual</td>
<td>505</td>
</tr>
<tr>
<td>Chapter 21</td>
<td>Figure 68</td>
<td>Time between pre-cirrhotic follow-up visits</td>
<td>514</td>
</tr>
<tr>
<td></td>
<td>Figure 69</td>
<td>Time between post-cirrhotic follow-up visits</td>
<td>516</td>
</tr>
<tr>
<td></td>
<td>Figure 70</td>
<td>Overall AFP timings</td>
<td>519</td>
</tr>
<tr>
<td></td>
<td>Figure 71</td>
<td>Overall scan timings</td>
<td>521</td>
</tr>
<tr>
<td></td>
<td>Figure 72</td>
<td>Number of AFP measurements per randomised patient by arm</td>
<td>528</td>
</tr>
<tr>
<td></td>
<td>Figure 73</td>
<td>Number of AFP measurements post-diagnosis per 6 month period</td>
<td>529</td>
</tr>
<tr>
<td></td>
<td>Figure 74</td>
<td>Number of USS per randomised patients by arm</td>
<td>531</td>
</tr>
<tr>
<td></td>
<td>Figure 75</td>
<td>Number of USS post diagnosis per 6 month period</td>
<td>532</td>
</tr>
<tr>
<td>Chapter 22</td>
<td>Figure 76</td>
<td>Leeds NIHR DEC diagnostic evaluation pipeline and</td>
<td>586</td>
</tr>
<tr>
<td>Chapter 23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter</td>
<td>Figure 76</td>
<td>Leeds NIHR DEC diagnostic evaluation pipeline and</td>
<td>586</td>
</tr>
<tr>
<td>24</td>
<td>development areas</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## List of tables

<table>
<thead>
<tr>
<th>Chapter 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 2</td>
<td></td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Table 1</td>
</tr>
<tr>
<td></td>
<td>Table 2</td>
</tr>
<tr>
<td></td>
<td>Table 3</td>
</tr>
<tr>
<td></td>
<td>Table 4</td>
</tr>
<tr>
<td></td>
<td>Table 5</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Table 6</td>
</tr>
<tr>
<td></td>
<td>Table 7</td>
</tr>
<tr>
<td></td>
<td>Table 8</td>
</tr>
<tr>
<td></td>
<td>Table 9</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Table 10</td>
</tr>
<tr>
<td>Chapter 6</td>
<td>Table 11</td>
</tr>
<tr>
<td></td>
<td>Table 12</td>
</tr>
<tr>
<td>Chapter 7</td>
<td>Table 13</td>
</tr>
<tr>
<td></td>
<td>Table 14</td>
</tr>
<tr>
<td></td>
<td>Table 15</td>
</tr>
<tr>
<td></td>
<td>Table 16</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>17</td>
<td>Results of strategies A-H</td>
</tr>
<tr>
<td>18</td>
<td>Results using retest monitoring strategy (strategy B) by observation point</td>
</tr>
<tr>
<td>19</td>
<td>Results using reduced frequency of monitoring strategy (strategy C) by observation point</td>
</tr>
<tr>
<td>20</td>
<td>Results using absolute increase from start value monitoring strategy (strategy D) by observation point</td>
</tr>
<tr>
<td>21</td>
<td>Results using absolute increase from last value monitoring strategy (strategy E) by observation point</td>
</tr>
<tr>
<td>22</td>
<td>Results using relative increase from start value monitoring strategy (strategy F) by observation point</td>
</tr>
<tr>
<td>23</td>
<td>Results using relative increase from last value monitoring strategy (strategy G) by observation point</td>
</tr>
<tr>
<td>24</td>
<td>Result using linear regression monitoring strategy (strategy H) by observation point</td>
</tr>
<tr>
<td>25</td>
<td>Results of using the reference strategy when changing estimates required for data simulation</td>
</tr>
<tr>
<td>26</td>
<td>Results using reference strategy with decreased measurement error by observation point</td>
</tr>
<tr>
<td>27</td>
<td>Results using reference strategy with decreased measurement error by observation point and PPV at 25%</td>
</tr>
<tr>
<td>28</td>
<td>Results using reference strategy with increased measurement error by observation point</td>
</tr>
<tr>
<td>29</td>
<td>Results using reference strategy with increased measurement error by observation point and PPV at 25%</td>
</tr>
<tr>
<td>30</td>
<td>Results using reference strategy with decreased between-individual variability by observation point</td>
</tr>
<tr>
<td>31</td>
<td>Results using reference strategy with decreased between-individual variability by observation point and PPV at 25%</td>
</tr>
<tr>
<td>32</td>
<td>Results using reference strategy with increased between-individual variability by observation point</td>
</tr>
<tr>
<td>Table 33</td>
<td>Results using reference strategy with increased between-individual variability by observation point and PPV at 25%</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Table 34</td>
<td>Results using reference strategy with decreased fibrosis progression rate by observation point</td>
</tr>
<tr>
<td>Table 35</td>
<td>Results using reference strategy with decreased fibrosis progression rate by observation point and PPV at 25%</td>
</tr>
<tr>
<td>Table 36</td>
<td>Results using reference strategy with increased fibrosis progression rate by observation point</td>
</tr>
<tr>
<td>Table 37</td>
<td>Results using reference strategy with increased fibrosis progression rate by observation point and PPV at 25%</td>
</tr>
<tr>
<td>Table 38</td>
<td>Results of strategies A-H for unadjusted fibrosis progression estimate data</td>
</tr>
<tr>
<td>Table 39</td>
<td>Results by observation point for the reference strategy (strategy A) for adjusted fibrosis progression estimate data</td>
</tr>
<tr>
<td>Table 40</td>
<td>Results using retest monitoring strategy (strategy B) by observation point for adjusted fibrosis progression estimate data</td>
</tr>
<tr>
<td>Table 41</td>
<td>Results using reduced frequency of monitoring strategy (strategy C) by observation point for adjusted fibrosis progression estimate data</td>
</tr>
<tr>
<td>Table 42</td>
<td>Results using absolute increase from start value monitoring strategy (strategy D) by observation point for adjusted fibrosis progression estimate data</td>
</tr>
<tr>
<td>Table 43</td>
<td>Results using absolute increase from last value monitoring strategy (strategy E) by observation point for adjusted fibrosis progression estimate data</td>
</tr>
<tr>
<td>Table 44</td>
<td>Results using relative increase from start value monitoring strategy (strategy F) by observation point for adjusted fibrosis progression estimate data</td>
</tr>
<tr>
<td>Table 45</td>
<td>Results using relative increase from last value monitoring strategy (strategy G) by observation point for adjusted fibrosis progression estimate data</td>
</tr>
<tr>
<td>Table 46</td>
<td>Results using linear regression monitoring strategy (strategy H) by observation point for adjusted fibrosis progression estimate data</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>47</td>
<td>Results of using the reference strategy when changing estimates required for data simulation</td>
</tr>
<tr>
<td></td>
<td>(difference to reference strategy with no change in simulation data)</td>
</tr>
<tr>
<td>48</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy</td>
</tr>
<tr>
<td></td>
<td>with decreased measurement error by observation point</td>
</tr>
<tr>
<td>49</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy</td>
</tr>
<tr>
<td></td>
<td>with decreased measurement error by observation point and PPV at 25%</td>
</tr>
<tr>
<td>50</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy</td>
</tr>
<tr>
<td></td>
<td>with increased measurement error by observation point</td>
</tr>
<tr>
<td>51</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy</td>
</tr>
<tr>
<td></td>
<td>with increased measurement error by observation point and PPV at 25%</td>
</tr>
<tr>
<td>52</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy</td>
</tr>
<tr>
<td></td>
<td>with decreased between-individual</td>
</tr>
<tr>
<td>53</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy</td>
</tr>
<tr>
<td></td>
<td>with decreased between-individual variability by observation point and PPV at 25%</td>
</tr>
<tr>
<td>54</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy</td>
</tr>
<tr>
<td></td>
<td>with increased between-individual variability by observation point</td>
</tr>
<tr>
<td>55</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy</td>
</tr>
<tr>
<td></td>
<td>with increased between-individual variability by observation point and PPV at 25%</td>
</tr>
<tr>
<td>56</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy</td>
</tr>
<tr>
<td></td>
<td>with decreased fibrosis progression rate by observation point</td>
</tr>
<tr>
<td>57</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy</td>
</tr>
<tr>
<td></td>
<td>with decreased fibrosis progression rate by observation point and PPV at 25%</td>
</tr>
<tr>
<td>58</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy</td>
</tr>
<tr>
<td></td>
<td>with increased fibrosis progression rate by observation point</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Table 59</td>
<td>Adjusted fibrosis progression sensitivity analyses – results using reference strategy with increased fibrosis progression rate by observation point and PPV at 25%</td>
</tr>
<tr>
<td>Table 60</td>
<td>Results of analysis of randomisation ELF and analysis of variance for ELF measurements at all time points</td>
</tr>
<tr>
<td>Table 61</td>
<td>Results of multilevel model of repeated ELF measures from ELUCIDATE trial</td>
</tr>
<tr>
<td>Table 62</td>
<td>Results of multilevel model of simulated repeated ELF measurements</td>
</tr>
<tr>
<td>Table 63</td>
<td>Results of multilevel model of simulated repeated ELF measurements with adjusted fibrosis progression rate</td>
</tr>
<tr>
<td>Chapter 8</td>
<td>Table 64 Impact of changing cut-points on cost and outcomes</td>
</tr>
<tr>
<td>Table 65</td>
<td>Parameter values for illustrative monitoring model</td>
</tr>
<tr>
<td>Table 66</td>
<td>Optimising cut-offs</td>
</tr>
<tr>
<td>Chapter 9</td>
<td></td>
</tr>
<tr>
<td>Chapter 10</td>
<td></td>
</tr>
<tr>
<td>Chapter 11</td>
<td>Table 67 Centres recruiting patients with renal cell carcinoma</td>
</tr>
<tr>
<td>Table 68</td>
<td>Renal Cell Carcinoma Study Schema</td>
</tr>
<tr>
<td>Table 69</td>
<td>Renal Transplant Study Centres</td>
</tr>
<tr>
<td>Table 70</td>
<td>Renal Transplant study Schema</td>
</tr>
<tr>
<td>Table 71</td>
<td>Sample processing SOPs compared to examples of published standards and guidelines</td>
</tr>
<tr>
<td>Table 72</td>
<td>Sample processing times and percentage compliance with SSOP</td>
</tr>
<tr>
<td>Table 73</td>
<td>RCC CRF compliance</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Table 74</td>
<td>Renal Transplant CRF Compliance</td>
</tr>
<tr>
<td>Table 75</td>
<td>Pros and Cons of Research Tissue Banks vs Project-Specific Approval</td>
</tr>
<tr>
<td>Table 76</td>
<td>Timescales for gaining national approval</td>
</tr>
<tr>
<td>Table 77</td>
<td>RCC study timescales for gaining local centre approval</td>
</tr>
<tr>
<td>Table 78</td>
<td>Renal Transplant study timescales for gaining local centre approval</td>
</tr>
<tr>
<td>Table 79</td>
<td>Characteristics of liver disease patients</td>
</tr>
<tr>
<td>Table 80</td>
<td>Characteristics of all patients recruited with suspected RCC</td>
</tr>
<tr>
<td>Table 81</td>
<td>Characteristics of the RCC patient cohort with clear cell subtype</td>
</tr>
<tr>
<td>Table 82</td>
<td>Healthy Control Characteristics</td>
</tr>
<tr>
<td>Table 83</td>
<td>Characteristics of the renal transplant patients</td>
</tr>
<tr>
<td>Table 84</td>
<td>List of available CLSI Evaluation Protocols</td>
</tr>
<tr>
<td>Table 85</td>
<td>Comparison of plasma and serum osteopontin concentrations in matched samples</td>
</tr>
<tr>
<td>Table 86</td>
<td>The recovery of OPN spiked into EDTA plasma samples</td>
</tr>
<tr>
<td>Table 87</td>
<td>Recovery of rCAIX spiked into EDTA plasma and serum samples</td>
</tr>
<tr>
<td>Table 88</td>
<td>Summary of the performance data for the five NGAL assays evaluated using urine samples</td>
</tr>
<tr>
<td>Chapter</td>
<td>Table</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td>14</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>93</td>
</tr>
<tr>
<td>15</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td>18</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>102</td>
</tr>
<tr>
<td>Chapter</td>
<td>Table</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>19</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Table 103</td>
</tr>
<tr>
<td></td>
<td>Table 104</td>
</tr>
<tr>
<td></td>
<td>Table 105</td>
</tr>
<tr>
<td>21</td>
<td>Table 106</td>
</tr>
<tr>
<td></td>
<td>Table 107</td>
</tr>
<tr>
<td></td>
<td>Table 108</td>
</tr>
<tr>
<td></td>
<td>Table 109</td>
</tr>
<tr>
<td></td>
<td>Table 110</td>
</tr>
<tr>
<td></td>
<td>Table 111</td>
</tr>
<tr>
<td></td>
<td>Table 112</td>
</tr>
<tr>
<td></td>
<td>Table 113</td>
</tr>
<tr>
<td></td>
<td>Table 114</td>
</tr>
<tr>
<td></td>
<td>Table 115</td>
</tr>
<tr>
<td></td>
<td>Table 116</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Table 117</td>
<td>ELF arm - Overall ELF test compliance</td>
</tr>
<tr>
<td>Table 118</td>
<td>Standard arm - ELF test compliance</td>
</tr>
<tr>
<td>Table 119</td>
<td>Overall compliance for patients who were expected to have at least one measurement</td>
</tr>
<tr>
<td>Table 120</td>
<td>Overall timings (for all AFP measurements due to take place 6 months after the previous one)</td>
</tr>
<tr>
<td>Table 121</td>
<td>Overall scan compliance for patients who were expected to have at least one measurement</td>
</tr>
<tr>
<td>Table 122</td>
<td>Overall scan timings (for all scans due to take place 6 months after the previous one)</td>
</tr>
<tr>
<td>Table 123</td>
<td>Overall compliance for patients who are expected to have at least one Endoscopy</td>
</tr>
<tr>
<td>Table 124</td>
<td>Diagnosis of Cirrhosis overall (by arm)</td>
</tr>
<tr>
<td>Table 125</td>
<td>Timing of diagnosis i.e. at/after randomisation (by arm)</td>
</tr>
<tr>
<td>Table 126</td>
<td>Method of diagnosis of Cirrhosis (overall)</td>
</tr>
<tr>
<td>Table 127</td>
<td>Method of diagnosis of Cirrhosis (for patients who were not diagnosed at randomisation)</td>
</tr>
<tr>
<td>Table 128</td>
<td>First method of diagnosis of Cirrhosis (overall)</td>
</tr>
<tr>
<td>Table 129</td>
<td>First method of diagnosis of Cirrhosis (for patients who were not diagnosed at randomisation)</td>
</tr>
<tr>
<td>Table 130</td>
<td>Number of randomised patients having at least one biopsy (pre or post diagnosis of cirrhosis)</td>
</tr>
<tr>
<td>Table 131</td>
<td>Number of AFP tests per randomised patient by arm</td>
</tr>
<tr>
<td>Table 132</td>
<td>Number of AFP tests per randomised patients by arm</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Table 133</td>
<td>Frequency of AFP before and after diagnosis of cirrhosis</td>
</tr>
<tr>
<td>Table 134</td>
<td>Number of Ultrasound Scans (USS) per randomised patient by arm</td>
</tr>
<tr>
<td>Table 135</td>
<td>Number of Ultrasound Scans (USS) per randomised patients by arm – Wilcoxon Test</td>
</tr>
<tr>
<td>Table 136</td>
<td>Frequency of USS before and after diagnosis of cirrhosis</td>
</tr>
<tr>
<td>Table 137</td>
<td>Number of randomised patients having at least one OGD since randomisation</td>
</tr>
<tr>
<td>Table 138</td>
<td>Number of patients having at least one OGD post diagnosis of Cirrhosis</td>
</tr>
<tr>
<td>Table 139</td>
<td>Number of OGDS by randomisation result</td>
</tr>
<tr>
<td>Table 140</td>
<td>Number of patients who have developed varices (by arm)</td>
</tr>
<tr>
<td>Table 141</td>
<td>Number of randomised patients being treated with beta blockers or band ligation</td>
</tr>
<tr>
<td>Table 142</td>
<td>Number of randomised patients receiving treatment to normalise LFTs</td>
</tr>
<tr>
<td>Table 143</td>
<td>Liver Biopsy</td>
</tr>
<tr>
<td>Table 144</td>
<td>Ultrasound scans</td>
</tr>
<tr>
<td>Table 145</td>
<td>AFP test</td>
</tr>
<tr>
<td>Table 146</td>
<td>Endoscopy OGD</td>
</tr>
<tr>
<td>Table 147</td>
<td>Treatment with beta-blockers</td>
</tr>
<tr>
<td>Chapter 22</td>
<td></td>
</tr>
<tr>
<td>Chapter 23</td>
<td>Table 148 ELUCIDATE Patient characteristics</td>
</tr>
<tr>
<td>Chapter 24</td>
<td></td>
</tr>
</tbody>
</table>
## List of boxes

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Box</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Criteria used to assess rigour of guideline development</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Criteria used to assess rigour of guideline development, with details and examples</td>
<td>74</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>Full trial reports</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Validity assessment</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Recommendations for future practice</td>
<td>112</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 21</td>
<td>Box 6</td>
<td>Unit costs and their sources</td>
<td>543</td>
</tr>
<tr>
<td>Chapter 22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 24</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Glossary

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACY-1</td>
<td>Serum aminoacylase-1</td>
</tr>
<tr>
<td>AFP</td>
<td>alpha-fetoprotein</td>
</tr>
<tr>
<td>AGREE</td>
<td>The Appraisal of Guidelines for Research &amp; Evaluation (AGREE) Instrument</td>
</tr>
<tr>
<td>AIDS</td>
<td>Acquired Immune Deficiency Syndrome</td>
</tr>
<tr>
<td>AKI</td>
<td>acute kidney injury</td>
</tr>
<tr>
<td>ALP</td>
<td>alkaline phosphatase</td>
</tr>
<tr>
<td>ALT</td>
<td>Alanine transaminase</td>
</tr>
<tr>
<td>AQP-1</td>
<td>aquaporin-1</td>
</tr>
<tr>
<td>AR</td>
<td>acute rejection</td>
</tr>
<tr>
<td>ASTRO</td>
<td>American Society for Therapeutic Radiology and Oncology</td>
</tr>
<tr>
<td>AUA</td>
<td>American Urological Association</td>
</tr>
<tr>
<td>AUC</td>
<td>area under the curve</td>
</tr>
<tr>
<td>Aus CN</td>
<td>Australian Cancer Network</td>
</tr>
<tr>
<td>B7-H3</td>
<td>B7 family ligand</td>
</tr>
<tr>
<td>BF</td>
<td>biochemical failure</td>
</tr>
<tr>
<td>bFGF</td>
<td>basic fibroblast growth factor</td>
</tr>
<tr>
<td>BMI</td>
<td>body mass index</td>
</tr>
<tr>
<td>BRISQ</td>
<td>Biospecimen Reporting for Improved Study Quality</td>
</tr>
<tr>
<td>CA-15-3</td>
<td>cancer antigen 15-3</td>
</tr>
<tr>
<td>CA-19-9</td>
<td>cancer antigen 19-9</td>
</tr>
<tr>
<td>CA-50</td>
<td>cancer antigen 50</td>
</tr>
<tr>
<td>CA-125</td>
<td>cancer antigen 125</td>
</tr>
<tr>
<td>CAF</td>
<td>C-Terminal Agrin Fragment</td>
</tr>
<tr>
<td>CAIX</td>
<td>carbonic anhydrase IX</td>
</tr>
<tr>
<td>ccRCC</td>
<td>clear cell renal cell carcinoma</td>
</tr>
<tr>
<td>CD4</td>
<td>cluster of differentiation 4</td>
</tr>
<tr>
<td>CADTH</td>
<td>Canadian Agency for Drugs and Technologies in Health</td>
</tr>
<tr>
<td>CCRNs</td>
<td>Comprehensive Clinical Research Networks</td>
</tr>
<tr>
<td>CE</td>
<td>Conformité Européene</td>
</tr>
<tr>
<td>CEA</td>
<td>carcinoembryonic antigen</td>
</tr>
<tr>
<td>CENTRAL</td>
<td>Cochrane Central Register of Controlled Trials</td>
</tr>
<tr>
<td>CIT</td>
<td>cold ischaemic time</td>
</tr>
<tr>
<td>CKD</td>
<td>chronic kidney disease</td>
</tr>
<tr>
<td>CLD</td>
<td>Chronic liver disease</td>
</tr>
<tr>
<td>CLSI</td>
<td>Clinical and Laboratory Standards Institute</td>
</tr>
<tr>
<td>CRP</td>
<td>C-reactive protein</td>
</tr>
<tr>
<td>CRR</td>
<td>creatinine reduction ratio</td>
</tr>
<tr>
<td>CSS</td>
<td>cancer-specific survival</td>
</tr>
<tr>
<td>CT</td>
<td>computed tomography</td>
</tr>
<tr>
<td>CTRU</td>
<td>Clinical Trials Research Unit</td>
</tr>
<tr>
<td>CUZD1</td>
<td>CUB and zona pellucida-like domains protein 1</td>
</tr>
<tr>
<td>CV</td>
<td>coefficient variation</td>
</tr>
<tr>
<td>DDB</td>
<td>donation after brain death</td>
</tr>
<tr>
<td>DCD</td>
<td>donation after circulatory death</td>
</tr>
<tr>
<td>DEC</td>
<td>Diagnostic Evidence Co-operative</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>DGF</td>
<td>delayed graft function</td>
</tr>
<tr>
<td>DMEC</td>
<td>Data Monitoring and Ethics Committee</td>
</tr>
<tr>
<td>DUA</td>
<td>Dutch Urological Association</td>
</tr>
<tr>
<td>EAU</td>
<td>European Association of Urology</td>
</tr>
<tr>
<td>ECD</td>
<td>expanded-criteria donors</td>
</tr>
<tr>
<td>ECOG PS</td>
<td>Eastern Co-operative Oncology Group performance status</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>ELF</td>
<td>enhanced liver fibrosis</td>
</tr>
<tr>
<td>ELISA</td>
<td>enzyme-linked immunosorbent assay</td>
</tr>
<tr>
<td>EPO</td>
<td>erythropoietin</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>EVPI</td>
<td>expected value of perfect information</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>FFPE</td>
<td>formalin-fixed paraffin-embedded</td>
</tr>
<tr>
<td>Flt-1</td>
<td>Fms-Like Tyrosine Kinase</td>
</tr>
<tr>
<td>FN</td>
<td>false negatives</td>
</tr>
<tr>
<td>FP</td>
<td>false positives</td>
</tr>
<tr>
<td>HAS</td>
<td>Haute Autorité de santé</td>
</tr>
<tr>
<td>HBV</td>
<td>Hepatitis B virus</td>
</tr>
<tr>
<td>HCC</td>
<td>hepatocellular cancer</td>
</tr>
<tr>
<td>HCG</td>
<td>human chorionic gonadotropin</td>
</tr>
<tr>
<td>HCV</td>
<td>Hepatitis C Virus</td>
</tr>
<tr>
<td>HIF</td>
<td>hypoxia-inducible factor</td>
</tr>
<tr>
<td>HIV</td>
<td>Human Immunodeficiency Virus</td>
</tr>
<tr>
<td>HLA</td>
<td>human leukocyte antigen</td>
</tr>
<tr>
<td>HSCIC</td>
<td>Health and Social Care Information Centre</td>
</tr>
<tr>
<td>IAP</td>
<td>immuno suppressive acidic protein</td>
</tr>
<tr>
<td>ICER</td>
<td>incremental cost-effectiveness ratio</td>
</tr>
<tr>
<td>IFCC</td>
<td>International Federation of Clinical Chemistry and Laboratory Medicine</td>
</tr>
<tr>
<td>IgA</td>
<td>Immunoglobulin A</td>
</tr>
<tr>
<td>IL-16</td>
<td>Interleukin-16</td>
</tr>
<tr>
<td>IL-18</td>
<td>Interleukin-18</td>
</tr>
<tr>
<td>IMDC</td>
<td>International Metastatic RCC Database Consortium</td>
</tr>
<tr>
<td>IOU</td>
<td>integrated Onstein-Uhlnbeck</td>
</tr>
<tr>
<td>IQR</td>
<td>interquartile range</td>
</tr>
<tr>
<td>IRI</td>
<td>ischaemia-reperfusion injury</td>
</tr>
<tr>
<td>ISBER</td>
<td>International Society for Biological and Environmental Repositories</td>
</tr>
<tr>
<td>ISRCTN</td>
<td>International Standard Randomised Controlled Trial</td>
</tr>
<tr>
<td>ITT</td>
<td>Intention to treat</td>
</tr>
<tr>
<td>LEPS</td>
<td>Laboratory Errors and Patient Safety</td>
</tr>
<tr>
<td>LLoQ</td>
<td>Lower Limit of Quantification</td>
</tr>
<tr>
<td>LTHT</td>
<td>Leeds Teaching Hospitals Trust</td>
</tr>
<tr>
<td>MFS</td>
<td>Metastasis free survival</td>
</tr>
<tr>
<td>MMP-7</td>
<td>Matrix metalloproteinase-7</td>
</tr>
<tr>
<td>MVI</td>
<td>microvascular invasion</td>
</tr>
<tr>
<td>NCCN</td>
<td>National Comprehensive Cancer Network</td>
</tr>
<tr>
<td>NCI PDQ</td>
<td>National Cancer Institute – Physician Data Query</td>
</tr>
<tr>
<td>NGAL</td>
<td>Neutrophil Gelatinase Associated Lipocalin</td>
</tr>
<tr>
<td>NHS</td>
<td>National Health Service</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>NH</td>
<td>natural history</td>
</tr>
<tr>
<td>NICE</td>
<td>National Institute for Health and Clinical Excellence</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>NIHR</td>
<td>National Institute for Health Research</td>
</tr>
<tr>
<td>NIHR CRN</td>
<td>National Institute for Health Research Clinical Research Network</td>
</tr>
<tr>
<td>NLR</td>
<td>neutrophil/lymphocyte ratio</td>
</tr>
<tr>
<td>OGD</td>
<td>oesophago-gastro-duodenoscopy</td>
</tr>
<tr>
<td>OPN</td>
<td>osteopontin</td>
</tr>
<tr>
<td>OS</td>
<td>overall survival</td>
</tr>
<tr>
<td>PAPP-A</td>
<td>Pregnancy-associated plasma protein</td>
</tr>
<tr>
<td>PBAC</td>
<td>Pharmaceutical Benefits Advisory Committee</td>
</tr>
<tr>
<td>PFS</td>
<td>Progression free survival</td>
</tr>
<tr>
<td>PIL</td>
<td>patient information leaflet</td>
</tr>
<tr>
<td>PIS</td>
<td>patient information summary</td>
</tr>
<tr>
<td>PLIN-2/ADFP</td>
<td>perlipin-2/adipophilin</td>
</tr>
<tr>
<td>PM</td>
<td>Phelps and Mushlin</td>
</tr>
<tr>
<td>PSA</td>
<td>prostate-specific antigen</td>
</tr>
<tr>
<td>PTEN</td>
<td>Phosphatase and tensin homolog</td>
</tr>
<tr>
<td>PTHLH</td>
<td>parathyroid hormone-like hormone</td>
</tr>
<tr>
<td>QALY</td>
<td>quality-adjusted life year</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>research and development</td>
</tr>
<tr>
<td>RCC</td>
<td>renal cell carcinoma</td>
</tr>
<tr>
<td>RCT</td>
<td>randomised controlled trial</td>
</tr>
<tr>
<td>RCV</td>
<td>reference change value</td>
</tr>
<tr>
<td>REC</td>
<td>research ethics committee</td>
</tr>
<tr>
<td>ROC</td>
<td>receiver-operator characteristic</td>
</tr>
<tr>
<td>RP</td>
<td>radical prostatectomy</td>
</tr>
<tr>
<td>RT</td>
<td>renal transplantation</td>
</tr>
<tr>
<td>RTB</td>
<td>research tissue bank</td>
</tr>
<tr>
<td>RTX</td>
<td>radical radiotherapy</td>
</tr>
<tr>
<td>RUSAE</td>
<td>related and unexpected serious adverse events</td>
</tr>
<tr>
<td>Rx</td>
<td>treatment</td>
</tr>
<tr>
<td>SAA</td>
<td>serum amyloid A</td>
</tr>
<tr>
<td>SCF</td>
<td>Stem Cell Factor</td>
</tr>
<tr>
<td>SCr</td>
<td>serum creatinine</td>
</tr>
<tr>
<td>SD</td>
<td>standard deviation</td>
</tr>
<tr>
<td>sIL-2R</td>
<td>soluble IL-2 receptor</td>
</tr>
<tr>
<td>SNOD</td>
<td>Specialist Nurse-Organ Donation</td>
</tr>
<tr>
<td>SPREC</td>
<td>Sample PREanalytical Code</td>
</tr>
<tr>
<td>SSIGN</td>
<td>Mayo Clinic Stage, Size, Grade and Necrosis score</td>
</tr>
<tr>
<td>SSOP</td>
<td>Study Site Operating Procedure</td>
</tr>
<tr>
<td>ST6Gall</td>
<td>Galbeta1-4GlcNAlpha2-6 sialytransferase</td>
</tr>
<tr>
<td>TKI</td>
<td>Tyrosine kinase inhibitor</td>
</tr>
<tr>
<td>TNM</td>
<td>tumour-node-metastasis</td>
</tr>
<tr>
<td>TN</td>
<td>true negative</td>
</tr>
<tr>
<td>TP</td>
<td>true positives</td>
</tr>
<tr>
<td>TPA</td>
<td>Tissue polypeptide antigen</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>UK NEQAS</td>
<td>United Kingdom National External Quality Assessment Service</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------------------------------------------------</td>
</tr>
<tr>
<td>UK PCWG</td>
<td>UK Prostate Cancer Working Group</td>
</tr>
<tr>
<td>US</td>
<td>United States</td>
</tr>
<tr>
<td>USA</td>
<td>United States of America</td>
</tr>
<tr>
<td>USS</td>
<td>ultrasound scan</td>
</tr>
<tr>
<td>UTI</td>
<td>urinary tract infection</td>
</tr>
<tr>
<td>VEGF</td>
<td>vascular endothelial growth factor</td>
</tr>
<tr>
<td>VEGFR-1</td>
<td>vascular endothelial growth factor receptor-1</td>
</tr>
<tr>
<td>VHL</td>
<td>Von Hippel Lindau</td>
</tr>
<tr>
<td>WBC</td>
<td>white cell count</td>
</tr>
<tr>
<td>WG-PRE</td>
<td>Working Group for Preanalytical Phase</td>
</tr>
<tr>
<td>WIT</td>
<td>warm ischaemic time</td>
</tr>
<tr>
<td>WS</td>
<td>workstream</td>
</tr>
</tbody>
</table>
Scientific summary

Protein biomarkers in body fluids which have demonstrable associations with the activity and outcome of a wide range of diseases are now being identified by modern proteomic technologies. They may be simple, accessible, cheap and safe tests which can inform diagnosis, prognosis, treatment selection, monitoring of disease activity and therapy. They may substitute for or augment more complex, invasive and expensive tests. However, their substantial potential to improve patient care and health service provision is not yet being realised because the pathway linking biomarker research to health services research is still quite poorly defined. Liver and renal diseases generate huge and growing patient and service burdens, are amenable to biomarker application.

Our NIHR Applied Programme addressed three workstreams which relate to the development pipeline for new biomarkers in renal and liver diseases and aimed to create a framework for research and innovation in this area:

WS1) METHODOLOGY. To define current best practice and explore innovations particularly in relation to the use of biomarkers to monitor disease activity.

WS2) CLINICAL TRANSLATION. To create and evaluate a framework of practice, samples and clinical data to rapidly identify protein biomarkers with the appropriate analytic and clinical validity and performance characteristics to justify evaluation of their clinical utility in the health service in liver and renal diseases.

WS3) ELUCIDATE RCT. A randomised trial on an established biomarker test, ELF, for liver fibrosis and cirrhosis for which clinical evidence for potential value in chronic liver disease is excellent, to determine whether its use will sufficiently alter the diagnostic timing and subsequent management of cirrhosis of the liver in order to change the process of care and reduce serious complications and improve outcomes for patients and service provision.

We assembled an outstanding internationally recognised multi-disciplinary team of methodologists, clinicians, clinical biochemists, statisticians and marker scientists to deliver these workstreams.
The methodology workstream evaluated published evidence on the quality of recommendations for using PSA to monitor patients with prostate cancer, systematically reviewed the use of randomised clinical trials to evaluate monitoring strategies, reviewed the monitoring biomarker literature and how monitoring can impact on patient outcomes, conducted simulation studies to evaluate monitoring strategies and how monitoring strategies can meet the requirements to improve the value of healthcare services. These studies confirm that the literature on the use of biomarkers in monitoring diseases is modest in scale and robust conclusions are infrequent and recommend improvements in research practice.

We considered the guidelines which were available for using PSA to monitor patients after they received either radical surgery or radical radiotherapy for localised disease. The guideline methods were assessed using a formal research evaluation framework which examined the systematic search methods used in the studies, the selection criteria, the clarity of the formulation of recommendations, the consideration given to relevant issues for monitoring in the recommendations, the explicit nature of the use of evidence, the use of external review and the description of updating procedures. Of the nine main guidelines evaluated, using an objective scoring system, the rigour of guideline development varied and scored best for the 2008 NICE guidelines. Only one guideline modified its recommendation to reflect the fact that a single PSA measurement may be technically unreliable and it did so by recommending re-testing within two months. Three recommended the use of the same assay on every test occasion. Only four guidelines attempted to justify the interval between tests that they recommended. Overall there was evidence of considerable inconsistency in guideline recommendations for the use of PSA even when they were published within a few years of each other. We concluded that general failings in the guideline development process are likely to contribute significantly to the variations between published guidelines. Only the National Institute for Health and Clinical Excellence (NICE) and the Australian Cancer Network cited handbooks on guideline development. It was notable that these guidelines scored relatively well on the evaluation instrument used in our study.

Randomised controlled trials of monitoring regimens are challenging to design and to deliver. They are complex, and involve serial testing. There are complex interactions between repeated test results, clinical decisions based on these results, the response of clinicians to the results and, of course, the identification through long follow up of important patient outcomes. We conducted a methodological review of RCTs of monitoring. Although the
target sample size was 60 RCTs, after a comprehensive search, 120 titles were selected for further evaluation. Following full text review 49 trials published in 58 publications were selected for inclusion. Cancer followed by cardiovascular disease and renal disease were the most frequently reported topics. Half of the trials evaluated patient related primary outcomes, one-third the impact on mortality and one-half aimed to report the impact of the monitoring strategy on the detection of new or recurrent disease. Process of care outcomes were evaluated primarily related to the number of patients treated between the different trial arms or the time taken to arrive at that treatment. Twelve trials reported statistically significant effects of monitoring on the primary outcome. There was only limited attention to test properties and intervention effectiveness in the population of interest, before trials were undertaken. There was a lack of detailed description of the protocols for trial monitoring and considerable evidence for lack of compliance for the monitoring strategies. The impact of the monitoring strategies on clinical behaviours such as to administer treatment or withhold treatment, was not always consistent with the test results. It appeared that the monitoring test was treated by clinicians as a guide to possible changes rather than a definitive indication for a particular change in care. There was an apparent lack of power to detect significant effects in the studies as a whole.

We reviewed the literature on monitoring strategies used to direct the care of patients with recurrent or progressive disease. After a formal search strategy and filtering, the literature was categorised and tabulated. The review identified a limited methodological literature for monitoring strategies. We then focused on the relationship between the monitoring care pathway and the points in that pathway where monitoring might be expected to affect patient outcomes. Three identified frameworks for this were reviewed. Clinical trials of relevance were grouped into three main categories (new monitoring strategies vs existing strategies; a monitoring strategy vs immediate treatment; a monitoring strategy vs no monitoring). Differences in study design from the use of biomarkers for screening and diagnosis were evaluated. Monitoring strategies considered included i) detection of significant clinical change earlier than conventional practice to deploy treatment early; ii) to reduce the invasiveness and cost of testing; iii) to reduce the volume and frequency of testing; iv) reduce over treatment; v) delay or avoid treatment. The analysis led to the recommendation that a test validation paradigm be adopted where a number of methods are used to determine whether the results of a test are going to be meaningful in practice and generate benefits for patients. RCTs will be needed in some settings but this level of evidence will not always be
essential. Strategic approaches need to be multidisciplinary involving evaluation of the performance of the tests in the laboratory, rigorous study design and analysis and close collaboration with clinicians and biochemists to determine the appropriate technical and clinical options for evaluation and the probability of changing clinical behaviours with test results.

Simulation studies are described in the report with the impact of the simulation studies on the conduct, re-design and extension of the ELUCIDATE RCT. Data sources were not always adequate for comprehensive simulation until quite late in the progress of the Programme and ELUCIDATE RCT. The modelling work allowed the accurate calculation of power based on observed and predicted event rates. This allowed the completion of trial recruitment, the reporting of process of care outcomes and the initiation of long term follow up strategies from healthcare informatics sources.

Introducing new biomarkers into clinical practice to promote the introduction of a more personalised, precise and stratified approach to patient management requires evaluation of the characteristics of the test and its impact upon clinical outcomes and the quality and cost of the care delivered. There is tremendous pressure to increase the efficiency of healthcare systems by introducing cost-effective new tests. The elimination of unnecessary tests is being explored. We focused especially on the role of monitoring tests and methods to evaluate their health economics. We compared the use of conventional clinical utilities with an approach based on cost effectiveness; described the framework for characterising personalised medicine technologies and drew on an existing method for optimising diagnostic tests to meet cost effectiveness targets and extended this to monitoring tests. This work demonstrated, among other things, that the cut points used for a test when used repeatedly for monitoring may under some circumstances be different from those when the test is used for diagnosis.

The findings were formulated to be shared with patients who strongly endorsed the need for robust and conclusive research in this area and for improved communication about test results between clinicians and patients.

The clinical translation workstream focused on the analytic and clinical validity of tests in renal disease. Prospective cohorts were established for renal cancer (RCC) and renal transplantation (RT), with samples and patient data from multiple NHS Centres and the
samples and patient data from WS3 (Liver disease) curated. The recruitment of patients obtaining high quality samples and clinical data was challenging but ultimately completed to target. These resources provide and will continue to provide a rapid access resource to evaluate the validity of biomarkers which are candidates to be evaluated to see if they can improve NHS services. In order to identify tests to be evaluated using this resource, all candidate biomarkers for RCC and RT were identified from the literature, the quality of studies evaluated and selected biomarkers were prioritised. Four selected biomarkers were studied further by rigorous evaluation of the validity of the tests and evaluation of their performance within the WS2 sample/data cohorts. Systematic evaluation of tests relevant to renal cancer in the literature suggested that Osteopontin, Vascular Endothelial Growth Factor, CAIX and C-Reactive Protein (CRP) were prioritised and evaluated further. For renal transplantation the most promising serum biomarkers for the early detection of delayed graft function appear to be NGAL and serum cystatin C and ACY-1, previously discovered by our group. The performance of available assays for the four prioritised biomarkers were rigorously evaluated including pre-analytic aspects and verification protocols. Therefore, specific biomarker technical evaluations were performed for all of the prioritised biomarkers studied within the programme. The important technical aspects of evaluating biomarker assays are illustrated in these studies as well as the critical importance of the principle that all assays must be technically robust before being employed in NHS diagnostics or in clinical trials. Without assay characterisation and validation as part of the early phase of biomarker translation the field will continue to move slowly and to waste resources. High quality biobanking and detailed consideration of pre-analytic factors are essential in this field. The four renal cancer biomarkers evaluated in the NIHR Applied Biomarker Programme cohorts showed promise but at this stage, we can only demonstrate C-Reactive Protein as adding value after multivariate analysis to the established panels of tests and clinical data (the Leibovich score) used in renal cancer practice. More importantly, however, we have demonstrated how to establish a streamlined approach to new biomarker validation. The duration of follow up was a limitation of the cohorts but they substantiated several existing findings and identified biomarkers which may be taken forward for clinical utility studies.

The ELUCIDATE RCT workstream delivered the design, conduct and analysis of a trial which registered 1303 patients with chronic liver disease and randomised 878 patients out of a target of 1000. The trial started late and recruited, initially, slowly. However, the trial team identified and opened additional centres, clinicians recruited patients energetically in most
centres and new modelling techniques and data collection approaches were introduced by the team so that the trial ultimately recruited an adequate number of patients with good statistical power to answer the key clinical questions. Analysis showed that within the trial, the use of the ELF monitoring strategy altered the patient process of care and led to the introduction of tests which will identify patients who should benefit from the early introduction of interventions to manage serious complications and improve outcomes. ELUCIDATE was an “exemplar” trial which has demonstrated the challenges of evaluating biomarker strategies in “end-to-end” RCTs which enter patients into randomisations between a new monitoring strategy or conventional care, and then follow them through to ultimate end-points including survival. Its lessons will inform future study design.

There were significant interactions between the three workstreams. Workstream 1 gave the programme investigators a clear insight into the historical, methodological and study design challenges within this field and the scope of previous contributions. Innovations in study design, simulation strategies and the applications of health economic methods to evaluating monitoring tests were developed. These informed the revision of study design for the ELUCIDATE RCT, provided innovative approaches to power calculations based on pre-existing published cohorts and the early trial data. The work in Workstream 2 showed clearly the importance of rigorous assay evaluation. This informed the development of the ELUCIDATE RCT and in particular the work to re-evaluate the performance of the ELF test in the context of intra-laboratory variation and inter-laboratory variation. The challenges of delivering the ELUCIDATE RCT have informed our recommendations for future methodological approaches. The interactions between workstreams bring out the advantages of conducting the development of the clinical cohorts and of the RCT within the context of a programme, which had a strong multidisciplinary team of methodologists, clinical biochemists, trialists and clinicians. However, incorporating all three workstreams in a single programme also meant that the work of the programme as a whole was potentially prone to delays. For instance, delays in the recruitment of cohorts and in the set up and recruitment into the ELUCIDATE RCT have limited our ability to feed back substantial bodies of data, and outcome evaluation, from the RCT to the methodology workstream.

The findings in the three workstreams were used to synthesise a strategy and framework for future biomarker evaluation by the investigators with a defined pipeline and innovative contributions in study design, health economics and health informatics which became the
basis of a successful application to become one of four NIHR Diagnostic Evidence Cooperatives in England.

Word count: 2398
Plain English summary

Protein biomarkers are substances that can be measured, often in body fluids to provide information about a patient and their illness. Measuring biomarkers in blood or urine is simple, safe and may help diagnose disease, its severity and help choose treatment. New research is discovering more biomarkers but there is no quick way to decide how useful they are.

Our research aimed at methods to assess the clinical usefulness of biomarkers as quickly and efficiently as possible by:

1. the identification of the best research methods for monitoring disease or treatment with biomarkers. We showed the literature is modest in scale and of variable quality. We made recommendations for improvements.

2. the creation of a sample "banking" system for collecting and storing patient samples and relevant clinical data from large numbers of patients, for renal cell carcinoma (RCC) and renal transplantation (RT). Biomarkers were identified, analysed and the system used to show their value.

3. conducting a trial of the "Enhanced liver fibrosis" test, in 878 patients. We showed that it alters patient care but longer follow up is needed to show if this improves long term outcomes.

Our experience is part of the basis of a new framework for evaluating diagnostic tests, in four centres in England.

Word count: 208
Chapter 1 - Introducing New Biomarkers for Renal and Hepatic Diseases into Healthcare Systems
Background

Protein biomarkers in body fluids are now being regularly identified using new techniques and are associated with the presence and activity of diseases and treatment benefits/toxicities. They are accessible, measurable in real time and inexpensive. However their potential benefit to the NHS is not being realised due to the absence of a defined pathway linking biomarker research to health services research. This Programme aimed to establish a process for stringent evaluation of promising biomarkers encompassing methodological developments, clinical evaluation and an RCT to enable assessment of impact on clinical outcome, the process of care, resource use and Service configurations.

Our specific objectives were to:
1) evaluate and develop methodology for the optimal use and benefits to patients and the NHS of biomarkers for disease monitoring
2) establish a sample and clinical data bank together with a robust system for evaluation of promising markers, to facilitate their rapid assessment prior to large-scale trialling in the NHS
3) conduct a randomised clinical trial (RCT) of an established panel of biomarkers of potential value in chronic liver disease to diagnose cirrhosis at an early stage when interventions may reduce dangerous complications and to determine patient and NHS benefits.

Biomarkers have major potential benefits for patients and the NHS, particularly in contributing to "personalised” and/or “stratified” medicine and improved safety. They may supplement or replace invasive procedures or imaging tests for:

- accurate and early diagnosis
- measurement of the activity and extent of disease
- indication of prognosis
- selection and prediction of optimal treatments
- monitoring for treatment response/toxicity or disease progression

Additionally biomarker information may inform patient counselling on lifestyle issues (eg alcohol avoidance or diet) and motivate patients towards healthy alternatives. Overall better biomarkers should lead to improvements in outcomes and more efficient, cost-effective and evidence-based use of NHS resources.
With ongoing technological developments, particularly in proteomics, the rate of identification of potential new biomarkers may be expected to increase. Various stages in the “biomarker pipeline” have been defined, but the translational work needed to progress through these - involving technology transfer, methodological considerations and aspirations of different stakeholders - present challenges. The lack of a clear evaluative infrastructure means that the route from stringent evaluation and/or validation to clinical implementation and then to evaluation of impact on outcomes and healthcare is not yet established, representing a major threat to achieving full patient benefit.\textsuperscript{1-9} As an indication, the number of new protein markers approved by the FDA has gradually declined with only 10 in the period from 1994 to 2002, only 2 of which were in 1998-2002.\textsuperscript{10} Encouragingly, the need for national strategies for the rapid evaluation and introduction of new biomarker tests is now better appreciated, for example by the NIH in the USA (www.nihroadmap.nih.gov/) and by the Royal College of Pathologists in the UK.\textsuperscript{11}

A framework is therefore required for the pipeline to justify and guide introduction of biomarkers - including specification and establishment of the infrastructure to acquire such evidence, appropriate assessment of test results, identification of whether RCTs are required and means of deciding when and how biomarker development and introduction should be accelerated.

**Selected Diseases**

Chronic liver disease (CLD) and renal diseases provided ideal subjects with which to work up such a framework. Liver and renal diseases generate huge and increasing burdens on patients and the NHS. The care of patients would be transformed with improved outcomes, more appropriate use of complex and expensive therapies and avoidance of expensive and invasive investigations if biomarkers of real healthcare value could be found.

**Chronic Liver Disease (CLD)**

CLD is the fifth most common cause of death in the UK and the second commonest cause of death in men aged 35 – 54, usually associated with alcoholic liver disease, fatty liver disease or hepatitis C infection, any of which may lead to fibrosis, cirrhosis and hepatocellular carcinoma (HCC).\textsuperscript{12-14} Life-threatening complications include variceal bleeding, recurrent
ascites, hepatic encephalopathy. Once cirrhosis has developed HCC arises in 1-6% of patients per annum.\textsuperscript{15, 16} Social issues such as inability to work constitute a huge healthcare and financial burden. Evidence shows that earlier cirrhosis detection results in better survival and reduced morbidity.

**Acute Renal Transplant**

Currently in the UK there are 23,000 patients with functioning transplants (see Chapter 10). Annually, almost 3,080 renal transplants are performed. Transplantation represents the best therapy to improve survival and quality of life and cost effectiveness, saving the NHS >£490M/yr compared with dialysis.\textsuperscript{17} Acute Rejection (AR, 25% patients) and delayed graft function (DGF, 40%) significantly reduce short and long term graft survival. Early diagnosis of AR/DGF is critical for optimal treatment. Serum creatinine is slow to respond and insensitive. Currently, renal biopsy is required for a definite diagnosis which is invasive and may not be available immediately.\textsuperscript{18-24} Biomarkers allowing the earlier diagnosis of DGF and AR and discrimination of subgroups, a strategic priority of the American Society of Nephrology, would allow earlier and more appropriate therapeutic intervention.\textsuperscript{25}

**Renal Cell Carcinoma (RCC)**

Accounting for ~3% of adult malignancies, RCC is increasing with approximately a third of a million new cases each year worldwide and 10,000 in the UK and more than 140,000 deaths worldwide (see Chapter 10).\textsuperscript{26, 27} Locally advanced or metastatic disease affects over 50% of patients, for which treatments are limited. New drugs have improved response rates and relapse-free survival but they are expensive and markers for diagnosis, prognosis and selection of expensive therapy are desperately needed.\textsuperscript{28-31}

**Protein Biomarkers**

Protein biomarkers in body fluids have substantial potential to improve quality of health care. There are active pipelines identifying them in both commercial and non-commercial sectors, but robust methodological approaches and well organised rapid clinical and health service evaluation is still limited.
In a clinical setting, the value of a protein biomarker depends on test performance and its relation to health improvements. Methods of reporting, and hence judging, test performance are well developed for biomarkers and other kinds of measures when these are used in a prognostic role, and in a diagnostic role.\textsuperscript{32, 33} The methodology for evaluating biomarker test performance when used in an individual patient monitoring role is however poorly developed.

After the initial discovery and preliminary evaluation of a protein biomarker, robust evaluation of its clinical characteristics (sensitivity, specificity, receiver-operator characteristic (ROC), etc) is often done slowly and in limited sample numbers. We set out to establish a robust system with samples and clinical data in adequate numbers to rapidly evaluate markers which may be useful in the NHS and select those that justify formal evaluation.

Chronic liver disease can frequently progress to cirrhosis and to life-threatening complication. Early diagnosis of cirrhosis with appropriate management can reduce these complications and a panel of protein biomarkers which directly evaluate fibrosis has excellent clinical characteristics (Enhanced Liver Fibrosis test). We proposed and have conducted a randomised controlled trial to establish whether its use can substantially improve patient outcomes and health care provision. Previous work on ELF show it has proven clinical associations with cirrhosis and its complications justifying its evaluation to see if it can monitor patients and allow diagnosis of cirrhosis at a time when interventions will reduce the morbidity and mortality associated with its complications.\textsuperscript{34-37}

**Monitoring Studies**

There is considerable research into the use of biomarkers for prognosis (including prediction of response to treatment). By contrast there is relatively small literature on the use of markers for monitoring.\textsuperscript{38} Methodological work has not yet been conducted on the design and interpretation of studies with repeated measurements of biomarkers, and we used Value of Information analysis in this context.

Monitoring may be undertaken for various purposes. In chronic diseases, it may assess whether interventions are keeping the disease and symptoms in control; assess the rate of
progression of disease (eg the ELF test in our RCT); to detect recurrence or evaluate efficacy of treatments. Monitoring may avoid adverse effects. There is considerable research into the use of biomarkers for prognosis (including prediction of response to treatment). By contrast literature on markers for monitoring is relatively sparse. This requires a strategy for the frequency of testing and rules for clinical actions (including retesting). Although each test result can be judged on its own there is potential also to learn from the change since the previous test or the rate of change over time, or to devise and calibrate a model of change over time. Rules may be devised using suitable cutpoints, changes between values, or values of fitted model parameters, or confidence intervals thereof. Also, it will be important to understand natural variability due to measurement error, prognostic factor related trends (eg in age), and the risks of incorrect decisions.

Few biomarkers have been studied in this context. Examples in cancer include CA-125 for monitoring patients with ovarian cancer for recurrence, PSA for monitoring men at risk of prostate cancer and for monitoring prostate cancer for recurrence. Serum creatinine measurements monitor patients with chronic kidney disease and transplants. An important question for monitoring by repeated measurement is whether prognosis relates to the actual marker level or change over time or more complex models.

Two different questions need to be considered, relating first to the interpretation of specific marker values and sequences of such values, and second to the implementation of such information, for example to determine how frequently measurements should be taken. There is a need for large, high quality data sets with repeated measurements to inform the development of decision rules for monitoring.

**Choice of the ELF test for the ELUCIDATE RCT**

Within our range of liver and renal diseases, we identified the Enhanced Liver Fibrosis (ELF) test as having ideal characteristics for our RCT because:

- the ELF test has convincing clinical evaluation data showing an association with cirrhosis (eg ROC >0.8) but the clinical and health service benefits from monitoring chronic liver disease for cirrhosis is untested;
- the burden of CLD and cirrhosis on patients and the NHS is huge and increasing;
• early diagnosis of cirrhosis allows effective surveillance and interventions to improve clinical outcomes and care;
• a simple blood test could radically improve and provide cost effective care for CLD patients;
• this model will provide an excellent prototype for health service biomarker research and vital datasets for our methodologists.

In the vast majority of cases liver fibrosis is asymptomatic and cirrhosis develops insidiously with non-specific symptoms so that opportunities for disease modification or cure are missed. Standard biochemical tests of liver function are not specific or sensitive. Liver biopsy is hazardous, inaccurate, subject to sampling error and variation in interpretation. Imaging has a major role in the detection and assessment of liver fibrosis. However all imaging modalities including ultrasound, elastography, cross-section imaging with X-rays or magnetic resonance require access to expensive technology and skilled operators.

We are in the relatively fortunate position, with cirrhotic diseases of the liver, of having a number of treatments (such as beta blocker therapy, or surgery for low-volume hepatocellular cancer) that are known to be effective at reducing complications of this disorder, if the cirrhotic condition is detected early enough. The ELF test seeks to identify a ‘pool’ of patients with a slowly progressing disorder that we can treat prophylactically, and reduce severe complications.

Evidence shows that early detection of varices and treatment with prophylactic use of beta blockers to reduce portal hypertension or band ligation reduces morbidity and increases survival and respected guidelines recommend surveillance because of its benefits and health economic justification. Similarly early detection of ascites and treatment has been shown to reduce the morbidity associated with bacterial peritonitis from 17% to 2%. The case for surveillance and early detection of hepatocellular cancer is more contentious with some randomised controlled trials showing evidence of benefit while other show none. International guidelines now advocate surveillance. Retrospective analyses have identified criteria, essentially small tumours, associated with better outcomes for HCC resection and liver transplantation but many patients are diagnosed after the growth of their tumours has ruled them out for curative resection or transplantation.
Studies and systematic reviews have demonstrated that single direct markers are less accurate than panels of markers in the detection of liver fibrosis. One such panel of direct markers is the Enhanced Liver Fibrosis (ELF) test, the only CE marked (EU Regulatory Approval) test for liver fibrosis measuring constituents of liver matrix (Hyaluronic acid and Procollagen III amino-terminal peptide) and an molecule critical to the regulation of matrix re-modelling (Tissue Inhibitor of Matrix metalloproteinase 1) using sensitive automated ELISA assays designed and manufactured specifically for this purpose. The 3 individual biomarkers were selected as optimal from among 20 candidates. The results of the individual assays are combined in an algorithm derived and validated in >1000 cases of liver fibrosis to generate a score that correlates with the severity of liver fibrosis on liver biopsy. ELF scores have been shown to be highly predictive of clinical outcomes, including variceal bleeding, ascites, HCC and mortality. Subsequent validation studies in hepatitis C, fatty liver disease, HIV-HCV co-infection, and primary biliary cirrhosis have confirmed the performance of the test. Although performance is best in the detection of advanced fibrosis and cirrhosis, the test can also detect mild and moderate degrees of fibrosis accurately with area under the curve (AUC) ROCs of 0.83 for 0-3 vs 4-6 and 0.86 for 0-4 vs 5-6 Ishak stages respectively. The ELF test is excellent in the detection of advanced fibrosis/cirrhosis in a range of chronic liver diseases and is thus well suited to be used for screening in populations at risk for cirrhosis. The ELF test has been developed by Siemens Medical Solutions (formerly Bayer Healthcare) in conjunction with the University of Southampton and iQur Limited.

The overall shape of our programme is given in Figure 1.
Figure 1. The Programme

**Workstream 1**
Methodology
- Systematic reviews of monitoring trials and the literature.
- Simulations of monitoring strategies and their optimisation.
- The patient perspective.

**Workstream 2**
Renal diseases – Biomarker evaluation
- Sample Banking & Clinical Data Collection
- Selection of Biomarkers
- Assay Development and Validation
- Clinical Evaluation

**Workstream 3**
ELUCIDATE Clinical Trial
The impact of ELF testing on the outcomes of patients with chronic liver diseases

**Deliverable**
The results of ELF testing:
Diagnosis of cirrhosis, changed process of care and potentially improved survival and economic outcomes

**Deliverable**
Future clinical trial designs and evidence-based use for patient benefit

**Deliverable**
Patient Perspective

**Deliverable**
Simulation models and their value

**Deliverable**
Biomarkers with potential clinical utility

**Deliverable**
Academic-NHS-Industry Partnerships

**Deliverable**
High quality biobanks enabling rapid future biomarker evaluation

**Deliverable**
Assay Reports, Position Papers, Publications

Influencing study designs.
Data sharing.
Chapter 2 - Introduction to the Methodological Workstream

(WS1)
In this chapter we introduce the methodological workstream on monitoring tests of the programme and indicate the work described in chapters three to nine.

Monitoring is the repeated application of a test, or set of tests, over time to assist in the management of a disease or condition. It is a fundamental element of patient care, comprising much of the clinical workload. Often thought of in terms of treatment titration and maintenance, where the aim is to keep a marker within predefined limits until treatment can be discontinued or an alternative treatment is required, monitoring is also used to manage individuals with a disease or condition that is likely to progress or recur at some time in the future, allowing timely decisions to be made regarding patient management. Patients are usually asymptomatic or mildly symptomatic but not yet receiving treatment, or they may experience symptoms of a disease that puts them at risk of developing other conditions. Monitoring often involves a general clinical assessment and physical examination of patients but is likely to include the application of specific tests, from tools assessing functional or psychological status, to blood or urine tests, physiological measurements such as blood pressure, imaging tests or more invasive assessments such as colonoscopy or biopsies. Although subsequent sections of this report focus on protein biomarkers in particular, the methodological considerations that underlie the development and evaluation of monitoring strategies are relatively universal and can be applied regardless of type of test.

Historically, methodological research around test evaluation has lagged behind that of intervention research. Over the last 10-15 years, however, considerable research effort has focused on identifying optimal methods for establishing diagnostic test accuracy and, more recently, on the evaluation of diagnostic tests and strategies in terms of their impact on patient management and outcomes. Monitoring tests, particularly in terms of treatment titration, are now beginning to receive attention in the literature. Our particular interest is in monitoring patients for disease progression or recurrence.

The development and evaluation of tests for monitoring purposes bears many resemblances to that for diagnostic tests but with a few key differences. Firstly, unlike tests for diagnosis, tests for monitoring often do not aim to detect present disease but some marker of preclinical or early stage disease that precedes the development of clinical disease progression or recurrence. Vitally, this latent stage of disease must be of a reasonable duration as to make the repeated application of a monitoring test worthwhile - too short and
recurrence or progression may be missed regardless of the monitoring schedule adopted, too long and frequent monitoring may not be of clinical benefit. Secondly, while diagnostic tests are often applied once, perhaps with a repeated application to confirm diagnosis, by the very nature of monitoring, tests are applied repeatedly over an indefinite period of time and according to some predetermined schedule. With patients’ true disease status often not established until clinical disease progression occurs, a cross-sectional evaluation of the diagnostic accuracy of a test can be impossible to establish and more longitudinal measures of capturing how well a test predicts clinical outcome have been suggested. The same principle of detecting true disease (recurrence and progression) while limiting false positive results applies, with the further consideration that the test should be able to differentiate long term change in disease status from short term measurement variability. The further in advance of the clinical event of interest the marker occurs, the less predictive it may be and the greater the potential influence from measurement variability on false positive and false negative results. More complex decision rules to determine the point at which some clinical action should be taken may also be relevant for monitoring. Although each test result can be judged on its own, as in a diagnostic context, serial values over time may provide valuable information. Rules may be devised using individual thresholds, the change in measurement since the previous test, the rate of change in measurement values, values of fitted model parameters, or confidence intervals thereof. Finally, whereas a diagnostic test may be applied to assist in the ruling in or out of a number of differential diagnoses and any number of therapeutic approaches may be indicated, a positive monitoring test often only increases the probability of a particular future clinical event, and furthermore, a series of further investigations may be initiated before a particular treatment approach is applied.

Many of these considerations are particularly true of protein biomarkers, where changes in biomarker levels may occur before any clinical symptoms or signs become apparent. Measuring biomarkers in blood or urine is relatively simple and safe for patients, making them an attractive alternative or complement to more complex, invasive or expensive tests. However, initially promising results at the biomarker identification stage do not necessarily translate into clinical benefit in practice.

Our aim was to identify and describe general methodological considerations for the development and evaluation of testing strategies to monitor for disease progression or recurrence, reviewing current best practice and exploring methodological innovations.
The chapters to follow all address aspects of the evaluation of monitoring biomarkers, but they do so from different starting points, depending on the amount and nature of the relevant literature being used as the starting point for the work. Each chapter therefore begins with a brief introduction presenting the background elements most relevant to explaining the work to be reported. This approach has created a degree of repetition, but does enable the chapters to be read separately without the need for extensive cross referencing to material presented elsewhere in the report. The chapters do also show some stylistic differences, reflecting in part the nature of the work reported, but also the discipline background of their lead authors.

Chapter 3 reviews monitoring strategies recommended in available clinical guidelines, with specific reference to the use of prostate specific antigen (PSA) for the detection of recurrent prostate cancer. PSA was chosen because of the extensive literature surrounding it, and in the event, there was so much material to consider, and so many generalisable considerations were emerging, that we decided to focus entirely on PSA rather than pursue our original plan to add a number of “mini” case studies from other clinical areas. The particular focus was on: the degree of consistency between guidelines, the explicit consideration of factors important for specifying a monitoring strategy, and the use of supporting evidence to justify any recommendations. Ultimately, monitoring strategies are employed in order to allow timely decisions to be made regarding patient management, thereby improving patient outcomes, for example, through earlier initiation of treatment to prevent or delay some clinical outcome. The randomised controlled trial (RCT) design is considered to be the gold standard approach to evaluation of patient benefit from therapeutic interventions, however testing strategies are complex interventions with many components, their evaluation presenting considerable challenges.

Chapter 4 reports a methodological review of RCTs of monitoring strategies to consider how successfully the design has been used to identify patient benefit from monitoring.

The methodological research is reviewed in Chapter 5. Although it is generally acknowledged that methodological work around monitoring tests has been lacking, there are areas of research that could be used or adapted for the development and evaluation of monitoring strategies for monitoring for disease progression or recurrence.
**Chapter 6** focuses on the wider impact of monitoring on patients. Ferrante di Ruffano and colleagues have produced a framework to assist those designing and evaluating trials of diagnostic tests to understand the ways in which changes to testing strategies can affect patient outcomes. We have adapted this framework to tests for monitoring, in the light of our review of randomised trials. In this chapter we consider the potential for benefit and harm from monitoring in broad terms, before considering the ways in which patient outcomes can be mediated by particular aspects of the monitoring care pathway, noting the similarities and differences between diagnostic and monitoring tests.

**Chapter 7** considers how simulation modelling can be used to identify optimal monitoring strategies, prior to or alongside a randomised trial. Simulation offers a powerful tool to design and evaluate monitoring rules. However, such models are data intensive, requiring many pieces of information to allow their construction. For many tests and diseases, limitations in the available data may affect the reliability of the final model. We explored how information obtained during an on-going study (the ELUCIDATE trial) could be incorporated into a simulation model of the Enhanced Liver Fibrosis (ELF) biomarker panel for monitoring patients with known liver fibrosis. The aim was to optimise monitoring rules to allow earlier detection of liver cirrhosis, and to consider whether any resulting adaptations to the design of the ongoing study that were suggested by the model could be implemented without compromising the validity or clinical value of the trial.

**Chapter 8** takes a Health Economic approach, modelling a method of optimizing a monitoring test to meet a cost-effectiveness target and exploring the feasibility of using Value of Information analysis to inform biomarker research and development.

The final chapter in this section, **Chapter 9**, brings together our findings and reports on a consultation with patient and public representatives, considering what we know from current practice in monitoring for disease progression and recurrence, what we have learned in terms of understanding the monitoring process and how this should inform future development and evaluation of monitoring strategies.
Chapter 3 - How is evidence being used to make recommendations about monitoring: the example of prostate specific antigen (PSA)

The work described in this chapter was published:


Table and figures are reproduced with permission
Introduction

Monitoring involves the scheduled, repeated use of a test or tests in an individual over time to make decisions about the management of a disease or condition. It is a central activity in the management of patients, taking up a considerable part of the clinical workload and associated cost. In contrast, the volume of published literature on the evaluation and use of tests for monitoring purposes is relatively sparse.

Mant and others have provided a framework for developing and evaluating a monitoring strategy with four main steps: deciding whether or not to monitor, choosing a test, specifying and assessing the monitoring strategy to be used, followed by an implementation phase. Underlying this is the key concept that the ‘signal’ from the test, reflecting the status of the underlying condition, should be greater than the surrounding ‘noise’, or measurement variability, that may affect test interpretation. If the ‘noise’ around a test measurement is too high in relation to the signal, one’s certainty in a given test result will be considerably reduced.

The repeated measurement of prostate specific antigen in men who have undergone primary treatment of prostate cancer is an apparently successful example of a rule-based monitoring strategy. The behaviour of prostate specific antigen following radical treatment varies, but in general, recurrence of disease is associated with the presence of (following radical prostatectomy), or some rise in (following radical radiotherapy), prostate specific antigen. When a predefined level of prostate specific antigen is reached, biochemical failure is said to have occurred. The usefulness of prostate specific antigen as a monitoring test is based on the assumption that biochemical failure predates clinical failure within some clinically meaningful time frame. The decision to initiate treatment for recurrence, however, will depend on multiple factors rather than on a single value alone.

We undertook a review of clinical guidelines’ recommendations for monitoring with prostate specific antigen for the detection of recurrent prostate cancer to determine the extent to which they take into account key factors that should inform rule-based strategies for monitoring. Our particular focus was on: the degree of consistency between guidelines, the explicit consideration of factors important for specifying a monitoring strategy, and the use of supporting evidence to justify any recommendations.
Methods

Inclusion criteria

Guidelines that considered the use of prostate specific antigen as a test for monitoring patients treated with either radical prostatectomy or radical radiotherapy for localised prostate cancer were eligible. Guidelines that considered only screening or treatment were excluded. Guideline recommendations regarding prostate specific antigen measurement following other potentially curative treatments or as part of active surveillance were not considered.

Literature searches

Medline was searched from 1999 to July 2009 using the MeSH terms (“Prostatic Neoplasms” OR “Prostate-Specific Antigen”) AND “Practice Guideline”, limited to English language publications. The National Library of Guidelines, the Trip database, and the Cochrane Library were also accessed and reference lists of retrieved papers checked. Titles and abstracts of retrieved records were assessed for inclusion by two authors independently (JD and JJD) with discrepancies resolved by consensus.

Data extraction

Recommendations or statements relating to the use of prostate specific antigen following treatment with curative intent were extracted and references to any supporting evidence noted. Guideline methods were assessed using the Appraisal of Guidelines for Research & Evaluation framework which contains 23 key items organised into six domains. We applied only the seven items included in the ‘Rigour of development’ domain (Box 1 & 2). We replaced the fourth item in this domain with one relevant to using tests for monitoring as opposed to consideration of benefits and harms of interventions.
Box 1 Criteria used to assess rigour of guideline development (Reproduced with permission from Dinnes et al, 2012.\textsuperscript{72})

<table>
<thead>
<tr>
<th><strong>Systematic search methods used</strong></th>
<th>Details of the strategy used to search for evidence should be provided including search terms used, sources consulted and dates of the literature covered.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selection criteria clearly described</strong></td>
<td>Criteria for including / excluding evidence identified by the search should be provided. These criteria should be explicitly described and reasons for including and excluding evidence should be clearly stated.</td>
</tr>
<tr>
<td><strong>Formulation of recommendations clearly described</strong></td>
<td>There should be a description of the methods used to formulate the recommendations and how final decisions were arrived at. Areas of disagreement and methods of resolving them should be specified.</td>
</tr>
<tr>
<td><strong>Consider relevant issues for monitoring in recommendations†</strong></td>
<td>Variability in measurements/need for re-testing, rationale presented for interval and prostate specific antigen threshold, and acknowledgement of the uncertainties in the natural history of prostate specific antigen following radical treatment.</td>
</tr>
<tr>
<td><strong>Explicit link with supporting evidence</strong></td>
<td>There should be an explicit link between the recommendations and the evidence on which they are based. Each recommendation should be linked with a list of references on which it is based.</td>
</tr>
<tr>
<td><strong>Pre-publication external review</strong></td>
<td>A guideline should be reviewed externally before it is published. A description of the methodology used to conduct the external review should be presented, which may include a list of the reviewers and their affiliation.</td>
</tr>
<tr>
<td><strong>Update procedure described</strong></td>
<td>Guidelines need to reflect current research. There should be a clear statement about the procedure for updating the guideline.</td>
</tr>
</tbody>
</table>


† Original criterion related to treatment outcomes, i.e. “The guideline should consider health benefits, side effects, and risks of the recommendations.”
Box 2 Criteria used to assess rigour of guideline development*, with details and examples (Reproduced with permission from Dinnes et al, 2012.72)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematic search methods used</td>
<td>Details of the strategy used to search for evidence should be provided including search terms used, sources consulted and dates of the literature covered. Sources may include electronic databases (e.g. MEDLINE, EMBASE, CINAHL), databases of systematic reviews (e.g. the Cochrane Library, DARE), handsearching journals, reviewing conference proceedings and other guidelines (e.g. the US National Guideline Clearinghouse, the German Guidelines Clearinghouse). Further point for judgement on completeness of search</td>
</tr>
<tr>
<td>Selection criteria clearly described</td>
<td>Criteria for including / excluding evidence identified by the search should be provided. These criteria should be explicitly described and reasons for including and excluding evidence should be clearly stated. For example, guideline authors may decide to only include evidence from randomised clinical trials and to exclude articles not written in English. Further point for judgement on application of criteria</td>
</tr>
<tr>
<td>Formulation of recommendations clearly described</td>
<td>There should be a description of the methods used to formulate the recommendations and how final decisions were arrived at. Methods include for example, a voting system, formal consensus techniques (e.g. Delphi, Glaser techniques). Areas of disagreement and methods of resolving them should be specified.</td>
</tr>
<tr>
<td>Consider relevant issues for monitoring in recommendations†</td>
<td>The guideline should consider factors relevant to test for monitoring, i.e. variability in measurements/need for repeat testing, rationale presented for interval frequency and PSA threshold, and acknowledge the uncertainties in the natural history of PSA following radical Rx.</td>
</tr>
<tr>
<td>Explicit link with supporting evidence</td>
<td>There should be an explicit link between the recommendations and the evidence on which they are based. Each recommendation should be linked with a list of references on which it is based.</td>
</tr>
<tr>
<td>Pre-publication external review</td>
<td>A guideline should be reviewed externally before it is published. Reviewers should not have been involved in the development group and should include some experts in the clinical area and some methodological experts. Patients’ representatives may also be included. A description of the methodology used to conduct the external review should be presented, which may include a list of the reviewers and their affiliation.</td>
</tr>
<tr>
<td>Update procedure described</td>
<td>Guidelines need to reflect current research. There should be a clear statement about the procedure for updating the guideline. For example, a timescale has been given, or a standing panel receives regularly updated literature searches and makes changes as required.</td>
</tr>
</tbody>
</table>


†Original criterion related to treatment outcomes, i.e. “The guideline should consider health benefits, side effects, and risks of the recommendations. For example, a guideline on the management of breast cancer may include a discussion on the overall effects on various final outcomes. These may include: survival, quality of life, adverse effects, and symptom management or a discussion comparing one treatment option to another. There should be evidence that these issues have been addressed.”
A generous approach to scoring items was taken. For example, if a systematic search was reported to have been carried out but was not reported in detail, the guideline would score three out of a possible four points. If a discussion of evidence was provided that appeared to relate to a recommended monitoring schedule, an explicit link with evidence was judged to have been provided without closer examination of the actual evidence cited. We did not make a judgement as to the acceptability of any rationale presented for test frequency or threshold but indicated whether a rationale was presented or not. A maximum score of 4 points was attached to each of the seven items for a maximum score of 28 points.

Synthesis

A narrative synthesis was undertaken.

Figure 2. Rigour of guideline development (Reproduced with permission from Dinnes et al, 2012.)

UK PCWG – UK Prostate Cancer Working Group; Aus CN – Australian Cancer Network; AUA – American Urological Association; DUA – Dutch Urological Association; NCI PDQ – National Cancer Institute – Physician Data Query (US); NICE – National Institute for Health and Clinical Excellence (UK); EAU – European Association of Urology; NCCN – National Comprehensive Cancer Network (US)

Results

Guidelines (n=7) or best practice statements (n=2) from 9 organisations were identified. Four were North American in origin, four from Europe and one from Australia. Nearly all of the
guidelines scored poorly on the framework criteria with between 9 and 16 out of a possible 28 (Figure 2). The National Institute for Health and Clinical Excellence guideline scored considerably higher with 22 points.\textsuperscript{78} The highest scoring item overall was the use of systematic searches, reported in most guidelines, even if it was often not described in any detail. Methods for recommendation formulation were described in only three. Only one guideline fully considered relevant issues for monitoring tests and was the only one to consistently provide clear links between its recommendations and the underlying evidence base. It also reported its methods in more detail than most of the other guidelines in the sample.\textsuperscript{78}

Table 1 shows the lack of consistency in guideline recommendations regarding the frequency of follow-up assessments and thresholds; there does not appear to be any clear pattern in recommendations over time.

Eight of the 9 guidelines acknowledged that prostate specific antigen levels may be affected by technical or biological variability but in most cases this was presented in the introductory sections of the guidelines, with only one tempering its recommendations with reference to the fact that a single prostate specific antigen measurement may be unreliable (recommending re-testing within two months).\textsuperscript{73-79, 81} Three acknowledged the potential impact from technical variation, recommending that the same assay be used at each measurement.\textsuperscript{77, 78, 81} Four guidelines made some attempt to justify the interval between tests and three discussed relevant issues affecting the choice of threshold.\textsuperscript{73, 76-78, 80} A further three stated that it was not possible to provide a recommendation on the most appropriate biochemical failure definition.\textsuperscript{74, 75, 77} Only three of the nine guidelines commented on the difficulty of using prostate specific antigen as a monitoring tool due to the uncertainties in its behaviour following radical treatment, with two clearly recognising that not all men with biochemical failure go on to experience clinical failure, such that evidence of the former alone may not be sufficient to alter treatment.\textsuperscript{75, 77, 78}

Many recommendations on frequency or threshold were made with no or with few supporting citations (Table 1). Only one guideline cited a primary study in support of its recommended monitoring intervals and only four of the nine indicated the level of evidence supporting their recommendations.\textsuperscript{76, 78-80} The levels of evidence suggested ranged from ‘Guideline Development Group consensus’ only to ‘well conducted clinical studies’ (Table 1),
suggesting that different guideline groups had varying views on the quality of the evidence available.
Table 1 Table of guideline statements or recommendations and indication of supporting evidence cited, if any (Reproduced with permission from Dinnes et al, 2012.72)

<table>
<thead>
<tr>
<th>Guideline Year</th>
<th>UK PCWG 199975</th>
<th>Aus CN 200274</th>
<th>AUA 200775</th>
<th>DUA 200776</th>
<th>NCI PDQ 200877</th>
<th>NICE 200878</th>
<th>AUA 200981</th>
<th>EAU 200982</th>
<th>NCCN 200983</th>
<th>Number of guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Frequency of follow-up visits after radical treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarterly for 1*or 2† years, then 6 monthly or annually</td>
<td>√</td>
<td>ASTRO⁴</td>
<td>-</td>
<td>-</td>
<td>3 prim studies (level 3)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>None cited (Level B)</td>
</tr>
<tr>
<td>Every 6 months (for 2* or 5† years) then annual</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>None cited</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>No direct evidence (Consensus)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2.1 Threshold for 'intervention' following prostatectomy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any detectable PSA</td>
<td>√</td>
<td>None cited</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3 prim studies</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PSA &gt;0.2 ng/mL</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1 prim study, 1 review (level 4)</td>
<td>-</td>
<td>-</td>
<td>3 prim studies, 2 reviews</td>
<td>√</td>
<td>1 prim studies, 1 review</td>
<td>-</td>
</tr>
<tr>
<td>No definite threshold recommended</td>
<td>-</td>
<td>√</td>
<td>None cited</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2.2 Threshold for 'intervention' following radiotherapy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three consecutive increases in PSA (ASTRO⁴ 1997 definition⁵)</td>
<td>√</td>
<td>ASTRO⁴</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>3 prim studies (level 4)</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>2 prim studies (Phoenix⁴, ASTRO⁴, Phoenix (Level B)</td>
</tr>
<tr>
<td>PSA nadir + 2 ng/mL (Phoenix 2005 definition⁶)⁷</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>2 prim studies, Phoenix⁴</td>
<td>-</td>
<td>2 prim studies, Phoenix⁴, ASTRO⁴, Phoenix (Level B)</td>
</tr>
<tr>
<td>PSA nadir + 4 ng/mL</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>3 prim studies, Phoenix⁴, 1 review</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>No specific recommendation</td>
<td>-</td>
<td>√</td>
<td>2 prim studies, 1 review</td>
<td>None cited</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>2 prim studies, ASTRO⁴, Phoenix⁴</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
3. Sources of PSA variability acknowledged and/or remedial action recommended

<table>
<thead>
<tr>
<th>3. Sources of PSA variability acknowledged and/or remedial action recommended</th>
<th>8/9</th>
</tr>
</thead>
<tbody>
<tr>
<td>technical variability possible</td>
<td>√</td>
</tr>
<tr>
<td>biological variability possible</td>
<td>-</td>
</tr>
<tr>
<td>remedial action recommended</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4. Acknowledgement of uncertainties in natural history of PSA and PCa following primary treatment</th>
<th>3/9</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>1 review</td>
</tr>
</tbody>
</table>

√ indicates whether factors were considered anywhere within the guideline document (√), within the guideline recommendations (√√), or not acknowledged in document (-). An indication of the amount and type of supporting evidence (if any) cited by the guideline is also provided, along with the level of evidence accorded to the recommendation by the guideline development group in question.

a Guideline development groups: UK PCWG – UK Prostate Cancer Working Group; Aus CN – Australian Cancer Network; AUA – American Urological Association; DUA – Dutch Urological Association; NCI PDQ – National Cancer Institute – Physician Data Query (US) ; NICE – National Institute for Health and Clinical Excellence (UK); EAU – European Association of Urology; NCCN – National Comprehensive Cancer Network (US)

b Levels of evidence as reported in individual guideline documents: DUA: level 3 - at least 1 RCT, other comparative or non-comparative study, level 4 - expert opinion from, for example, working group members;76 NICE: consensus – Guideline Development Group consensus;78 EAU: Grade B - well-conducted clinical studies, but without RCTs; 79 NCCN: level 2a - lower-level evidence and uniform NCCN consensus.80

c Consensus threshold definitions: ASTRO – American Society for Therapeutic Radiology and Oncology 1997 consensus statement;82 Phoenix – 2005 revision of the ASTRO consensus statement83
Despite the general lack of citations in individual guideline documents, a wide range of papers were cited across guidelines. A total of 48 papers were cited (Tables 2-4); 29% (14/48) were reviews and the remainder were primary studies, almost exclusively retrospective in nature. Of the primary studies, we judged half to have studied the natural history of prostate specific antigen following treatment and eight (25%) to have evaluated the effect of different biochemical failure definitions on clinical outcomes. Only two primary studies examining measurement variability were cited.

**Table 2 Type of studies used to support guideline recommendations** (Reproduced with permission from Dinnes et al, 2012.72)

<table>
<thead>
<tr>
<th>Used to support guideline recommendations on:</th>
<th>Number of studies per group</th>
<th>Test frequency</th>
<th>Threshold (RP)</th>
<th>Threshold (RT)</th>
<th>Variability</th>
<th>Uncertainty in natural history</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASTRO consensus statements</td>
<td>3</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best practice statement</td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reviews</td>
<td>10</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Primary studies:</td>
<td>34</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Follow-up (acceptability of)</td>
<td>3</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up (optimal frequency)</td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natural history of PSA post-Rx</td>
<td>15</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Natural history of PSA w/out Rx</td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salvage RTX outcomes</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing BF definitions</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measurement variability</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of guidelines citing evidence</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

RP – radical prostatectomy; RTX – radical radiotherapy; ASTRO – American Society for Therapeutic Radiology and Oncology; Rx – treatment; BF – biochemical failure

Most studies were cited by only one or two of the guidelines, but a handful was cited three or more times (Table 3). Two consensus statements were amongst the most frequently cited, as was a review of biochemical failure definitions.70, 82, 83 The four primary studies had amongst the largest sample sizes of all cited primary studies: three evaluated the use of different biochemical failure definitions and one studied the natural history of disease progression in men with raised prostate specific antigen.84-87
Table 3 Most commonly (≥3 times) cited studies supporting guideline statements
(Reproduced with permission from Dinnes et al, 2012.72)

<table>
<thead>
<tr>
<th>Author (year)</th>
<th>Study design/aim (extracted from abstract)</th>
<th>Used to support statements on:</th>
<th>No. times cited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roach 2006</td>
<td>Reports second consensus conference to revise the ASTRO definition of BF</td>
<td>Threshold (RTX)</td>
<td>5</td>
</tr>
<tr>
<td>Pound 1999</td>
<td>Retrospective review of a large surgical series (n=1997) to examine the natural history of progression to distant metastases in men with raised PSA following surgery</td>
<td>Frequency, Threshold (RP), NH</td>
<td>5</td>
</tr>
<tr>
<td>Kuban 2006</td>
<td>Primary study of patients treated with radioisotopic implant as solitary treatment for localised PCa (n=2,693). Multiple PSA failure definitions were tested for their ability to predict clinical failure.</td>
<td>Threshold (RTX), NH</td>
<td>4</td>
</tr>
<tr>
<td>ASTRO 1997</td>
<td>Consensus statement providing guidelines for PSA following radiation therapy</td>
<td>Frequency, Threshold (RTX)</td>
<td>3</td>
</tr>
<tr>
<td>Cookson 2007</td>
<td>AUA review of the variability in published definitions of biochemical recurrence; recommends a standard definition in patients treated with radical prostatectomy</td>
<td>Threshold (RP), NH</td>
<td>3</td>
</tr>
<tr>
<td>Horwitz 2005</td>
<td>Determined the sensitivity and specificity of several BF definitions using pooled data on 4,839 patients treated with external beam RT alone</td>
<td>Threshold (RTX)</td>
<td>3</td>
</tr>
<tr>
<td>Stephenson 2006</td>
<td>Tested 10 definitions of BF on 3,125 patients who underwent RP, to identify the one that best explains metastatic progression</td>
<td>Threshold (RP)</td>
<td>3</td>
</tr>
</tbody>
</table>

RP – radical prostatectomy; RTX – radical radiotherapy; NH – natural history; ASTRO – American Society for Therapeutic Radiology and Oncology; AUA - American Urological Association; Rx – treatment; BF – biochemical failure
<table>
<thead>
<tr>
<th>Body</th>
<th>Brief description</th>
<th>Systematic search methods used</th>
<th>Subcategory criteria clearly described</th>
<th>Formulation of recommendations clearly described</th>
<th>Consider relevant issues for monitoring in recommendations</th>
<th>Explicit link with supporting evidence</th>
<th>Pre-publication external review</th>
<th>Update procedure described</th>
<th>Total</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUA 200775</td>
<td>Localised PCa management guideline</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>13</td>
<td>One database used, search poorly reported; inclusion criteria described but grounds for later exclusion of papers not clear; limited description of recommendation formulation; no basis for interval between measurements; evidence-based recommendations for threshold not possible; NH uncertainty acknowledged in recommendations, variability acknowledged but not in recommendations; no link between recommendations and evidence; external review carried out but not described; update recommended and to include only RCT evidence</td>
</tr>
<tr>
<td>AUA 200981</td>
<td>PSA best practice statement</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>No systematic search; inclusion criteria not described; some description of recommendation formulation; no interval between measurements recommended; consensus definition of threshold used; NH uncertainty not acknowledged; variability acknowledged but not in recommendations; some supporting evidence cited; peer review carried out but not described in detail; no mention of update</td>
</tr>
<tr>
<td>Aus CN 200274</td>
<td>Localised PCa management evidence based recommendations</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>Comprehensive and systematic search described; inclusion criteria not described; no interval between measurements recommended; states no widely accepted biochemical range applicable; NH uncertainty not acknowledged; variability acknowledged but not in recommendations; some supporting evidence cited; internal review carried out but not described in detail; update recommended but procedure not described</td>
</tr>
<tr>
<td>DUA 200776</td>
<td>PCa management guideline</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>16</td>
<td>Systematic search carried out but not fully described; some description of inclusion criteria; interval between measurements based on evidence; consensus threshold used for post-RT, no justification for post-RP; NH uncertainty not acknowledged; variability acknowledged but not in recommendations; some link to supporting evidence; external review partly described; update recommended but procedure not described</td>
</tr>
<tr>
<td>EAU 200979</td>
<td>PCa management guideline</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>Systematic search carried out but only partly described; no description of inclusion criteria; formulation of recommendations not described; no basis for interval between measurements; consensus thresholds used; NH uncertainty not acknowledged; variability acknowledged but not in recommendations; some link to supporting evidence; external review conducted but not described; no mention of update</td>
</tr>
<tr>
<td>NCI PDQ 200877</td>
<td>PCa treatment evidence based summary for health professionals</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>No search described; no description of inclusion criteria; formulation of recommendations not described; no interval between measurements recommended; evidence-based recommendations for threshold post-RT not possible; basis for post-RP threshold given; NH uncertainty acknowledged in recommendations; variability acknowledged but not in recommendations; threshold/variability/uncertainty; some links to supporting evidence; external review not described; limited description of update procedures</td>
</tr>
<tr>
<td>NCCN 200980</td>
<td>PCa management guideline</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>No search described; no description of inclusion criteria; no description of formulation of recommendations; some justification given for interval between measurements; consensus threshold used for post-RT, no justification for post-RP; NH uncertainty and variability not acknowledged; some links to supporting evidence; external review not described; no mention of update</td>
</tr>
<tr>
<td>Country</td>
<td>Guideline Type</td>
<td>Quality Score</td>
<td>Recommendation Formulation</td>
<td>Disagreement Reporting</td>
<td>Evidence Justification</td>
<td>External Review</td>
<td>Update Procedure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------------------</td>
<td>---------------</td>
<td>----------------------------</td>
<td>------------------------</td>
<td>-----------------------</td>
<td>----------------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NICE 2008</td>
<td>PCa diagnosis and treatment guideline</td>
<td>4 3 3 4 3 2 3 22</td>
<td>Systematic search carried out and fully described; inclusion criteria developed for each question but not reported; recommendation formulation described but methods used to deal with disagreement not reported; attempted to find evidence to justify interval between measurements; relevant discussion regarding choice of thresholds; NH uncertainty and variability acknowledged in recommendations; clear link to supporting evidence; external review no described; update recommended but procedure described.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK PCWG 1999</td>
<td>PCa management guideline</td>
<td>3 1 2 2 1 1 1 12</td>
<td>Systematic search carried out and fully described; inclusion criteria developed for each question but not reported; recommendation formulation described but methods used to deal with disagreement not reported; interval between measurements justified; relevant discussion regarding choice of threshold post-RT only; NH uncertainty not acknowledged, variability acknowledged but not in recommendations; clear link to supporting evidence; external review not described; update recommended but procedure described.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion

We found considerable inconsistency in guidelines’ recommendations for the use of prostate specific antigen as a monitoring test, even when they were published within a few years of each other. Factors considered to be important when specifying a monitoring strategy were given limited attention and were not well supported with reference to primary literature.

Recommendations on when to test and what action to take consequent on a given test result were very much considered in isolation from each other. ‘When to test’ appeared to be almost exclusively determined by standard follow-up schedules rather than on any scientific basis. Although most guidelines acknowledged the potential presence of measurement variability, they did not attempt to account for its potential effect on test interpretation. A systematic review of biological variation in prostate specific antigen found mean biological variability of 20%. Using reference change value methodology it concluded that in order to be 95% sure that a change in total prostate specific antigen is not due to random variation, the change needs to be around 50% on the previous measurement. This review was not cited by any of the eight guidelines subsequently published.

Recommendations on when to take action were based on consensus statements or retrospective case series with little attention paid to variations in the definition of the threshold, the definition of clinical failure, and the frequency and length of follow-up between studies, all of which can affect the accuracy of any given cut-off. Sensitivity and specificity are also known to be affected by differences in patient case-mix between studies. This was not acknowledged by any of the identified guidelines. However, a 2005 review of prostate specific antigen for monitoring prostate cancer found it impossible to recommend any single definition of biochemical failure following either treatment for the reasons listed above. This review was cited by only one of the nine guidelines, possibly due it not being fully systematic. Given the lack of description of inclusion criteria used in the guidelines, it is difficult to reconcile why an individual study or review was included or not.

Reviews of guidelines in other areas have shown similar findings regarding the presentation of evidence for recommended monitoring schedules. Reviews of treatment and diagnostic guidelines have identified a similar inconsistency in recommendations between
guidelines and variation in evidence cited with some referring to a substantial body of evidence whilst others presented very little.\textsuperscript{93-97}

A number of factors are likely to contribute to these findings. In the first instance, although the area is beginning to receive more attention, there is a lack of high quality evidence for, and indeed lack of clear methodological guidance on, what to consider when establishing monitoring strategies.\textsuperscript{38} It is therefore perhaps not surprising that relevant evidence has not been used to inform guidelines.

Secondly, the various pieces of information needed to inform a monitoring strategy are not usually available from a single study. Ideally, one or more monitoring strategies should be evaluated in a randomised controlled trial or some form of prospective comparative study. Where there is high quality evidence, greater consensus between guideline recommendations and stronger guideline recommendations have been found.\textsuperscript{93} Randomised trials of monitoring, however have their own challenges and are consequently relatively rare. Instead, evidence has to be gathered from various sources. Although the diversity of evidence needed to inform coherent monitoring strategies makes the identification of relevant pieces of evidence a challenge for guideline developers and likely adds to the inconsistency in recommendations between guidelines, guideline developers have a responsibility to highlight recommendations where there is a lack of evidence or the evidence is inconsistent.

Efforts to improve the evidence base for monitoring are ongoing. For example, a Bayesian hierarchical changepoint model has been used to simulate multiple post-radiotherapy prostate specific antigen series from primary data; the sensitivity and specificity of different definitions of biochemical failure were then compared, allowing characteristics that might affect accuracy to be controlled.\textsuperscript{99} More pertinently, statistical models using estimates of mean change and variability in a measurement over time to suggest optimal monitoring intervals are being developed. A review of four case studies found that for each topic the results suggested over frequent monitoring.\textsuperscript{100-104} There is clear potential for the extension of this work to monitoring in other settings.

Finally, general failings in the guideline development process are likely to contribute significantly to the variations between published guidelines. In a review of hypertension guidelines, Campbell and colleagues found a lack of methodological rigour in the guideline
development process. In our sample, the National Institute for Health and Clinical Excellence and the Australian Cancer Network were the only organisations to cite published handbooks on guideline development, which may explain their higher ratings on the evaluation instrument; those clearly based on expert consensus tended to score considerably lower. Others suggest that the greater the involvement of clinical experts in the development process of the guideline, the less the recommendations reflect the research evidence. It is likely that in the absence of clear methodologies for assessing monitoring strategies, greater involvement of methodologists on guideline panels would be beneficial.

**Strengths and limitations**

Our research has some limitations. Our literature search was limited to one major medical database, supplemented with searches of more specialist resources and records were limited to English language only. We believe, however, that we have identified key guidelines that provide a good representation of the methodologies in use by well known agencies. Although other guidelines may be available, they are unlikely to have used alternative methods or to report on evidence that the included guidelines have missed.

Secondly, our use of the original Appraisal of Guidelines for Research & Evaluation instrument may be criticised given that it is now 10 years since it was published, however at the time the framework was chosen, the update to the original instrument and other potentially useful frameworks were not yet available. Nevertheless, our approach to assessing the development and content of the guidelines was systematic and provides a reasonable means of comparison between guidelines.

We were not able to comment fully on the state of the underlying literature cited in support of monitoring schedules, as we did not retrieve copies of all cited primary studies. Furthermore, our ‘generous’ approach to associating citations to recommendations may have inadvertently led to citations incorrectly associated with recommendations. This may have led to some bias in favour of the guidelines, however could only be avoided by a full review of all evidence cited or by direct contact with the guideline authors to determine which aspect of the recommendations were supported by the citations given, both of which were outwith the scope of this review.
Finally, our use of only one case study may limit the generalisability of our results to other topic areas, however we have no reason to believe the picture would be any better or worse for other fields, and indeed Moschetti and colleagues had similar findings for monitoring in cardiovascular disease.92

Our systematic approach to assessing the development and content of the guidelines provides a valuable insight into how strategies for monitoring are developed and reported and we were able to present a general picture of the type of evidence that has been cited. The true picture may be even worse given our attempt to attribute citations to recommendations wherever possible.

Conclusions

Our findings highlight the lack of a scientific or systematic approach to the development of monitoring schedules for the use of prostate specific antigen as reported in clinical guidelines. This was due both to inadequacies in the evidence base and to inappropriate use of the available evidence, resulting in considerable inconsistencies between guidelines.

Guideline developers should be encouraged to adopt systematic approaches to guideline development, such as those developed in Australia and the United States, and should take care to explicitly consider each element of a recommended monitoring schedule (interval, threshold and action to be taken on crossing that threshold) and the standard of its evidence base.107, 108
<table>
<thead>
<tr>
<th>Study design/aim (extracted from abstract)</th>
<th>Focus of study</th>
<th>Used to support guideline statements on:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Frequency</strong></td>
<td><strong>Threshold (RP)</strong></td>
<td><strong>Threshold (RT)</strong></td>
</tr>
<tr>
<td>ASTRO 1997**</td>
<td>Consensus statement providing guidelines for PSA following radiation therapy</td>
<td>ASTRO consensus statement</td>
</tr>
<tr>
<td>Cox 1999**</td>
<td>Report of ASTRO consensus panel to develop evidence-based guidelines for (1) prostate re-biopsy after radiation and (2) radiation therapy with rising PSA levels after radical prostatectomy in the management of patients with localized prostatic cancer</td>
<td>ASTRO consensus statement</td>
</tr>
<tr>
<td>Roach 2006**</td>
<td>Reports second consensus conference to revise the ASTRO definition of BF</td>
<td>ASTRO consensus statement</td>
</tr>
<tr>
<td>Carroll 2001**</td>
<td>Best practice statement (AUA)</td>
<td>Best practice statement</td>
</tr>
<tr>
<td>Aus 2006**</td>
<td>Review of high-intensity focused ultrasound (HIFU) and cryosurgery as the primary treatment option in patients with prostate cancer.</td>
<td>Review</td>
</tr>
<tr>
<td>Bott 2004**</td>
<td>Review of management of recurrence following RP</td>
<td>Review</td>
</tr>
<tr>
<td>Catton 2003**</td>
<td>Review/comment paper examining follow-up strategies</td>
<td>Review</td>
</tr>
<tr>
<td>Cookson 2007**</td>
<td>AUA review of the variability in published definitions of biochemical recurrence; recommends a standard definition in patients treated with radical prostatectomy</td>
<td>Review</td>
</tr>
<tr>
<td>Edelman 1997**</td>
<td>Review of available data on follow-up strategies</td>
<td>Review</td>
</tr>
<tr>
<td>Lee 2005**</td>
<td>Review of PSA kinetics in addition to clinical factors in the selection of patients for salvage local therapy</td>
<td>Review</td>
</tr>
<tr>
<td>Nelson 2003**</td>
<td>Review of RP for PCa</td>
<td>Review</td>
</tr>
<tr>
<td>Polascik 1999**</td>
<td>Review of PSA</td>
<td>Review</td>
</tr>
<tr>
<td>Selley 1997**</td>
<td>HTA review of PCa mgmt</td>
<td>Review</td>
</tr>
<tr>
<td>Vicini 2005**</td>
<td>Review of PSA for monitoring pts after radical Rx</td>
<td>Review</td>
</tr>
<tr>
<td>Yao 2003**</td>
<td>Review/comment paper examining follow-up strategies</td>
<td>Review</td>
</tr>
<tr>
<td>Reference</td>
<td>Type of Study</td>
<td>Subjects</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------</td>
<td>----------</td>
</tr>
<tr>
<td>Albertsen 2004&lt;sup&gt;126&lt;/sup&gt;</td>
<td>Retrospective(?)</td>
<td>Study of 1136 men undergoing surgery or radiation</td>
</tr>
<tr>
<td>Amling 2001&lt;sup&gt;121&lt;/sup&gt;</td>
<td>Retrospective(?)</td>
<td>Analysis of 2,782 men who had undergone radical prostatectomy to attempt to determine the best PSA cut point for defining BF</td>
</tr>
<tr>
<td>Booker 2004&lt;sup&gt;125&lt;/sup&gt;</td>
<td>Study of telephone follow-up led by a specialist nurse for pts undergoing RT</td>
<td></td>
</tr>
<tr>
<td>Buyyounouski 2005&lt;sup&gt;125&lt;/sup&gt;</td>
<td>Retrospective(?)</td>
<td>Review of 688 men who had undergone RT to compare three definitions of biochemical failure (BF) in terms of sensitivity, specificity etc for detecting of clinical progression</td>
</tr>
<tr>
<td>Cathala 2003&lt;sup&gt;124&lt;/sup&gt;</td>
<td>Feasibility study on 140 patients undergoing RP</td>
<td>To determine acceptability of an internet FU service</td>
</tr>
<tr>
<td>Cheung 2005&lt;sup&gt;128&lt;/sup&gt;</td>
<td>Retrospective(?)</td>
<td>Analysis of 101 men who received salvage RT to determine</td>
</tr>
<tr>
<td>Crook 1997&lt;sup&gt;77&lt;/sup&gt;</td>
<td>Prospective study of 207</td>
<td>To correlate the failure pattern after radiotherapy (RT) with pretreatment PSA and post-RT nadir PSA</td>
</tr>
<tr>
<td>D’Amico 2004&lt;sup&gt;127&lt;/sup&gt;</td>
<td>Retrospective review of 8669 men who had undergone radical Rx</td>
<td>To determine whether a short post-Rx PSADT is a suitable surrogate endpoint for prostate cancer specific mortality</td>
</tr>
<tr>
<td>Eastham 2003&lt;sup&gt;128&lt;/sup&gt;</td>
<td>Retrospective analysis of an unscreened population of 972 men over 4 years to determine whether year-to-year fluctuations in PSA levels are</td>
<td></td>
</tr>
<tr>
<td>Frazier 1993&lt;sup&gt;73&lt;/sup&gt;</td>
<td>Analysis of 226 patients who underwent radical perineal prostatectomy to identify whether raised serum PSA infers failure of the procedure</td>
<td></td>
</tr>
<tr>
<td>Horwitz 2005&lt;sup&gt;51&lt;/sup&gt;</td>
<td>Determined the sensitivity and specificity of several BF definitions using pooled data on 4,839 patients treated with external beam radiation therapy (RT) alone</td>
<td></td>
</tr>
<tr>
<td>Klotz 2008&lt;sup&gt;54&lt;/sup&gt;</td>
<td>Reports PSADT in a series of 299 patients undergoing active surveillance for PCa</td>
<td></td>
</tr>
<tr>
<td>Kuban 2006&lt;sup&gt;44&lt;/sup&gt;</td>
<td>Primary study of patients treated with radioisotopic implant as solitary treatment for T1-T2 prostate adenocarcinoma (n=2,693). Multiple PSA failure definitions were tested for their ability to predict clinical failure.</td>
<td></td>
</tr>
<tr>
<td>Leibman 1995&lt;sup&gt;11&lt;/sup&gt;</td>
<td>Retrospective review of 628 patients who underwent RP to determine whether PCA recurrence can occur without an increase in serum PSA</td>
<td></td>
</tr>
<tr>
<td>Nielsen 2008&lt;sup&gt;12&lt;/sup&gt;</td>
<td>Retrospective review of data from 2,570 men who had undergone RP to examine the effect of applying the 2005 ASTRO definition of BF (for RT pts) to surgical series.</td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Study Details</td>
<td>FU/Defns</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Niwakawa 2002</td>
<td>Study of 221 patients treated with RP to determine the optimal frequency and method of follow-up to minimize medical cost</td>
<td>FU - optimal frequency</td>
</tr>
<tr>
<td>Oefelein 1995</td>
<td>Retrospective review of data from 394 men who underwent RP to characterize the incidence of recurrent carcinoma despite undetectable serum PSA levels</td>
<td>NH of pts post-Rx</td>
</tr>
<tr>
<td>Patel 2005</td>
<td>Retrospective review of 48 patients who had undergone salvage RT for biochemical relapse after RP to determine whether PSA is a suitable selection criterion for salvage radiotherapy</td>
<td>Prognosis following salvage Rx</td>
</tr>
<tr>
<td>Pickles 2006</td>
<td>An analysis of a ‘prospective’ database of 2030 patients who underwent EBRT or brachytherapy to determine the false call rate for PSA relapse according to nine different PSA relapse definitions after a PSA bounce has occurred</td>
<td>Testing defns of BF</td>
</tr>
<tr>
<td>Pound 1999</td>
<td>Retrospective review of a large surgical series (n=1997) to examine the natural history of progression to distant metastases in men with raised PSA following surgery</td>
<td>NH of pts post-Rx</td>
</tr>
<tr>
<td>Ragde 1997</td>
<td>Study of 126 patients with localised PCA to determine the efficacy of treatment with iodine-125 radionuclides (2 definitions of PSA failure used)</td>
<td>Testing defns of BF</td>
</tr>
<tr>
<td>Ray 2006</td>
<td>Retrospective review of 4839 patients treated definitively with RT to determine the significance of PSA nadir and time to PSA nadir in predicting biochemical or clinical disease-free survival</td>
<td>NH of pts post-Rx</td>
</tr>
<tr>
<td>Ritter 1992</td>
<td>Study of the prognostic value of the PSA in pretreatment evaluation and posttreatment follow-up in 63 patients undergoing RT for localised PCa</td>
<td>NH of pts post-Rx</td>
</tr>
<tr>
<td>Rose 1996</td>
<td>To identify patients' symptoms following completion of radiotherapy for common cancers by a nurse-managed telephone interview in 111 pts treated with RT.</td>
<td>FU acceptability</td>
</tr>
<tr>
<td>Sandler 2000</td>
<td>Retrospective database study of 1844 pts who had undergone RT and had a minimum of 2 post-RT PSAs separated by at least 1 week to determine the significance of biochemical failure i.e. in terms of survival</td>
<td>NH of pts post-Rx</td>
</tr>
<tr>
<td>Sartor 1997</td>
<td>Primary study of 400 patients treated with radiotherapy to determine whether the rate of PSA rise could differentiate future local versus metastatic failure.</td>
<td>NH of pts post-Rx</td>
</tr>
<tr>
<td>Stamey 1989</td>
<td>Study of pre- and post-Rx serum PSA in 102 men who underwent radical prostatectomy to determine usefulness of PSA as a pre-operative marker.</td>
<td>NH of pts post-Rx</td>
</tr>
<tr>
<td>Stephan 2006</td>
<td>Assessed 5 frequently used commercial assay combinations in sera from 314 patients with prostate cancer (PCa) and 282 men with no evidence of prostate cancer to identify the interchangeability of the PSA values</td>
<td>Measurement variability</td>
</tr>
<tr>
<td>Stephenson 2004</td>
<td>Retrospective review of 501 patients who underwent salvage RT following RP to identify those variables indicative of a durable response</td>
<td>Prognosis following salvage Rx</td>
</tr>
<tr>
<td>Stephenson 2006</td>
<td>Tested 10 definitions of BF on 3,125 patients who underwent RP, to identify the one that best explains metastatic progression</td>
<td>Testing defns of BF</td>
</tr>
<tr>
<td>Reference</td>
<td>Summary</td>
<td>NH of pts post-Rx</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Trapasso 1994</td>
<td>Primary study of patients undergoing radical retropubic prostatectomy (n=601) and followed with serial PSA determinations. Evaluated rate of detectable PSA (greater than 0.4 ng./mL) as an indicator of cancer progression.</td>
<td></td>
</tr>
<tr>
<td>Trock 2008</td>
<td>Retrospective analysis of a cohort of 635 men undergoing RP and who experienced biochemical and/or local recurrence to determine the effect of salvage RT and to identify subgroups for whom salvage treatment is most beneficial</td>
<td>Prognosis following salvage Rx</td>
</tr>
<tr>
<td>Ward 2004</td>
<td>Retrospective cohort study of 211 men with detectable PSA following RP to determine whether PSADT predicts outcomes following salvage radiotherapy</td>
<td>NH of pts post-Rx</td>
</tr>
<tr>
<td>Zagars 1997</td>
<td>Analysis of 841 men with serial PSA determinations who underwent external beam radiation without androgen ablation to determine the kinetics of serum PSA after RT and to evaluate whether such kinetics provide prognostic information.</td>
<td>NH of pts post-Rx</td>
</tr>
</tbody>
</table>
Chapter 4 - Has the randomised controlled trial design been successfully used to evaluate strategies for monitoring disease progression or recurrence? An assessment of experience to date
Introduction

Clinical consultations between patient and clinician usually involve the use of tests; often starting with a general clinical assessment and physical examination but also including the application of specific tests, from tools assessing functional or psychological status, to blood or urine tests, physiological measurements such as blood pressure, imaging tests or more invasive assessments such as colonoscopy or biopsies. Testing can inform a diagnosis or can be used for monitoring whereby a test, or set of tests are applied repeatedly over time to assist in the management of a disease or condition. Our particular interest is in monitoring individuals with (or at risk of) a disease or condition that is likely to progress or recur at some time in the future. This is distinct from monitoring in a treatment titration context, where the aim is to keep a marker within predefined limits until treatment can be discontinued or an alternative treatment is required. Although monitoring for disease progression or recurrence can serve many purposes including reassurance to patients or clinicians, it is usually undertaken to allow timely decisions to be made regarding patient management. Management decisions include the initiation of treatment to prevent some clinical outcome from occurring (e.g. variceal bleeding subsequent to cirrhosis of the liver or infertility due to Turner’s Syndrome), to delay a clinical event (e.g. progression to AIDS in HIV infection) or to otherwise improve outcome (e.g. through earlier treatment of cancer recurrence), or the goal may be to avoid or delay treatment in those who may not need it (e.g. surveillance of mild hip dysplasia in infants).

Whatever the goal, monitoring is a central activity for patient and disease management, and, just as for therapeutic interventions and for tests used in a diagnostic context, it is important to identify its impact on patient outcomes: “the primary purpose of using...tests should...be to prevent premature death and suffering and restore functional health”. Given the advantages of the randomised controlled trial (RCT) design for the evaluation of therapeutic interventions, it is tempting to assume that the same approach must be the gold standard for the evaluation of all monitoring strategies.

Trials of monitoring regimes present considerable challenges, however. As for diagnostic tests, trials of monitoring evaluate a strategy, with tests applied at specific intervals, with defined thresholds for changing patient management, and effective interventions prescribed, all of which should be specified in advance, and ideally supported by previous research. The
complexity of such strategies, and in particular the serial nature of testing, and consequent potential for “interactions between tests, repeated tests, test results and the decisions based on these results”, may necessitate unfeasibly large sample sizes in order to detect an effect on important patient outcomes. Furthermore, even with careful planning it may be difficult to capture the wider patient impact of testing in an RCT, whether it is used for diagnostic, screening, or monitoring purposes. Ferrante di Ruffano and colleagues have outlined a range of effects from testing in a diagnostic context, including emotional, cognitive and behavioural effects, which also have applications in monitoring.

We conducted a methodological review of RCTs of monitoring in order to gain some insight into how successfully the design has been used to identify patient benefit from monitoring.

**Methods**

**Literature search**

Our target sample size was 60 RCTs. The Cochrane Central Register of Controlled Trials (CENTRAL) was searched to retrieve relevant records (last updated 21-07-11; details available from authors). The search was supplemented by screening all RCTs funded by the NHS Health Technology Assessment Programme and those published in Trials journal (to December 2011). After assimilation of studies meeting the inclusion criteria, a search of the U.S. National Institutes of Health clinicaltrials.gov database was carried out (using keywords monitor#, surveill# or early or immediate treatment) and purposively sampled to include trials conducted in topic areas that would complement those already identified. The sampling was not made on the basis of trial quality. Attempts were made to identify publications related to these trials using Google Scholar and by contacting trial principal investigators to request copies of their protocols or trial reports.

**Inclusion criteria**

Trials were eligible for inclusion if they considered monitoring of a disease or condition that is likely to progress or recur at some time in the future. RCTs where the main purpose of monitoring was treatment titration or improvement in adherence to a treatment regimen or those evaluating methods of delivering monitoring were excluded as were trials of tests used for population-based screening or for diagnosis. Trials had to compare at least one formal
monitoring strategy to no formal monitoring, to an alternative monitoring strategy, or to an immediate treatment option. All clinical topics, test types and outcomes were eligible.

Trials reported only as protocols were included, but those available only in abstract form and non-English language papers were excluded. Multiple reports of a single trial were assimilated through cross-referencing.

The search was conducted by and search results screened by one reviewer (JaD).

Data extraction and analysis

A data extraction form was designed and piloted. Data were extracted on items including: the study population and topic area, monitoring strategies or interventions in experimental and control arms (including details of testing frequency, threshold and intervention), the citation of evidence to support these features of the monitoring strategies, study design and validity criteria (Box 1). Details of the primary outcomes used were also extracted and classified as patient, process or composite outcomes; if not clearly reported the outcome used in the study’s power calculation was extracted, or failing that, the outcome most closely related to the study aim. Where final analyses were reported, the result for the primary outcome was extracted and, where possible, compared graphically to that predicted in the sample size calculation. For 30 trials, data were extracted independently by two authors (JaD and AS or JP); for the remaining 28, data were extracted by one author (JaD) and checked by a second (AS or JP). Any disagreements were resolved by consensus.

Studies were considered according to topic area, types of tests and monitoring strategies, the study aim and change in patient care that was evaluated, study validity and primary outcomes and results.

Results

The CENTRAL search retrieved 4697 potentially eligible records (Figure 3). 120 titles were selected for further evaluation, along with 9 trials identified from the HTA database or Trials Journal. Following full-text review, 49 trials published in 58 publications were selected for inclusion. Twenty trials identified from the clinicaltrials.gov database were selected and trial
principal investigators contacted. Documents related to 12 of these were successfully retrieved, of which 7 met the inclusion criteria. Reference list screening identified a further two eligible trials. Of the 58 included trials, five were reported in two publications, and 19 prior or related publications were also identified (for a total of 74 papers). Figure 4 shows a general upward trend in the number of trials published per year.\textsuperscript{152, 153}

Figure 3 Flow diagram of the trial selection process
General description of included trials

The trials were primarily conducted in the fields of cancer (29%), cardiovascular disease (16%), and renal disease (16%) (Table 6). A further 9% of trials were conducted in patients with aneurysm – either abdominal aortic or cranial - and in transplant recipients, including stem cell or bone marrow (9%). Most were parallel in design except for six multi-arm trials; the total number of experimental arms was 68.

Of the 58 trials, 34 were available as full trial reports, 11 as interim analyses and 13 as trial protocols. Twelve trials were stopped early (21%); interim analyses were available for 9 and protocols only for the remaining three. The most common reason for trial stopping was a lower than expected event rate in the control groups (33%; 4/12). Three-quarters of trials reported sample size calculations (78%; 45/58) and a quarter were industry-sponsored (24%; 14/58). The median sample size was 272 [IQR 120, 599] and median follow-up 21 months [IQR 12, 60].
### Table 6 General description of included trials (n=58)

#### Patient group

<table>
<thead>
<tr>
<th>Group</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>17</td>
<td>29%</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>9</td>
<td>16%</td>
</tr>
<tr>
<td>(all with implanted devices)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Renal disease</td>
<td>9</td>
<td>16%</td>
</tr>
<tr>
<td>(all haemodialysis recipients)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aneurysm</td>
<td>5</td>
<td>9%</td>
</tr>
<tr>
<td>(4 AAA and 1 cranial)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transplant recipients</td>
<td>5</td>
<td>9%</td>
</tr>
<tr>
<td>(3 stem cell/bone marrow; 2 solid organ)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other (&lt;3 per group)</td>
<td>13</td>
<td>22%</td>
</tr>
</tbody>
</table>

#### No. of study arms

<table>
<thead>
<tr>
<th>Arms</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>52</td>
<td>90%</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>7%</td>
</tr>
</tbody>
</table>

#### Publication type

<table>
<thead>
<tr>
<th>Publication type</th>
<th>No.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full trial report</td>
<td>34</td>
<td>59%</td>
</tr>
<tr>
<td>Interim analysis</td>
<td>11</td>
<td>19%</td>
</tr>
<tr>
<td>Protocol</td>
<td>13</td>
<td>22%</td>
</tr>
</tbody>
</table>

**Trial early stopping reasons (n=12)**

- Recruitment difficulties: 3 (25%)
- Technology issues: 3 (25%)
- Interim analyses showed:
  - early superiority: 1 (8%)
  - no evidence of benefit: 1 (8%)
  - low event rate in CG: 4 (33%)

#### Sample size calculations

- Reported: 45 (78%)
  - Full trial reports (n=34): 25 (74%)
  - Interim analyses (n=10): 8 (80%)
  - Protocol only (n=14): 12 (86%)

#### Study funding

- Industry sponsored: 14 (24%)
- Non-Industry sponsored: 33 (57%)
- Not reported: 11 (19%)

#### Total number of participants randomised (median [IQR])*

272 [120, 599] (range 64 to 4439)

#### Length of follow-up in months (median [IQR])*

21 [12, 60] (range 1 to 240 months)

* Full trial reports only (n=34)
Description of monitoring strategies

One hundred and thirty-nine tests were applied in the control arms of 55 RCTs, excluding those with no formal surveillance or where all patients were treated (Table 7). After clinical examination, imaging tests were the most commonly used (39%). The 55 ‘new’ tests applied in the experimental arms included biochemical (35%), imaging (34%), or physiological measurements (16%), and implanted devices (9%).

Figure 5 shows that while the frequency of application of the tests was well reported, the method of application of the tests was provided for only two thirds of experimental arms and less than half of control arms. The test thresholds, describing when a change is patient management was indicated were reported for 79% and 51% respectively. A simple threshold approach (i.e. where the patient crosses a predefined threshold on a single measurement) was used to judge an abnormal test result in 36% of control tests and 42% of experimental tests (Table 7). Few trials reported using test measurements over time to define an abnormal result, although percentages were higher for the experimental tests: the change from the previous measurement was reported for 1% of control tests and 7% of experimental tests; and a more complex algorithm taking account of more than one test result was reported for 7% and 11% respectively. There was also limited reporting of repeated testing to confirm abnormal or indeterminate test results; reported for less than 20% of tests in both arms (Table 7).

The recommended change in patient management following a positive monitoring test was the same in experimental and control groups for 69% of trials. The change in management prescribed was usually treatment (45% of both usual care and experimental arms). Some form of confirmatory testing was indicated in 31% and 41% respectively (Table 7), many of which also recommended the treatment to be used following a positive confirmatory test. Details of the actual intervention given and method of application of the intervention were provided for 74% and 28% of intervention groups, compared to only 62% and 22% of control arms. (Figure 5).

Citation of evidence to support the various elements of the monitoring strategies were higher for the experimental arms than for control (Figure 6), however much of the mention of previous studies was found in the introduction or discussion sections of the papers rather than
explicitly supporting the particular test frequency, threshold or interventions reported in the Methods.

Figure 5 Adequate description of monitoring strategy elements

Figure 6 Citation of evidence to support features of monitoring strategy

<table>
<thead>
<tr>
<th>Element</th>
<th>Control</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test frequency</td>
<td>Yes (Methods)</td>
<td>Yes (Methods)</td>
</tr>
<tr>
<td>Threshold</td>
<td>Yes (Intro/Disc)</td>
<td>Yes (Intro/Disc)</td>
</tr>
<tr>
<td>Intervention</td>
<td>Yes (Methods)</td>
<td>Yes (Methods)</td>
</tr>
</tbody>
</table>

Legend:
- Yes (Methods)
- Yes (Intro/Disc)
- No
- n/a - all Rx
### Table 7 Details of monitoring schemes

<table>
<thead>
<tr>
<th>Tests</th>
<th>Control (n=139)</th>
<th>Experimental (n=55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biochemical</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>Clinical</td>
<td>53</td>
<td>0</td>
</tr>
<tr>
<td>Cytological</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Histological</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Imaging</td>
<td>46</td>
<td>16</td>
</tr>
<tr>
<td>Imaging (invasive)</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Implanted device</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Physiological</td>
<td>12</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of threshold used (excluding clinical assessment)</th>
<th>Control (n=139)</th>
<th>Experimental (n=55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not reported</td>
<td>41</td>
<td>18</td>
</tr>
<tr>
<td>Simple threshold</td>
<td>30</td>
<td>21</td>
</tr>
<tr>
<td>Change from prior</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Algorithm</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>&gt;1 threshold</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Repeat measure taken to confirm abnormal result (excluding clinical assessments)</th>
<th>Control arm (n=86 tests‡)</th>
<th>Exp arm (n=55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All abnormal</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Indeterminate only</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Not repeated</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Not reported</td>
<td>71</td>
<td>46</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in patient management following positive monitoring test (n=58 trials)</th>
<th>Control (n=139)</th>
<th>Experimental (n=55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirmatory testing (noninvasive)</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Confirmatory testing (invasive)</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>Treatment</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>Treatment or further investigation</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>More intensive surveillance + Rx option</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Not described</td>
<td>11</td>
<td>5</td>
</tr>
</tbody>
</table>

Excludes one trial where control arm undergo immediate treatment and two where no formal surveillance is standard practice.

Out of a total of 68 experimental arms, 13 experimental arms where no new test was introduced (test frequency varied), 9 'treat all on recruitment' arms and one study evaluating no formal surveillance are not included. Those treatment trials where randomisation was conditional on crossing a predefined threshold (n=4) were included; in 8 arms more than one test was added or replaced.

Excludes 53 'clinical assessments' includes 1 trial where experimental arm was 'treat all on recruitment' but confirmatory test was undergone first.

Includes 12 trials where experimental arms was immediate treatment.
What was the aim of the monitoring evaluation and what change in patient care was implemented?

In 78% of trials (45/58), usual care in the control arm was based on some form of clinical assessment (Table 8), often with a focus on one main test (12/58), or more commonly a battery of other tests (21/58). Across the 68 experimental arms, the most common change in monitoring was the addition of a new test to an existing monitoring strategy (n=30) or as triage to a more invasive test (n=3). In 12 experimental arms, there was no change in the tests used, but test frequency was increased (n=5) or decreased (n=7). In 53% of the experimental arms (36/68), the frequency of outpatient visits undertaken by patients was the same as for the control arm.

For most trials, the change in patient management was intended to improve patient outcomes (75%; 51/68), either through earlier initiation of treatment or better selection of patients requiring treatment (Figure 7). This was generally achieved through the addition of a new test to an existing monitoring strategy (55%; 28/51), although in 12 (24%), triallists evaluated earlier treatment by enrolling patients who had not yet reached the standard (implicit or explicit) threshold for treatment and randomising them to either an immediate treatment option or continued surveillance. In four there was an explicit underlying evaluation of a monitoring strategy as patients in both groups were monitored with a specific test following recruitment and randomised only when their test result crossed a predefined threshold for intervention. 154-157

For the remainder, the goal was to maintain the same patient outcomes (25%; 17/68) but to reduce the amount of testing undertaken, either by reducing the number of tests (n=9) or reducing the amount of invasive testing undergone (n=7), or introducing surveillance to avoid treatment (n=1) (Figure 7). This was achieved by, for example, replacing an existing test (4/16), reducing the frequency of testing (8/16) or by adding a new triage test (3/16) to select patients for a more invasive test. The remaining study was a non-inferiority trial aiming to demonstrate that additional testing did not improve survival of patients with colorectal cancer. 158

In terms of study validity, sequence generation and allocation concealment were judged to be adequate in 48% and 45% of trials respectively (Figure 8). Blinding of study participants, of
study personnel and of outcome assessment was rarely implemented; in the majority of trials, blinding was not described and a judgement had to be made as to the likelihood of blinding being present or not. Uniform and unbiased outcome assessment (where the primary outcome was assessed in the same way in both arms and was not determined by the monitoring test under evaluation) was present in 83% of studies. In eight trials, however, the presence of the primary outcome was clearly measured by the monitoring test, so that the outcome was defined differently between arms.
Table 8 What change in patient care was evaluated?

<table>
<thead>
<tr>
<th>Change in testing strategy</th>
<th>Experimental arms (n=68)</th>
<th>Effect on office visits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Addition of test</td>
<td>Triage to existing test</td>
</tr>
<tr>
<td>No formal monitoring</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>Monitoring focused on single test</td>
<td>11</td>
<td>19%</td>
</tr>
<tr>
<td>Clinical assessment</td>
<td>12</td>
<td>21%</td>
</tr>
<tr>
<td>Clinical assessment plus one main test</td>
<td>12</td>
<td>21%</td>
</tr>
<tr>
<td>Clinical assessment plus multiple tests*</td>
<td>21</td>
<td>36%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th>58</th>
<th>29</th>
<th>3</th>
<th>10†</th>
<th>13‡</th>
<th>1</th>
<th>12§</th>
<th>68</th>
<th>8</th>
<th>36</th>
<th>12</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>43%</td>
<td>4%</td>
<td>15%</td>
<td>19%</td>
<td>1%</td>
<td>18%</td>
<td></td>
<td>12%</td>
<td>53%</td>
<td>18%</td>
<td>3%</td>
<td></td>
</tr>
</tbody>
</table>

58 control arms and 68 experimental arms due to 6 multiarm trials

* includes one trial where standard care is treatment on recruitment with subsequent regular FU; the experimental arm undergo treatment only if the surveillance tests continue to show abnormality at later follow-up

† Test frequency was also reduced in three arms

‡ Test frequency was increased in 5 experimental arms, reduced in 7 and stayed same in 1 (this was a trial of treatment versus surveillance, but following treatment, patients in the control arm underwent the same surveillance as those in the experimental arm\textsuperscript{59})

§ Includes 3 trials where all patients were monitored with a new test following trial recruitment and were only randomised on crossing a given threshold. Participants in the control arms of these trials continued to undergo standard monitoring following randomisation and were treated according to usual criteria.
Primary outcomes and results

Approximately half of trials chose to evaluate patient-related primary outcomes (Table 9). A third (9/31) aimed to assess the impact of monitoring on mortality, while over half aimed to detect either new (9/31) or recurrent (7/31) disease. Process outcomes evaluated were primarily related to the number of patients treated between arms (8/13) or the time to treatment (4/13).

Around a third of full trial reports (12/34) reported statistically significant effects on the primary outcome (Box 3). The 16 trials reporting power calculations and reporting results as risk differences generally found smaller effects on the primary outcome compared to those predicted (Figure 9).
**Table 9 Description of primary outcomes used and main result**

<table>
<thead>
<tr>
<th>Type of primary outcome measures</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>31</td>
<td>53%</td>
</tr>
<tr>
<td>Process</td>
<td>13</td>
<td>22%</td>
</tr>
<tr>
<td>Composite</td>
<td>10</td>
<td>17%</td>
</tr>
<tr>
<td>Unclear/Not described</td>
<td>4</td>
<td>7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient outcomes used</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>4</td>
<td>13%</td>
</tr>
<tr>
<td>Function</td>
<td>1</td>
<td>3%</td>
</tr>
<tr>
<td>New disease rate</td>
<td>9</td>
<td>29%</td>
</tr>
<tr>
<td>Recurrent disease rate</td>
<td>7</td>
<td>23%</td>
</tr>
<tr>
<td>Mortality</td>
<td>9</td>
<td>29%</td>
</tr>
<tr>
<td>Psychological morbidity</td>
<td>1</td>
<td>3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Process outcomes used</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic yield</td>
<td>1</td>
<td>8%</td>
</tr>
<tr>
<td>Therapeutic yield</td>
<td>8</td>
<td>62%</td>
</tr>
<tr>
<td>Timing of care</td>
<td>4</td>
<td>31%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Was effect on PO statistically significant? (n=34*)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>12</td>
<td>35%</td>
</tr>
<tr>
<td>No</td>
<td>20</td>
<td>59%</td>
</tr>
<tr>
<td>Not reported</td>
<td>2</td>
<td>6%</td>
</tr>
</tbody>
</table>

* Full trial reports only
Box 3 Full trial reports

<table>
<thead>
<tr>
<th></th>
<th>Non-significant result (n=22)</th>
<th>Statistically significant result (n=12)</th>
<th>Chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean sample size</td>
<td>405</td>
<td>778</td>
<td></td>
</tr>
<tr>
<td>Median ss (range)</td>
<td>274 (64, 1340)</td>
<td>225 (79, 4439)</td>
<td></td>
</tr>
<tr>
<td>Industry funded</td>
<td>4/22 18%</td>
<td>4/12 33%</td>
<td>P=0.912</td>
</tr>
<tr>
<td>Median FU(mos)</td>
<td>35</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Adequate sequ gen</td>
<td>18/22 82%</td>
<td>2/12 17%</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>Adequate alloc conc</td>
<td>12/22 55%</td>
<td>4/12 33%</td>
<td>P=0.24</td>
</tr>
<tr>
<td>Outcome ass blinded</td>
<td>3/22 14%</td>
<td>3/12 25%</td>
<td>P=0.41</td>
</tr>
<tr>
<td>No uniform outcome ass</td>
<td>3/22 14%</td>
<td>2/12 17%</td>
<td>P=0.81</td>
</tr>
<tr>
<td>PO – patient outcome</td>
<td>16/22 73%</td>
<td>5/12 42%</td>
<td>P=0.07</td>
</tr>
<tr>
<td>PO – process outcome</td>
<td>4/22 17%</td>
<td>5/12 41%</td>
<td>P=0.14</td>
</tr>
<tr>
<td>Evidence cited</td>
<td>11/22 50%</td>
<td>4/12 33%</td>
<td>P=0.35</td>
</tr>
<tr>
<td>Evidence cited for exp arm</td>
<td>6/22 27%</td>
<td>3/12 25%</td>
<td>P=0.89</td>
</tr>
<tr>
<td>Evidence cited for control arm</td>
<td>11/22 50%</td>
<td>5/12 42%</td>
<td>P=0.64</td>
</tr>
<tr>
<td>Any evidence cited for exp arm</td>
<td>16/22 73%</td>
<td>9/12 75%</td>
<td>P=0.89</td>
</tr>
<tr>
<td>Any evidence cited for control arm</td>
<td>11/22 50%</td>
<td>4/12 33%</td>
<td>P=0.35</td>
</tr>
<tr>
<td>Any evidence cited in either arm</td>
<td>17/22 77%</td>
<td>10/12 83%</td>
<td>P=0.67</td>
</tr>
</tbody>
</table>

Figure 9 Comparison of observed versus predicted effects
Discussion

This is the first review that we know of to have examined the use of the randomised trial design to evaluate monitoring tests. We found that trialists have made valiant attempts to evaluate a wide range of different types of monitoring strategies in various medical fields and clinical settings and the number published per year appears to be increasing over time. Many of the strategies were relatively complex, involving the addition or replacement of a test within an existing battery of tests or changing the frequency of one or more tests, presenting considerable challenges for their evaluation. Only a small proportion of trials reported statistically significant results for the primary outcome; various possible reasons for this and the impact of other key features of the trials are worth exploring. From data presented in Hopewell et al’s (2009) publication bias paper, between 55% and 75% of published trials reviewed in 5 studies demonstrated statistically significant effects.

In the first instance, a striking lack of scientific basis for the monitoring strategies that were evaluated was presented. Just as reviews of monitoring strategies specified in clinical guidelines have shown, any existing evidence base for the strategies was poorly cited. Potential issues with the tests, thresholds, or interventions which could have been avoided with a thorough evaluation of each component of the new monitoring strategy in the context in which it is to be applied were identified only in retrospect.

Key to this is the consideration of relevant evidence related to the test(s) to be used (in terms of both accuracy and ability to predate the appearance of clinically relevant disease), the interval between test applications (which should be influenced by the expected natural history of disease and by the degree of measurement variability associated with the test in the given setting) and the change in patient management that is to be implemented (particularly in terms of establishing the effectiveness of any intervention in the trial population). Providing evidence (or at the very minimum, a rationale) for each of these components is key to the interpretation of trials of monitoring, as others have pointed out, “a mediocre test could improve outcomes when it is coupled with effective management; similarly a quality test could fail to improve outcomes in the absence of effective management”.

It appears, however, that insufficient attention was paid to establishing test properties and intervention effectiveness in the population of interest before trials were undertaken. For example, although one might expect both the rate of disease progression and the degree of measurement variability associated with a given test to be taken into account when setting
test frequency, test intervals were apparently determined by convenience or by fear of missing a key clinical event.\textsuperscript{38} There was little acknowledgement of the potential for false-positive results, limited use of repeated testing to confirm abnormal or indeterminate test results and examples of ‘personal tailoring’ of decisions based on risk from previous tests, for example, by changing the frequency of testing or altering thresholds for intervention on the basis of prior test results, were rare. A small number of studies did acknowledge problems with technical aspects of the tests evaluated, or with the test thresholds evaluated, with one reporting a change to the threshold in the trial protocol (in order to minimise the number of biopsies undertaken in the experimental arm).\textsuperscript{161-164} Others found that although the new monitoring strategies identified potential disease earlier in the disease pathway, as was intended, either the disease was detected too early to warrant intervention, the intervention itself may have been ineffective or the lack of observed benefit from monitoring may have been due to the effectiveness of treatment in the control group.\textsuperscript{165-167} It is difficult to assess how widespread such issues might be without a more in depth examination of each topic area, however it is clear that before a trial is undertaken, the specifics of the monitoring strategies to be evaluated should be established using appropriate methods and in a similar population to those eligible for the trial.

The second characteristic of the trials was the lack of detailed description of the protocols for monitoring, particularly for the control groups. Overall the methods of application of the tests and interventions, and of how test results should be used to inform downstream management were particularly poorly reported. Given the multiple components of a monitoring strategy and multitude of ways in which these might interact to affect outcome, a clearly defined protocol for testing and subsequent management is essential, not only so that the ‘intervention’ could be replicated in future, but also so that the mechanisms by which outcomes are affected can be better understood and the generalisability of the result beyond the trial can be judged.\textsuperscript{150, 160} The reasons behind the lack of description are not clear. It may be partly down to poor reporting but could also be related to a lack of acknowledgment of testing strategies as complex interventions or to an unwillingness to standardise a complex intervention.\textsuperscript{168}

Even for those elements of the intended monitoring strategies that were described, we found some evidence of lack of compliance, for example, trials reporting the use of the experimental test or the use of extensive additional testing in the control arm; failure to
administer treatment despite positive test results; and the prescribing of treatment despite negative test results.\textsuperscript{162, 165, 169-171} This apparent lack of ‘buy in’ by clinicians is difficult to interpret. One explanation might be that a monitoring test is considered more of a guide to potential changes in management rather than a definitive indication for a particular intervention. The often lengthy time frame of monitoring may also increase the likelihood of trial fatigue making the intended intervention difficult to implement over long periods. Delaney and colleagues advocate that a variety of methods might be used to ensure that complex interventions are delivered reliably over a period of time.\textsuperscript{160} Alternative designs might also be considered, particularly where monitoring relies primarily on a single test. For example, monitoring of all enrolled patients with randomisation to immediate treatment or a continued surveillance option on crossing a particular threshold leaves less to chance in terms of downstream management.\textsuperscript{154, 156, 157, 172}

In terms of study validity, trials of tests should be subject to the same validity standards of randomisation and allocation concealment as the wider RCT literature; however in many instances, the same standards of blinding of patients and particularly clinicians will be difficult to achieve. In theory, patient or clinician blinding might be more easily implemented for simpler tests such as blood tests or where tests are applied by a non-treating clinician such as a radiographer, but in many contexts, blinding could be more difficult and even inappropriate given that it is not just the test \textit{per se} that is being evaluated but its interaction with other components of the overall monitoring process. Although we found two examples where stringent attempts at blinding patients mitigated potential benefit from monitoring (both using implanted devices), the full impact of not using blinding in a monitoring trial is as yet unclear.\textsuperscript{173, 174} Nevertheless, its impact may be mitigated through the use of objective outcome measures and blinded and \textit{uniform} outcome assessment. Blinded outcome assessment should be feasible for most trials, but its use was reported for less than a fifth of our sample, and in over half of these the blinding related to outcome adjudicators rather than those collecting the outcome data. Furthermore, we observed a small number of trials with a ‘fatal flaw’ in outcome assessment, whereby the presence of the primary outcome, e.g. recurrence of cancer, was defined by a different test in each arm, thereby introducing an additional source of bias. It is of fundamental importance for the primary outcome to be uniformly defined in the same way between groups, ideally at the same point in time.
A final characteristic of these monitoring trials was an apparent lack of power to detect significant effects. This is not a phenomenon limited to trials of tests; previous authors have found sample size calculations to be based on inaccurate assumptions for the control group, with others suggesting that up to 10-20% of trials might be discontinued due to insufficient recruitment.\textsuperscript{175-177} However, statistically significant effects are easiest to achieve when two groups are allocated to different treatments; in trials of testing strategies all ‘test-positive’ patients usually undergo the same treatment reducing the potential to demonstrate clear differences in outcome between groups. We have not yet conducted an in-depth look at the sample size calculations of trials in our sample and cannot yet comment on the assumptions made around the predicted benefit from monitoring, however we did find evidence of lower than expected control group event rates both in trials that were stopped early and others with some forced to revise their sample calculations or change their primary outcome from overall survival to a surrogate outcome of number of recurrences treated surgically with curative intent.\textsuperscript{156, 173, 174, 178, 179}

Our research has some limitations. We are unlikely to have retrieved all available eligible trials. This is in part due to our focus on one main database, although this was supplemented with searches of more specialist resources, and was also partly due to a lack of standard terminology for trials of monitoring making searching a challenge. However, an exhaustive search is not as important for a methodological review as for a systematic review of effectiveness. We did not aim to systematically identify all available trials but to retrieve a sample of trials that provide a good representation of those available. Although other trials may be available, it seems likely that our review has flagged up many of the key issues.

Our review has identified a range of problems with available randomised trials of monitoring, raising real questions regarding the feasibility and appropriateness of this approach for evaluating monitoring. Trial investigators have perhaps underestimated the complexity of the interventions they were trying to evaluate and the multitude of ways in which the effect of an intended change in monitoring strategy might be mediated by other factors. The recommendations in Box 5 provide guidance for future researchers evaluating monitoring strategies.
Box 4 Validity assessment

The adequacy of the random sequence generation and allocation concealment were judged according to the Cochrane Collaboration’s ‘Risk of bias’ assessment tool.\textsuperscript{1160}

1. Sequence generation
   - Adequate – random number table, computer random number generator, coin tossing, shuffling of cards/envelopes, throwing dice, drawing lots, minimisation

2. Allocation concealment
   - Adequate – central allocation (including telephone, web-based and pharmacy-controlled randomisation), sequentially numbered, opaque, sealed envelopes

The presence of blinding was judged using the ‘Risk of bias’ supplemented with the instructions for estimating unclearly reported blinding status outlined by Akl and colleagues:\textsuperscript{13}

3. Blinding (assessed for four groups: patients, treating clinicians (i.e. those making subsequent management decisions), non-treating clinicians (e.g. those undertaking the monitoring test), and outcome assessment (both primary and secondary, e.g. outcome adjudicators)
   - Explicit statement that a group was blinded – definitely Yes
   - Explicit statement that a group was not blinded – definitely No
   - Explicit statement that investigators were blinded – definitely Yes for clinicians and outcome assessors
   - Explicit description of trial as ‘open’ or ‘unblinded’ – definitely No
   - If no explicit statement about blinding status – probably No
   - Described as single or double-blinded – use best judgement to assign ‘Probably Yes’ to one or more groups as appropriate

The method/definition of the primary outcome was appraised for both groups

4. Uniform outcome assessment
   - Present if primary outcome was defined/measured in the same way in all groups or if attempts were made to ensure that the primary outcome definition captured relevant events in both groups
   - Absent if differences in measurement of primary outcome was likely to have led to bias between groups, i.e. if monitoring test result was used to define the primary outcome. For example, where presence of recurrence was defined by existing test in one arm and by new test in experimental arm.
   - Unclear – if insufficient information available to judge
**Box 5 Recommendations for future practice**

**Triallists should:**

* Provide a scientific basis, or at a minimum a carefully considered rationale, for the monitoring strategy to be evaluated, including:
  - Test interval
  - Test threshold
  - Intervention(s) following a positive test result

*Ensure that the test(s) operate as you expect them to and interventions are effective in your intended patient population

*Provide clear guidance to clinicians taking part in the trial regarding how you expect them to respond to a positive, indeterminate or negative monitoring test result

*Make stringent attempts to avoid known biases, for example using:
  - Proper randomisation
  - Adequate concealment of allocation
  - Blinded outcome assessment

*Avoid additional bias from non-uniform outcome assessment between study arms, i.e. the (primary) outcome must not be determined by the monitoring test under evaluation

*Ensure trial follow-up is sufficiently long to allow important events to occur

*Be realistic with estimates used in sample size calculations

*Follow the CONSORT guidelines for trial reporting and clearly report the care provided in both experimental and control arms of the trial
Chapter 5 - A review of monitoring-related methodology literature
Introduction

Monitoring strategies used to direct the care of patients with potential recurrent or progressive disease are rarely evidence based. Monitoring strategies specify the frequency of observations, duration of monitoring, decision rule and threshold for a positive result, with a positive result prompting a change in patient management. There is a need for monitoring strategies to be developed based on evidence of how the disease will progress and the performance of the monitoring test to be used. Too often test frequencies are based on routine care schedules with decision rules and thresholds chosen in an ad hoc manner.

Methods

Methodological information related to monitoring was firstly sought from the first edition of the book ‘Evidence-based medical monitoring’, edited by Paul Glasziou, Les Irwig and Jeffrey Aronson. Literature searches of MEDLINE and Science Citation Index were undertaken to identify relevant papers.

A variety of searches were performed to identify relevant literature using various combinations of the following text words:

i. monitor*
ii. measure* or biomarker* or marker*
iii. serial or repeat* or periodic or longitudinal or trajectory*
iv. recurrence or progression
v. rule* or threshold* or trigger
vi. statistical process control or control chart* or reference change value or critical difference
vii. screen* with frequenc* or intensit* or interval*

Where necessary, results were filtered using either:

a. statistical or epidemiological journal titles (Biostatistics, Biometrics, Statistics in Medicine, Methods of Information in Medicine, Lifetime Data Analysis, Journal of Clinical Epidemiology, American Journal of Epidemiology, Annals of Epidemiology)

b. randomised controlled trials (sensitive search)

Additional searches were undertaken to identify literature related to statistical process control, reference change values and to the methodology of health screening.
The papers selected for review are summarised in the sections below and in Table 10.
Table 10: Summary of reviewed studies

<table>
<thead>
<tr>
<th>Reference</th>
<th>Design</th>
<th>Analysis</th>
<th>Citations†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test frequency</td>
<td>Test thresholds</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Decision rules</td>
<td>Review of methods</td>
<td>General data structure</td>
</tr>
<tr>
<td>Stevens et al. (2010)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(2010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thompson and Pocock</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(1990)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buclin et al. (2011)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(2011)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glasziou et al. (2007)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2007)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bell et al. (2008)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glasziou et al. (2008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bell et al. (2011)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2011)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bell et al. (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bell et al. (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keenan et al. (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Powers et al. (2011)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2011)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Takahashi et al. (2012)</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>(2012)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Takahashi et al. (2010)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oke et al. (2012)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2012)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proust-Lima et al.</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(2014)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proust-Lima and Taylor</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(2009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li and Gatsonis (2012)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2012)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slate and Tumbull</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(2000)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bellera et al. (2008)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(2008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bellera et al. (2009)</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(2009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inoue et al. (2004)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(2004)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtil and Rabilloud</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taylor et al. (2005)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(2005)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thiébaut et al. (2003)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(2003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wolbers et al. (2010)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(2010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soléormos et al</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bellera et al. (2008)</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(2008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeLong et al. (1985)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>(1985)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Design</td>
<td>Analysis</td>
<td>Citations</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>-------------------------</td>
<td>-------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>Test frequency</td>
<td>Test thresholds</td>
<td>Decision rules</td>
</tr>
<tr>
<td><strong>Monitoring</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>When to start consortium</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cole <em>et al.</em> (2004)&lt;sup&gt;118&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Ahdieh-Grant (2003)&lt;sup&gt;206&lt;/sup&gt;</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Biomarker design, analysis and development</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pepe <em>et al.</em> (2001)&lt;sup&gt;207&lt;/sup&gt;</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sturgeon <em>et al.</em> (2009)&lt;sup&gt;208&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Baker <em>et al.</em> (2006)&lt;sup&gt;209&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Pepe <em>et al.</em> (2008)&lt;sup&gt;210&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Sturgeon <em>et al.</em> (2010)&lt;sup&gt;211&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Baker (2000)&lt;sup&gt;212&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Baker (2009)&lt;sup&gt;213&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Ransohoff (2007)&lt;sup&gt;1&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Ransohoff and Gourlay (2010)&lt;sup&gt;214&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Lumbreras <em>et al.</em> (2008)&lt;sup&gt;215&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Parker <em>et al.</em> (2010)&lt;sup&gt;216&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Screening</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walter and Day (1983)&lt;sup&gt;217&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Day and Walter (1984)&lt;sup&gt;218&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Etzioni and Shen (1997)&lt;sup&gt;219&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Zelen (1993)&lt;sup&gt;220&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Lee and Zelen (1998)&lt;sup&gt;221&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Frame and Frame (1998)&lt;sup&gt;222&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Lee <em>et al.</em> (2004)&lt;sup&gt;223&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>McIntosh <em>et al.</em> (2002)&lt;sup&gt;224&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>McIntosh and Urban (2003)&lt;sup&gt;225&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Design</td>
<td>Analysis</td>
<td>Citations</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------------------</td>
<td>-------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>Test frequency</td>
<td>Test thresholds</td>
<td>Decision rules</td>
</tr>
<tr>
<td><strong>Time-dependent ROC curves</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zheng and Heagerty (2004)228</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtil et al (2009)229</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parker and DeLong (2003)230</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slate and Turnbull (2000)231</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Variability</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macaskill (2008)233</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smellie (2008)235</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petersen (2005)236</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fraser (2001)237</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fraser et al (1990)238</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klee (2010)239</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petersen et al (2012)240</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clerico and Emdin (2004)243</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SPC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macaskill (2008)244</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tennant et al (2007)245</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thor et al (2007)246</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gavit et al (2009)247</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Design</td>
<td>Analysis</td>
<td>Citations</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-----------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>Test frequency</td>
<td>Test thresholds</td>
<td>Decision rules</td>
</tr>
<tr>
<td>Karnon et al (2007)</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Baker (1998)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parmigiani (1997)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Real options approaches</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palmer and Smith (2000)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Driffield and Smith (2007)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Meyer and Rees (2012)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Shechter et al (2010)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Whynes (1995)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

SNR is Signal to Noise ratio.
† citations from Scopus search 24th July 2014
* review of current practice; † reporting guidelines; ‡ review of reporting standards; ‡ review of literature.
Results – Development and evaluation of monitoring strategies

Limited methodological literature was identified that provides guidance for the design of studies to evaluate monitoring tests. Work has focused more on analytic techniques to assist with the design of monitoring strategies; primarily through analysis of existing data in order to make recommendations on monitoring frequency or decision rules, or simulation work, with both approaches being specific to the disease area researched.

Designing studies to evaluate monitoring strategies

A small number of papers from the biomarker development field, cover study design issues of relevance to the monitoring field, however these largely take the form of commentary papers highlighting study design issues often not considered in biomarker development rather than empirical work.

Pepe et al discuss the five stages of biomarker development for cancer: preclinical exploratory studies, clinical assay development for clinical disease, retrospective longitudinal repository studies, prospective screening studies and cancer control studies. The initial stage is the primary search for promising biomarkers, this stage is pre-clinical and generally involves comparison of diseased and non-diseased tissue for many potential biomarkers. At this initial stage, biomarkers are assessed on their ability to produce stable results and to discriminate between disease and non-disease. The next stage sees the development of a clinical assay which can be gained non-invasively, with the discriminative ability of the assay assessed. The third stage of the process involves comparing the ability of the biomarker to differentiate between samples taken from patients with disease prior to the diagnosis of disease and samples taken from patients that are free of disease, with the aim of this stage being to evaluate if the biomarker can detect pre-clinical disease and understand what biomarker values should be used to classify a result as positive. The fourth stage uses the biomarker to prospectively screen patients, with those obtaining a positive result also receiving definitive diagnostic tests; this stage of the process allows the stage of disease the biomarker is able to detect to be known, as well as the likely positive and falsely positive yield of the test. The final stage of the process is to test the benefit of screening to the population in terms of reducing mortality. At this last stage participants should be representative of the population to be screened and a randomised controlled trial approach
can be taken to understand the difference between groups offered screening with the biomarker and not undertaking screening.

Biomarker development studies have been criticised on a number of counts,\textsuperscript{1} for producing novel findings that are often unreliable and not replicable, and being open to bias both before and after the laboratory receives the study samples, primarily due to a lack of standardisation. A number of papers have discussed issues of biases in both randomised and non-randomised study designs and to improve the development and evaluation of biomarkers, which include recommendations on both study design and analysis.\textsuperscript{1, 209, 213} Baker and colleagues, Baker, Ransohoff and Ransohoff and Gourlay further discuss the issues with biomarker development studies.\textsuperscript{1, 209, 213, 214}

Lumbreras et al presents a tool for evaluating the quality of test accuracy studies of new biomarker (or ‘omics’) technologies and Parker et al reports the results of using the QUADOMICS tool.\textsuperscript{215, 216}

In terms of specific study designs, in an overview paper of study design and methods for evaluating biomarkers for early detection of cancer, Baker and colleagues cover four different types of prospective study design:\textsuperscript{209}

- a cohort type design where asymptomatic patients are tested with a biomarker and followed up to clinical diagnosis may be an option where there is limited available evidence of benefit of the biomarker and no ethical grounds to make decisions based on the result.
- a longitudinal accuracy study where all asymptomatic patients are tested with a biomarker and all patients receive biopsy or definitive testing; however some concerns raised around this design due to over-diagnosis.
- a randomised trial design, where sensitivity is estimated using the number of positive tests and the number who develop disease in the time frame.
- a design in which only those asymptomatic patients with a positive biomarker result receive definitive testing, with discussion of the issue of negative patients not receiving definitive testing.

Pepe et al, further discuss a design entitled the PReBE design (prospective-specimen-collection, retrospective-blinded-evaluation).\textsuperscript{210} In this paper they suggest the prospective
nested case-control design as a means of ensuring that biomarker development is rigorous and robust. In the first instance, specimens should be prospectively collected and stored from a randomly selected cohort reflective of the population in which the proposed biomarker would be used, prior to development of the outcome of interest. Cases and controls (i.e. those who do/do not experience the clinical event of interest) are then randomly selected, their specimens retrieved from storage and tested for the biomarker of interest, blinded to case/control status. The authors also suggest the necessary performance of the new biomarker be established in clinical context. The design has clear potential for application in a monitoring context, with the possibility of obtaining repeat biomarker measurements on a relatively frequent schedule, so that the best threshold and test interval could be determined from the data.

A further design proposed for the evaluation of a biomarker as a screening test, so potentially applicable for monitoring disease, is the paired design, where different screening strategies are introduced at different centres and screening performance is assessed by comparing the number of interval cases observed.²¹³

Analytic approaches to developing monitoring strategies

The literature in the area of monitoring focusses on modelling approaches (linear mixed modelling, joint modelling and non-linear modelling). The review of the literature has also identified some simulation studies and some work on the evaluation of monitoring strategies.

Linear mixed effects models and estimation of signal to noise ratio

Glasziou and colleagues question the need for randomised controlled trials of monitoring under certain circumstances, pointing to the need to understand the background variation and evaluate the signal to noise ratio when assessing treatment effects.¹⁸² They suggest large estimated treatment effects would be required to demonstrate an effect.

General description of models

Stevens et al review statistical models used for the control phase of monitoring and explain how models can be fitted to observed monitoring data providing details of maximum likelihood methods, moment-based methods and literature based methods, where parameter
estimates are obtained from reviewing the literature. They introduce a generic model for monitoring data, defining $Y_{it}$ as the observed monitoring values including assay noise and variability, and $U_{it}$ as the ‘true’ underlying and unobserved values. $U_{it} = \alpha_i + \beta_{it}$ and $Y_{it} = U_{it} + \omega_{it}$, where $\alpha_i$ is the true value at time 0, $\beta_{it}$ is the change in the true value over time and $\omega_{it}$ is random error.

**Signal and noise**

Modelling methods are used with repeated test data in the hope of distinguishing ‘signal’ from ‘noise’. The ‘noise’ is normal fluctuation in test results for patients (caused by the measurement variability of the test) and the ‘signal’ is a change in test results signifying a true change in disease state.

Thompson and Pocock present their findings following the analysis of repeated serum cholesterol measurements on 14,600 men and women. Their work focuses on the impact of within-individual variability on screening and monitoring. Using the cholesterol data a single observed measurement did not reflect the true underlying measure. Thompson and Pocock showed how the probability of a measure being classed as ‘high’ varied with the true underlying value, and whether the classification was based on a single measure or the mean value of multiple measures. The use of multiple measures was shown to improve classification. The authors identified regression to the mean when analysing multiple measures and variability in the measures for untreated individuals over time leading them to doubt whether repeated measures would be able to identify the benefit of treatment. The authors state the use of repeated measures could be ‘very discouraging’ for some patients.

Buclin et al define two decision rules that could be used to guide the treatment of patients with HIV infection with antiretroviral treatment based on CD4 cell measurements using a review of longitudinal analyses of CD4 cell trajectories. The first decision rule is a ‘snap-shot rule’—dependent on a single CD4 measure—and the other is a ‘track-shot rule’—where multiple CD4 measures are required. The devised rules are then tested using clinical data, with the view of minimising false findings, and recommendations are made regarding the frequency of testing.

Bell and colleagues developed a framework to identify when monitoring of initial response to treatment would be beneficial using data from RCTs. The findings showed monitoring of
initial response to treatment would only be useful when there is variation in treatment effect between patients and not all treated patients achieved results at the level targeted.\textsuperscript{183}

Other examples of the use of mixed modelling and signal to noise ratio estimation to understand when it is appropriate to monitor response to treatment, thresholds or monitoring frequency are monitoring of cholesterol, bone mineral density, blood pressure, lipids, and diabetes.\textsuperscript{101, 102, 183-190}

**Joint modelling of longitudinal and outcome data**

*Joint linear class models*

When fitting a joint linear class model subjects are split into a finite number of latent subgroups. The trajectory of biomarker measurements and the risk of event are specific to each latent class, meaning the joint linear class model allows for the dependency of biomarker values and risk of event. Biomarker measurements and time-to-event are conditionally dependent given latent class. More specifically, a multinomial logistic regression model is used to assign subjects to subgroups. A linear mixed model is then used to model repeated biomarker measurements given the assigned latent class of the subject; and, a survival model is used to model the time-to-event, again given the latent class of the subject. The model is fitted using maximum likelihood estimation.\textsuperscript{191}

Examples of the use of joint linear class models can be seen in work by Proust-Lima and Taylor and Li and Gatsonis, both applied to monitoring with prostate specific antigen (PSA) prostate cancer recurrence.\textsuperscript{192, 193}

Proust-Lima and Taylor discuss the derivation of a posterior probability of recurrence from a joint class linear model to identify a ‘dynamic prognostic tool of recurrence’.\textsuperscript{192} The posterior probability obtained from the joint linear class model gives the probability of an event occurring between time $s$ and time $s + t$ (where the subject is event free at time $s$). Estimating the probability of an event after a certain time requires fitting survival models to subjects at each time being estimated with only covariates available at time $s$. As biomarker data is often discrete imputation techniques are used to allow predictions of an event to be obtained at multiple time points. Proust-Lima et al also discuss the validation of predictive tools and the lack of consensus in this area.
Li and Gatsonis use a joint linear class model to develop a strategy that modifies monitoring intervals.\textsuperscript{193} Li and Gatsonis use a two-stage approach when fitting the joint linear class model, where the model used to identify latent classes is fitted separately. Bayesian Information Criteria is used to select the number of classes. The two-stage approach has the advantage of being less computationally intensive and the EM-algorithm can be used to estimate parameters at each point of monitoring. The uncertainty of latent class assignment is evaluated using multiple imputation assuming latent class is missing completely at random. For prospective studies the two-stage procedure is repeated as new information is collected (measures, events and study end). Li and Gatsonis demonstrate the method using simulated PSA measurements for 150 patients with prostate cancer with testing to identify recurrence. Predictions from the model inform a utility function which is used to identify the appropriate monitoring intervals for each patient. The expected value of the utility function used is $U(t) = a \cdot P(\text{event at time } t)$, where $a$ is a negative value if the event occurs and zero otherwise. The optimal monitoring interval can be identified for individuals or groups of patients; the authors advocate optimising by latent class as these intervals can then be adapted for new patients.

**Bayesian hierarchical change-point models**

Bayesian hierarchical change-point models model the trajectory of test results prior to the onset of disease, the onset of disease and the trajectory of test results after the onset of disease simultaneously. Bayesian hierarchical change-point models also allow for the within-individual correlation, as individuals have multiple test measurements, between-subject variation in trajectories and the random change-point. Bayesian hierarchical change-point models use a piecewise or segmented linear model where the parameters of the model are the trajectory of test results prior to the change-point, the test result value at the time of the change-point, the time of the change-point and the trajectory of results after the change-point; each of the parameters is a random effect within the model. Non-informative prior distributions are used for the parameters in the model; with the parameters describing the distributions of the parameters used the model being drawn from non-informative prior distributions.\textsuperscript{194, 195}

Slate and Turnbull discuss and demonstrate the use of Bayesian hierarchical change-point models by analysing PSA data from the Nutritional Prevention of Cancer Trial (NPCT).\textsuperscript{194} Slate and Turnbull state the advantages of using Bayesian hierarchical change-point models
are the ‘borrowing of strength’ when estimating parameters specific to individuals whilst also accounting for the correlation of measures and, by obtaining posterior distributions using Gibbs sampling, the model can give the probability an individual has reached the change-point.

Bellera et al also demonstrate the use of Bayesian hierarchical change-point modelling using PSA data. Bellera and colleagues state additional advantages of this type of modelling are the ability of the model to provide precise estimates compared with simpler models, the parameters used by the model are all of clinical importance, estimates of test measurement variability can be estimated as a function of test result value, and the model is flexible and can be easily adapted. Bellera and colleagues do, however, comment that the model can be influenced by the timing of and the number of test measurements for individuals with the potential for this to cause bias, as participants with more results will provide more information for the model and participants with more test results may be different to those with fewer test results. Subsequent work by Bellera et al uses an empirical simulation approach with Bayesian hierarchical change-point modelling to evaluate and compare different rules used in detecting the recurrence of prostate cancer based on PSA measures. Bayesian hierarchical change-point models were used to identify if the rules used in practice were able to adequately classify patients with real progression of PSA values and those with stable PSA.

Inoue et al combined longitudinal PSA measurements from three different studies using a non-linear Bayesian hierarchical model. At the individual level a non-linear model is used to model PSA over time and the hierarchical model component then accounts for the variability between studies.

**Non-linear mixed models**

Non-linear mixed models allow more flexibility in modelling as linearity of the parameters is not necessary which may be appropriate for modelling longitudinal test data in some conditions. Multiple measures for each individual can also be accounted for by non-linear mixed models with the incorporation of random effects.

Examples of the use of non-linear mixed models can be seen in work by Subtil and Rabilloud and Taylor et al.
Alternative modelling approaches

Alternative modelling approaches are used by Thiébaut et al and Wolbers et al in the area of CD4 cell monitoring to understand when to initiate treatment for patients with HIV.\(^\text{199, 200}\)

Machine learning methods

We have considered machine learning methods. These arise from, essentially, a set of pattern recognition processes or algorithms. They are primarily concerned with non-continuous data, such as x-ray images, or electronic health records, and for machine learning methods to provide high-quality predictions these pattern recognition processes would usually be applied to large data sets.\(^\text{256, 257}\) The biomarker data is continuous, and there is a limited amount of data available. The statistical analysis and modelling approach also enables testing of different trial design strategies, as discussed extensively in chapter 7. Our conclusion was that it is not clear how this would be achieved through machine learning methods. Machine learning methods are still in the developmental stage, and may become useful in the field of biomarkers in the future, as the field develops. There are still issues to do with such methods not always providing estimates of uncertainty, and being hard to interpret. The logic and mathematics to underpin these processes are still being developed, and are not trivial. The task of modelling these data sets is very complex, and requires many assumptions, which need verifying, as for statistical approaches.\(^\text{256, 258}\) Use of machine learning methods is currently limited by the lack of availability of simple off-the-shelf application packages.

Monitoring simulation studies

With knowledge of the progression of disease and the variability of the test used to monitor the disease, data can be simulated allowing the evaluation and comparison of decision rules and testing frequency. Simulation approaches are used by Sölétermos et al and Bellera et al.\(^\text{201, 202}\)

Evaluation of monitoring strategies

How the performance of a monitoring strategy is measured will be different and more complex compared with testing at a single time point, due to repeated testing and the potential for patients to change disease state. DeLong et al discuss sensitivity and specificity
for monitoring testing and Li and Gatsonis provided guidance on the evaluation of monitoring strategies.\textsuperscript{193, 203}

**Pooled analysis of prospective cohort studies**

In some disease areas data sets from multiple cohort studies exist and it is possible to combine the data and analyse the pooled group of patients. With information on the patients within the cohort studies it is possible to use the combined data as a proxy for trial data. The cohort approach is used to evaluate the appropriate CD4 cell level at which to begin antiretroviral treatment for patients with HIV by The When To Start Consortium (using a method introduced by Cole et al) and Ahdieh-Grant et al.\textsuperscript{204-206}

**Results – Screening literature**

The aim of screening is to benefit patients by detecting disease prior to the onset of symptoms as is the case with monitoring. The detectable pre-clinical stage of disease is the time when screening may detect asymptomatic disease; this is also known as the sojourn time. The delay time is the period of the sojourn time when the screening has not detected disease and lead time is the period of sojourn time after screening has detected disease. The greater the lead time, the greater the potential benefit of screening.

Walter and Day discuss the biases that need to be considered when analysing screening data. Firstly, the population that participate in screening may vary from the population that do not participate in screening, as they may be at higher or lower risk of having the disease the screening process aims to detect.\textsuperscript{217} This is likely to be less of an issue for monitoring populations, although it is conceivable that there will be differences between those who do participate in monitoring and those who either drop out (perceiving themselves to be at low risk of the event in question) or who demand some form of treatment (perceiving themselves to be at high risk of the event in question). Other biases that may affect monitoring studies include length-biased sampling and lead time bias. Length-bias sampling occurs as patients with more aggressive disease will be in the pre-clinical phase of disease where screening will detect disease (sojourn time) for a shorter length of time than those with less-aggressive disease. Screening is most likely to detect cases with a longer sojourn time, hence cases of less-aggressive disease which will likely have a better prognosis. Lead time bias is when survival times for screened cases appear to be greater than survival times for cases identified
by different means when there is no difference in survival; the only difference is cases identified by screening are detected earlier.

There is a body of work in the area of screening which focuses on estimating the duration of the pre-clinical stage of disease (Walter and Day, Day and Walter and Etzioni and Shen), which enabled further work into the optimal frequency of screening (Zelen, Lee and Zelen, Frame and Frame and Lee et al.223).217-223 Others have considered how to set the optimal decision rule for a screening strategy using a new test when the length of the sojourn period is not known (McIntosh and colleagues and McIntosh and Urban).224, 225

**Results - Biomarker development process**

**Methods for analysis and study design used in the biomarker development process**

The methods for analysis and design of studies of biomarker development are discussed by Baker et al, Pepe et al and Sturgeon et al.209, 210, 259 Baker discuss a method for evaluating multiple biomarkers for selection for further study.212 Baker and colleagues, Baker, Ransohoff and Ransohoff and Gourlay discuss the issues with biomarker development studies.1, 209, 213, 214

Sturgeon et al.208 provide information on the biomarkers that have been developed to identify cancer and the extent of their use in practice.

**Results - Time-dependent ROC curves**

When a test gives a binary result (positive or negative) the performance of the test is usually assessed by calculating the sensitivity and specificity. When the result of a test is a continuous value the performance of the test is evaluated for various cut-offs by calculating the sensitivity and specificity of the test at each possible result value and plotting sensitivity against 1-specificity, a receiver operator characteristic (ROC) plot. The ROC plot and the areas under the curve produced can then be used to assess the performance of the test and identify optimal thresholds for the use of the test in practice. When allowing for time in ROC analysis, time-dependent ROC methods are used.
Sensitivity and specificity

Pepe and colleagues undertook a review of time-dependent ROC curves. The definition of sensitivity of a test is dependent on the time when the test is performed. As it is assumed diseased cases will present with positive test values early in the testing process it is thought sensitivity will decrease with time. Pepe et al also discussed cumulative sensitivity, which would provide the sensitivity for a test for an interval of time, and how this can be derived. The false positive fraction, or 1-specificity, is problematic to define as the disease status of individuals can change over time making it difficult to classify individuals as diseased or non-diseased, especially in situations where all individuals will have an event at some point. One approach is to choose a time point specific to the context assessed and individuals are treated as non-diseased if they are free of the event at the specified time (the static false positive fraction). Another approach is to allow the false positive fraction to vary with the time since the test was performed (the dynamic false positive fraction). When using the dynamic false positive fraction test performance may be misleading as a positive result will be falsely positive shortly before an individual develops disease. For tests with continuous results time-dependent ROC curves compare individuals with and without disease at each time point. If using the dynamic false positive fraction ROC curves are difficult to interpret due to the non-diseased group changing over time.

Cai et al present equivalent time dependent definitions of sensitivity and 1-specificity, but with the emphasis on the time of an event occurring, defining sensitivity and 1-specificity to be functions of time relative to the time of disease or time of an event. The authors state most research in the area assumes the test and assessment of disease status are carried out simultaneously raising the issue of the predictive accuracy of a test being dependent on the time it is carried out in comparison to the onset of disease, assuming an increase in accuracy if the test is used closer to the time of an event. Cai and colleagues also fit semi-parametric models using longitudinal test data to separately estimate sensitivity and 1-specificity. Zheng and Heagerty discuss sensitivity and specificity for time-dependent ROC analysis as functions of both the time of testing and the time of event. Zheng and Heagerty also discuss the difference between estimating incident and prevalent ROC curves, restricting their work to incident ROC curves.
Subtil et al discuss how incident sensitivity requires a test to be performed a given number of days prior to the onset of disease and offers a way of taking the variation of time between individuals receiving a test and developing disease into account. Subtil and colleagues introduce a Bayesian method to allow for the interval-censored measurements. Results of using this method compared with the method without adjustment suggest the ‘crude’ method underestimates sensitivity.

Parker and DeLong provide a method to convert estimates of sensitivity and specificity for monitoring tests for ROC curve analysis. The estimates of sensitivity and specificity used are those introduced by DeLong et al, which are derived using partial likelihood estimation under the assumption that diseased participants can have at most one test result when in the diseased state.

**Modelling to produce time-dependent ROC curves**

Slate and Turnbull review methods used to analyse repeated test data when the test is used to screen or monitor the onset of disease in a population. These methods are used to estimate the ROC curve for each test and the resulting ROC curves are compared. The review discusses the use of time dependent Cox proportional hazards modelling, joint modelling of longitudinal test data and time of diagnosis, Weibull methods to model two time events, random effects models and integrated Onstein-Uhlnbeck (IOU) stochastic processes, multi-state models and Markov models, and change-point models.

Zheng and Heagerty discuss a semi-parametric regression approach used to estimate ROC curves and an approach based on asymptotic distribution theory, which will allows covariates to change the distributional shape of test results.

Etzioni et al introduce and demonstrate two methods for modelling the effect of lead time on the ROC curve. The first approach requires modelling of longitudinal test data and then using parameter estimates from the model the ROC curve can be estimated at varying time points. The second approach directly models the ROC curve as a function of covariates including time of the test relative to the time of diagnosis. Etzioni and colleagues discuss how the methods can be adapted to compare two tests. The first method requires separate fitting of models using data for the two tests with comparison of the derived ROC curves after; whereas, the approach of modelling the ROC curve directly more easily allows for
comparison of tests and allows the difference between tests to be tested. Other advantages of the direct modelling approach are: fewer distributional assumptions with the method using the ranking of data points, robustness and flexibility and ease of implementation.

**Results – Differentiating measurement change from measurement variability**

**Variability—reference change values and coefficients of variation**

The variability of repeated test measures for an individual can be broken down into three components: pre-analytical variability, analytical variability and individual variability. Analytical variability is the variation in result due to the laboratory test, and individual variation is the within patient variability. Pre-analytical variability can be minimised by ensuring collection and storage of the samples is carried out in a standard way, and so is regularly not used when calculating total variability of successive measurements. Analytical and individual variability are combined using $SD_T = \sqrt{SD_A^2 + SD_I^2}$, where $SD_A^2$ represents analytical variation, $SD_I^2$ represents individual variation and $SD_T^2$ represents the total variation of repeated measures for an individual. The coefficient of variation (CV) is calculated by dividing the standard deviation by the mean and is commonly used in place of standard deviations as this allows for a reference change value (RCV) to be calculated to reflect percentage changes rather than absolute changes. Normality is assumed and the RCV is given by $\sqrt{2Z \sqrt{CV_A^2 + CV_I^2}}$, where $CV_A$ is analytical variation expressed as a CV, $CV_I$ is the within-individual variability expressed as CV, and $Z$ refers to the z-statistic (value from the standard normal distribution). Given the values of analytical and within-individual variation, a difference between two results greater than the RCV suggests a real change in condition.\(^\text{232}\) Sölétormos et al use this rule in a computer model for monitoring of progression to metastatic breast cancer with cancer antigen 15.3 (CA15.3), carcinoembryonic antigen (CEA) and tissue polypeptide antigen (TPA).\(^\text{233}\)

The use of RCVs are further discussed by Smellie, Petersen, Fraser, Fraser et al, Klee, Petersen et al and Omar et al.\(^\text{234-240}\) Biosca et al report a study of biological variability to identify the appropriate RCV to use in their specific clinical situation, and Clerico and Emdin highlight differences in analytical sensitivity across studies carried out in differing populations.\(^\text{241, 242}\)
Statistical process control and statistical rules for interpretation of sequential tests

Statistical process control methods (first developed by Shewhart) are often used in manufacturing and can be used for medical applications when a process can be measured directly or via a biomarker. Statistical process control procedures measure variability across time, where variability can be split into common cause and special cause variability (or assignable cause variability). Special cause variability is akin to signal and signifies true change in the disease state of an individual. Common cause variability, as noise, reflects random variability in measures.\(^\text{232}\)

X charts are used to display measurements over time for an individual. If a process is stable, measurements are expected to fluctuate around the mean, and the standard deviation of observed measures is expected to be constant over time. Estimates of the mean ($\mu$) and standard deviation ($\sigma$) can be taken from stable processes, with an unbiased estimate of the standard deviation obtained using a moving range (the difference between consecutive measures) and dividing the mean of the moving range estimates by a constant ($d_2=1.128$). Estimates of the mean and standard deviation of a stable process can then be used to identify control limits. The control limits can be identified using many criteria and should be modified depending on the situation; it may be target values are safety driven. Moving range charts and exponentially weighted moving average charts (moving averages are calculated with greater weight given to the most recent observations) are also used in similar ways. The variability of a process can be quantified using the capability index, the difference between the upper and lower limit divided by $6\sigma$. The off-target ratio, $S_T = (\mu - T)/\sigma$ where $T$ is the target value, measures how far the process is from the specified target value in terms of standard deviations. Process control charts use the assumption of independent normally distributed outcomes and generally require at least 20-25 observations.\(^\text{232}\)

Tennant et al review studies where patients are monitored using statistical process control methods and compare the use of statistical control methods with currently used rules and guidelines.\(^\text{243}\) Clinical areas found to use process control methods are peak flow measurements for patients with asthma, blood pressure measurements for patients with hypertension and serum creatinine measurements for patients after undergoing a kidney transplant. Thor et al also review studies using statistical control processes to monitor patients and highlight the disadvantages of using these methods.\(^\text{244}\) Thor and colleagues discuss that in
some studies methods had been employed where there was a clear lack of understanding. The authors also comment on issues with auto-correlated measures, collection of data and application of the methods.

Gavit et al discuss a slightly different approach to process control in change point analysis. Change point analysis uses cumulative sum charts of the difference between the mean value and the recorded value. Change points are then analysed as bootstrapping methods are used to generate a confidence interval for the change point. The change point method can also be used to identify differences in variability. An advantage of the change point method is the ability to analyse non-normal data due to the lack of distributional assumptions. Gavit and colleagues also claim the change point method is able to identify subtle changes that would not be picked up by control charts.

Results – Health economic approaches

Decision analytic models

Decision analytic modelling evaluates the cost, outcomes and cost effectiveness of interventions. In the case of repeated testing appropriate techniques need to be used for this evaluation.

Karnon et al review models for measuring the cost effectiveness of screening regimes, with the Baker and Parmigiani approaches featuring. Sutton et al introduce comprehensive decision modelling.

Real options approaches

Palmer and Smith introduce real options approaches, inspired by methods used in financial markets, which aim to include the uncertainty around the use of a new technology along with health economic evaluation. The approach uses the potential to delay introducing a new technology (akin to a change in management) and the irreversibility of using a new technology. Analyses factor in deferring using a technology and the better evidence that may be available after deferral using expected value of perfect information methods (EVPI).
Real options approaches are further discussed and expanded on by Driffield and Smith, Meyer and Rees and Shechter et al, whilst Whynes and Lasserre et al discuss a similar method. 251-255

**Summary and conclusions**

This review has revealed limited methodological literature around the design of monitoring strategies. Work has focussed primarily on analysis of data where subsequently recommendations of monitoring frequency or decision rules could be made or simulation work, with both approaches being specific to the disease area researched. The area of screening has developed methods with the focus being identifying the optimal frequency of screening which could be used for designing monitoring strategies. There is some work on the design of biomarker development studies which could potentially be adapted to allow for the evaluation of a monitoring strategy using previously collected specimens. It appears thresholds are often developed by analysis of the variability of the test being used, identified by the literature describing signal to noise ratio, biomarker development studies, statistical process control and reference change values.

The study by Buclin et al shows an approach where decision rules were devised by a review of the literature and then using an obtained data set and analysis of signal and noise the rules were refined to minimise false results.104 Following this, recommendations of the decision rule and frequency of monitoring could be made. Takahashi et al, Takahashi et al and Oke et al also use signal and noise methods when analysing data and subsequently recommendations can be made for future monitoring strategies.188-190

A number of applications of the signal and noise approach were identified, 12-23 largely in the area of treatment titration.182 The limitations of this approach for monitoring disease progression or recurrence are that rules and thresholds are devised purely by analysing the variability of test measures and the minimisation of false findings rather than detection of disease at the earliest point possible and the impact on patients.

The simulation approach proposed by Li and Gatsonis uses a joint linear class model which combines predictions from the model along with a utility function to identify optimal monitoring frequencies.193 The results of a simulation study reported by Li and Gatsonis
appear promising; however, the approach has not been widely adopted perhaps due to the complex nature of the model. Other simulation approaches may also have potential under certain circumstances, particularly if measurement error can be included, and a link between biomarker values and true disease state can be included.

The biases that are well documented in the screening literature are applicable to the area of monitoring also. Length-time bias and lead time bias should be considered when analysing monitoring data and when designing monitoring studies. There is also the issue of post-screening noise which is again important to take into consideration when evaluating a monitoring strategy; the time point at which monitored and non-monitored patients are compared should be selected to minimise the issue of incidence after the final testing point and should also consider the number of likely events. Harm to patients is vitally important in screening and monitoring as this harm may be at several time points and this must be thought of when designing strategies.

A further consideration for the analysis of monitoring data concerns the number of test measurements and the timing of test measurements: people with more results will contribute more data to the model but they may be very different to those with fewer results. Measurement error and particularly biological variability also requires consideration. Studies have shown that reference change values from biological variability studies of healthy participants are not necessarily reflective of the true RCV for a diseased population. As methods to derive test thresholds used in monitoring rely heavily on the variability of test results it is important that estimates from biological variability studies are accurate. Also, the quality of studies undertaken when developing new biomarkers is not always rigorous; however, there are new methods and guidelines for reporting that hope to improve quality.
Chapter 6 - How can monitoring impact on patient outcomes?
Much of the test evaluation literature centres on establishing key test properties such as test accuracy. However, the ultimate use of any test in clinical practice should be based on the knowledge that testing does more good than harm to patients. Comparison of patient outcomes resulting from different interventions is ideally assessed using a randomised controlled trial (RCT) design, and the same design can be applied to evaluations of tests. RCTs are less commonly used for assessing medical tests but are increasing in number both for diagnostic and for monitoring tests (see Chapter 4), such that a thorough understanding of the ways in which testing can affect patient outcome is important.

Patient monitoring is undertaken for many purposes, most obviously within the context of ongoing treatment as the main tool for treatment titration and maintenance; the goal being to maintain test results within certain limits of a given marker until such a time as treatment can be discontinued or an alternative treatment needed. Our particular interest is in monitoring people who have a known disease or condition that is likely to progress or recur at some point in the future but that does not yet require treatment. Patients are usually asymptomatic (for example, following primary treatment for the first occurrence of a disease), but may be mildly symptomatic but not yet receiving treatment, or may experience symptoms of a disease that puts them at risk of developing other conditions. The primary goal is usually earlier treatment or the avoidance or delay of treatment, the crux of monitoring being to detect the need for a change in patient management in a timely manner.

Particular challenges to evaluating the impact of monitoring tests on patient outcomes are: firstly, that the effect on outcomes will be relatively small, thus requiring large samples of patients to demonstrate statistically significant effects; and secondly that changing patient outcomes is reliant on patients and clinicians following potentially complex protocols both for testing and for treatment.

Over the last 10-15 years a number of framework papers related to the development and evaluation of tests for screening, diagnosis, prognosis and treatment monitoring purposes have been published, many of which have been comprehensively reviewed by previous authors. We have selected three frameworks of particular relevance to the consideration of patient outcomes in monitoring. The first, by Adriaensen and colleagues, presents a stepwise evaluation process for new screening strategies, which includes a consideration of the trade-off between the harms and benefits from a new test. The second, by Ferrante di
Ruffano and colleagues, aims to assist those evaluating diagnostic tests to understand the ways in which changes to testing strategies can affect patient outcomes. The third, by Lord and colleagues, considers the circumstances in which randomised evidence of patient impact from a new diagnostic test may be needed. With these in mind, our aim was to consider the potential impact of monitoring on patient outcomes, illustrated by our review of randomised trials of monitoring strategies.

**Methods**

A monitoring care pathway was outlined to identify in simple terms the points at which monitoring might affect outcomes (Figure 10). The three identified frameworks were reviewed in terms of their relevance to this monitoring context and trials from a review of 58 RCTs in which monitoring was carried out in at least one arm of the trial (see section 3) were used for illustration purposes. The trials were grouped into three main categories in terms of the change in patient care under evaluation and the intended impact on patient outcome:

i. a new monitoring strategy versus an existing monitoring strategy, such that the current monitoring test might be replaced by a new and more accurate test, a new test might be added to the strategy, or the currently used test applied at a different intensity or with an alternative threshold for intervention. Depending on the associated change in patient care, the new monitoring strategy may be intended to detect patients at an earlier stage of disease, to more accurately detect those in need of treatment or to reduce the invasiveness or frequency of testing.

ii. a monitoring strategy versus immediate treatment of all patients at risk of an adverse outcome, where monitoring may be used to avoid or delay treatment in those who do not need it.

iii. a monitoring strategy versus no monitoring, where patients are usually treated on the basis of clinical presentation only and the likely aim of monitoring is the detection and treatment of disease at an earlier stage.

In the following sections, we first consider the similarities and differences between monitoring, screening and diagnosis, before broadly outlining the potential for benefit and harm from monitoring, and considering the ways in which patient outcomes can be mediated by particular aspects of the monitoring care pathway according to the aim of monitoring and the change in strategy under evaluation.
Figure 10 Monitoring care pathway (adapted from Ferrante di Ruffano et al, 2012 with permission.68)

a) Pathway

b) Detail of ongoing monitoring process

Monitoring versus screening or diagnosis

The monitoring care pathway outlined in figure 10 bears close resemblance to that for diagnosis and for screening.68, 262 In a monitoring context, (1) a test is administered according to a predetermined schedule to detect a target condition or some precursor or marker of that condition, (2) the test result is considered (often in relation to previous measurements and
with the potential for repeat testing to confirm abnormal or indeterminate results), (3) alongside other evidence (usually including the results of further investigations) to decide whether therapeutic intervention is needed, and (4) the necessary intervention is implemented.

Where the pathway diverges from that for diagnosis is with the added dimension of repeated testing over time, and a merging of the ‘diagnostic’ and ‘management’ decisions outlined in the diagnostic care pathway by Ferrante di Ruffano and colleagues. The serial nature of testing for monitoring purposes can affect patient outcomes in a number of ways, most obviously by increasing the physical and psychological burden of testing on patients but also potentially impacting on other outcomes, for example via patient and physician compliance with testing protocols. Furthermore whilst a diagnostic test informs both a ‘diagnostic decision’ (often where more than one differential diagnosis may be available) and a ‘management decision’ (assisting in the choice of a range of therapeutic options), a monitoring test is often relatively less definitive, providing more of a guide to the need for changes in patient management. A positive monitoring test result frequently triggers further investigation to determine whether and when a particular treatment should be implemented, rather than informing the choice of one of a range of therapeutic options. In this respect, monitoring is more akin to screening, where a test is applied repeatedly over time to detect and treat a particular clinical condition, rather than to differentiate between diagnoses, with the caveat that monitoring populations have a higher risk of the clinical event of interest occurring and that whilst the ultimate goal of a new screening programme is usually a reduction in disease-specific mortality, monitoring can be implemented for a range of reasons.

Potential benefits and harms from monitoring

Figure 11 uses the concept of a 2x2 contingency table to illustrate that assigning potential benefits and harms from a monitoring test is not as straightforward as might be imagined. For simplicity the following is set mainly in the context of a new monitoring strategy to allow earlier detection and treatment.

In general terms, for those patients with a ‘true’ result, benefits accrue both to those who would otherwise have experienced a poor clinical outcome but for the new test (A) and to
those who would have been detected clinically and successfully treated but the new test allows this to happen at an earlier point in the disease process (B). Benefit also occurs for those with no disease and for whom a negative monitoring test result has a reassurance value, increasing a patient’s sense of control over the disease (G). Positive benefits might also be experienced by all patients, for example with the use of a less invasive test, or less frequent testing (H).

Patients who have ‘false’ monitoring test results will experience harm from the new testing strategy in a similar way to false negative or false positive diagnostic tests. False negative test results can lead to a false feeling of security, delayed detection of disease and potentially delay of effective treatment until the disease becomes clinically apparent (D). False positive test results can lead to unnecessary further investigation and/or unnecessary treatment (E). For monitoring tests that aim to detect preclinical or very early stage disease, “early” false-positive results (i.e. in patients whose disease would not have progressed to clinically overt disease within a clinically meaningful timeframe) will lead to a longer period of time in a diseased state and potentially in overdiagnosis and unnecessary treatment (F). Similarly, ‘early’ positive monitoring tests in those with true positive test results (C) can cause harm for those patients who go on to experience a poor clinical outcome and who undergo a longer period of treatment (with associated side effects and a longer period of time in a ‘diseased’ state).

More direct harms can also be incurred by the testing experience, relating to repeated applications of the monitoring test, to any confirmatory testing and to any intervention that is implemented (I and J). These can be physical or psychological in nature, and may result from either positive or negative test results; the serial nature of testing in a monitoring context necessarily multiplying the potential impact. In particular, the ongoing monitoring process (even with repeated negative results) may raise general levels of anxiety and distress and can also have a ‘labelling’ effect that can have a negative influence on patients’ perceptions of themselves and their disease.265, 266
Figure 11 Summary of potential harms and benefits from a new monitoring strategy*

<table>
<thead>
<tr>
<th>True positives</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Would have experienced clinical outcome (e.g. death) but are cured, due to (earlier) detection and effective treatment (A)</td>
<td>- Do not have the disease, or any precursor, and undergo unnecessary further investigation and treatment (E)</td>
</tr>
<tr>
<td>+ Would have been successfully treated for disease anyway, but quality of life is improved due to detection (at an earlier stage of disease), ± less debilitating treatment (B)</td>
<td>- Have preclinical or early stage disease that would not have progressed to clinically overt disease within a 'reasonable' timeframe (or could potentially even have regressed), resulting in overdiagnosis and unnecessary treatment and a longer period of time in a 'diseased' state (F)</td>
</tr>
<tr>
<td>- Would have experienced clinical indications of disease at a later time point, but clinical outcome not improved and quality of life potentially decreased by earlier detection and treatment (C)</td>
<td>+ Do not have the disease or preclinical indicator of disease and are reassured by the negative results of a monitoring test that correctly shows that they do not have the disease (G)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>False negatives</th>
<th>True negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Have the disease (or a progressive precursor of disease) but have a negative monitoring test resulting in a false feeling of security, delayed detection and delay of effective treatment (D)</td>
<td>+ Experience benefit from the monitoring experience, either psychologically or due to less frequent or less invasive testing (H)</td>
</tr>
<tr>
<td>+ Experience direct harm from the monitoring test(s) or from any confirmatory testing (I)</td>
<td>- Experience direct harm from the monitoring test(s) or from any confirmatory testing (I)</td>
</tr>
<tr>
<td>- Experience psychological impact from increased anxiety or labelling effects (J)</td>
<td>- Experience psychological impact from increased anxiety or labelling effects (J).</td>
</tr>
</tbody>
</table>

*Harms and benefits are similar to those identified from a screening context in Adriaensen and colleagues."
When is randomised evidence needed?

The randomised controlled trial is the gold standard approach to assessing impact on patient outcomes but given the challenges to implementing the design for the assessment of test impact, its use requires careful consideration.

Lord and colleagues determined the need for randomised evidence for a new diagnostic test based firstly on whether the cases detected by the new test represent a similar spectrum of disease to those detected by the old test and secondly, whether treatment has been shown to be, or can be assumed to be, as effective in the new group of patients or regardless of disease spectrum. Incorporating the time dimension of monitoring into this framework: for a new monitoring strategy one must determine firstly, whether the cases detected by the new strategy represent a similar spectrum of disease, both in terms of the biological characteristic that is measured by the test and in terms of the time point in the disease process that disease recurrence or progression is identified; and secondly, whether treatment has been shown to be, or can be assumed to be, as effective in the new group of patients, regardless of disease spectrum and timing of detection in relation to stage of disease.

Notwithstanding the simple appeal of this approach, testing strategies are necessarily complex interventions, with various components and possible interactions that can combine to affect patient outcomes; even a ‘perfect’ test and highly effective treatment will not necessarily improve patient outcomes. The Ferrante di Ruffano framework Table 11 outlines 14 mechanisms by which testing can affect health outcomes which can be broadly considered in five categories: timing, test properties, treatment effectiveness, potential to change practice, and the patient experience. Many of these can apply equally to patients undergoing monitoring, however, some require more or less emphasis or need to be adapted to the monitoring context.

We consider the extent to which these factors might affect a monitoring evaluation according to the aim of the new strategy and the way in which it fits with standard care, in order to demonstrate how this could inform the need for randomised evidence (Table 12).
Table 11 Patient outcome framework for monitoring tests (adapted from Ferrante di Ruffano et al\textsuperscript{68})

<table>
<thead>
<tr>
<th>Care Pathway Component</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Delivery</td>
<td></td>
</tr>
<tr>
<td>1 Test feasibility</td>
<td>Completion of the test process, where reasons for non-completion might include: (a) Counter-indication (clinician refusal to administer test) (b) Technical failure (ability of test equipment to produce data)</td>
</tr>
<tr>
<td>2 Test procedure</td>
<td>Patients’ interaction with the test procedure, potentially causing physical or psychological harms or benefits.</td>
</tr>
<tr>
<td>3 Test frequency</td>
<td>Patients’ response to serial testing, potentially multiplying the impact of any physical harms and incurring additional psychological impact.</td>
</tr>
<tr>
<td>Test Result</td>
<td></td>
</tr>
<tr>
<td>4 Interpretability</td>
<td>After successful completion of the test process, the likelihood of high frequencies of indeterminate or unreadable test results (distinct from the measurement variability associated with an individual test result which will affect the ability of the test to detect true changes in disease status).</td>
</tr>
<tr>
<td>5 Clinical validity</td>
<td>The ability of a test to predict the presence of, or development of, clinical or overt disease.</td>
</tr>
<tr>
<td>6 Timing of test result</td>
<td>The ability of the test to predict overt disease within a clinically meaningful timeframe</td>
</tr>
<tr>
<td>7 Detection of long-term change</td>
<td>The ability of a test to differentiate true changes in patients’ disease status from short-term variations.</td>
</tr>
<tr>
<td>Management Decision</td>
<td></td>
</tr>
<tr>
<td>8 Added clinical value</td>
<td>The degree to which the test contributes to a change in management (\cdot) Indication for treatment (\cdot) Indication for further confirmatory testing (\cdot) Indication for closer monitoring (\cdot) Indication for less frequent monitoring [Also incorporates any other information used by a clinician to formulate a change in management (such as prior or additional test results).]</td>
</tr>
<tr>
<td>9 Timeframe of management decision</td>
<td>The timeframe within which patients undergo a change in management</td>
</tr>
<tr>
<td>10 Clinical confidence</td>
<td>The degree of confidence clinicians have in the validity or applicability of a test result.</td>
</tr>
</tbody>
</table>

Treatment Implementation
<table>
<thead>
<tr>
<th></th>
<th>Timing of treatment</th>
<th>The timeframe within which patients receive treatment.</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Efficacy</td>
<td>The ability of the intervention to improve patient outcomes at the particular stage of disease detected.</td>
</tr>
<tr>
<td>13</td>
<td>Adherence</td>
<td>The extent to which patients participate in the management plan, as advised by their physician, in order to attain the therapeutic goal.</td>
</tr>
</tbody>
</table>
Table 12 Analysis of the need for randomised evidence of new monitoring strategy (adapted from Lord and colleagues\textsuperscript{263})

<table>
<thead>
<tr>
<th>Goal of strategy Comparison</th>
<th>Change in testing</th>
<th>Key effects on patient outcomes</th>
<th>Example</th>
<th>Is an RCT necessary?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earlier detection and treatment</td>
<td>Add or replace test</td>
<td><strong>Timing</strong>&lt;br&gt;Test properties - limit early FP results&lt;br&gt;Rx effectiveness - at earlier stage of disease?&lt;br&gt;Added clinical value – what does test or Δ in threshold/freq add to clinical decision making?&lt;br&gt;Pt experience - Test more/less invasive?</td>
<td>Addition of PET scans at 3 and 15 mos to standard surveillance for detection of recurrent colon or rectal cancer (includes CT scans at 3 and 15 mos)\textsuperscript{267}</td>
<td>Longitudinal studies could identify additional cases detected by new test if management decision made after standard surveillance compared to management decision following new test or if ethical to blind result of additional (interim) investigations and act only on those carried out according to original FU schedule. RCTs needed to determine effect on patient outcomes from earlier treatment if randomised evidence does not already exist</td>
</tr>
<tr>
<td>New monitoring vs existing monitoring strategy</td>
<td>Δ existing test threshold</td>
<td></td>
<td>Surveillance and treatment of patients with HIV infection at higher CD4 threshold\textsuperscript{157}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>More frequent testing</td>
<td></td>
<td>Increased frequency of Doppler US for detection of HCC in patients with compensated cirrhosis of the liver\textsuperscript{165}</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{1} Note for this example, a collaborative analysis of 18 cohort studies from the When to Start Consortium\textsuperscript{31} supported the higher threshold for ART initiation.
<table>
<thead>
<tr>
<th>Monitoring vs no monitoring</th>
<th>New testing strategy</th>
<th>Timing</th>
<th>Test properties</th>
<th>Rx effectiveness</th>
<th>Added clinical value</th>
<th>Pt experience</th>
<th>Pt experience</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>limit FP results</td>
<td>at earlier stage of disease?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New monitoring vs existing monitoring</td>
<td>Replacemen t test</td>
<td>Timing</td>
<td>should be similar to existing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triage test</td>
<td></td>
<td>Timing</td>
<td>potential increase in time to treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduce volume of testing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Introduction of new endoscopic surveillance compared to endoscopy on demand to allow earlier detection of oesophageal carcinoma in patients with Barrett’s oesophagus. Akin to screening context; randomised evidence needed to determine benefit from formalised monitoring.

Reduce invasiveness of testing

Less invasive biopsy approach for patients at risk of colorectal cancer. No. Unlikely to be change in disease spectrum. Accuracy/predictive ability of new biopsy approach is key; value of the new test may be inferred from assessment of safety and/or cost.

Gene expression profiling as triage to endomyocardial biopsy to detect cardiac transplant rejection. Yes. Only subset of patients will be identified by new test. Rx will be as effective in subset but clinical outcomes unknown in those not selected by new test.

Reduce volume of testing
| New monitoring vs existing monitoring | Fewer tests or less frequent testing | **Timing** - potential increase in time to treatment  
Test properties  
**Rx effectiveness** – possible Rx at later stage  
**Added clinical value** - Pt/clin confidence in less testing  
**Pt experience** – possible psychological impact | Reduction in CT scans from 5 over 36 mos to 2 in 12 mos for detection of recurrence of non-seminomal testicular cancer \(^{271, 272}\) | Yes. Beyond 3 months FU spectrum will change to later stage of disease, especially for those whose disease has not recurred by 12 mos. Rx options/effectiveness likely similar for those detected in first 12 months. Longitudinal study could indicate number of positive tests that would be missed by reducing test frequency but could not compare clinical outcomes between strategies. |
|----------------------------------|-----------------------------------|--------------------------------------------------------------|-----------------------------------------------------------------|
| Reduce over-treatment            | Addition or replacement test;     | **Timing**  
Test properties - Similarly predictive with fewer FPs  
Rx effectiveness – does test detect same marker?  
**Added clinical value** - Clinician confidence key  
**Pt experience** – Impact depends on nature of new test | DNA-based test versus antigenemia test to allow pre-emptive instead of prophylactic Rx of HCMV infection following solid organ transplant \(^{159, 272, 273}\) | No. The two tests reflect different aspects of virus replication but apparently no indication of differential treatment response in this example. RCT needed only if treatment on the basis of the new test is expected to incur a different treatment response. |
| Delay/avoid treatment by introducing new surveillance | New testing strategy | **Timing** – Possible unnecessary delay in treatment  
Test properties - Should minimise FNs  
Rx effectiveness - Possible Rx at later stage of disease  
**Added clinical value** - Patient/clinician confidence key  
**Pt experience** – Increase testing but reduction in Rx | Sonographic surveillance vs immediate treatment of mild hip dysplasia in newborns \(^{65, 159, 160}\) | Yes. Spectrum changes to more severe disease as some cases resolve without treatment. RCT may be needed to establish whether later/no treatment leads to poorer outcomes. |
Earlier detection and treatment

Evaluations of monitoring strategies that aim to detect and treat disease at an earlier stage or time point will generally take the form of a new versus an existing monitoring regime or the initiation of monitoring where no previous monitoring was undertaken. In this setting, the goal of earlier detection almost necessarily implies that those patients detected and treated by a new monitoring strategy will have a different spectrum to those previously treated. Although patient outcomes may be affected by all of the identified mechanisms, it is the clinical validity of the test(s) used, its ability to detect long-term change within a clinically meaningful timeframe and the effectiveness of the treatment at the particular stage of disease, that are of overarching importance. Longitudinal studies can establish test properties and identify the spectrum of patients detected by the test. If no randomised evidence for treatment in this group of patients exists or if it is not clear whether the existing evidence will apply in the new group of patients, a new RCT may be needed, as for example in a trial evaluating a lower CD4 threshold for the initiation of antiviral treatment in HIV. Alternatively, the evidence may be such that no trial is indicated. A trial of ultrasound to detect small hepatocellular carcinomas (HCC) in patients with cirrhosis of the liver found that many of the lesions detected were too small to warrant treatment, with some even regressing rather than progressing. This high rate of “early false-positive” results could potentially have been identified in a longitudinal study without the need for an RCT.

Even where clinically valid, timely monitoring tests and effective treatments are available, the potential for a monitoring strategy to have a positive impact on patients will be influenced both by the degree to which it can substantively add clinical value over and above usual clinical practice and the degree of clinical and patient confidence in the new strategy. Where an individual monitoring test provides a clear guide to future management, as in the CD4 example above, the added contribution of the change in strategy might be relatively easy to discern. However, where a new monitoring test is one component of a bigger surveillance programme, for example the addition of biochemical tests and/or imaging tests to an existing surveillance programme, its added value may be more difficult to ascertain. In some circumstances, the accuracy of the confirmatory test could also mitigate any impact from the monitoring test, especially if the new test is able to detect very early stage disease that is not detectable by the confirmatory test.
Clinician and patient confidence in and compliance with a prescribed monitoring strategy can be vital to the success of a monitoring evaluation. For example, if clinicians have a high degree of faith in a new test, its ‘off protocol’ use in the control arm of a trial may dilute the observed effect. This was observed in the trial of ultrasound for the detection of HCC which also attempted to evaluate the added value of serial AFP measurements; high rates of serum AFP assay use in the two groups not randomised to AFP (60.5% and 54.8%, respectively) precluded reliable interpretation of the data and led to a final analysis restricted to ultrasound randomisation only.\textsuperscript{165}

The effect of the patient experience of monitoring on outcome will depend on the nature of the test involved and of the care that would otherwise have been received, especially if no monitoring was previously carried out. In the Tombola trial, for example, non-attendance was higher in the cytological surveillance arm: 10.6% did not attend first cytological surveillance vs 6.8% did not attend immediate colposcopy.\textsuperscript{270} Of the 10%, 2% did not attend at all and 8% attended more than 6 months after it was due. Some tests (or subsequent confirmatory testing) will carry a risk of immediate or long-term physical harm, for example, the introduction of routine endoscopic surveillance with biopsies in patients with Barrett’s oesophagus to allow earlier detection of oesophageal carcinoma, or the initiation of six monthly CT in patients at risk of colorectal cancer recurrence, with the potential to affect patient compliance.\textsuperscript{179, 268} Strategies that afford patients more control over a disease however, might increase adherence to treatment regimens, as in a trial of daily foot skin temperature monitoring by patients with diabetes where those who were compliant with the monitoring strategy for at least 50% of the time were significantly less likely to develop a foot ulcer.\textsuperscript{180}

Patients will also experience a sometimes complex psychological impact from the monitoring experience. A qualitative study of patients’ attitudes to and understanding of CA125 for monitoring for ovarian cancer recurrence found that both negative and positive test results can reassure patients if the result appears to legitimise patients’ own subjective experience of the disease: a positive result confirming their worst suspicion, or a negative result providing reassurance that they are as ‘well’ as they feel.\textsuperscript{278} Positive, or rising, test results can mediate a patient’s experience of any clinical symptoms, the knowledge that s/he is probably experiencing a relapse potentially making symptoms more unbearable or causing them to reinterpret prior ‘symptoms’ that had previously been discounted.\textsuperscript{278} Similar findings in other monitoring situations have also been observed.\textsuperscript{266}
Reduce the invasiveness of testing

Patients’ exposure to invasive tests can be reduced in two ways: by replacing an existing test with a less invasive one, or introducing a triage test to select the most appropriate patients for invasive testing. For the former option, assuming the new test aims to detect disease at a similar stage or time point, it will be important to establish its properties in relation to the existing test, and to ensure that it identifies a similar spectrum of patients. If so, treatment can be expected to have similar effectiveness, and the impact of the test on patient outcomes will result from the less invasive nature of the test, as for example with the use of a less invasive biopsy approach in patients at risk of colorectal cancer.  

If a new triage test is introduced to prevent harm from the testing process, such as with gene expression profiling as a triage to endomyocardial biopsy when monitoring for acute cardiac transplant rejection, or the introduction of cytological surveillance to reduce the number of colposcopies undertaken in women with mild dyskaryosis, the need for a randomised trial will be greater. Only a subgroup of patients will be selected for further investigation and treatment, such that treatment effectiveness could vary, and outcomes in those no longer selected for treatment must also be assessed. In this setting, the properties of the new test (in particular the rate of false negative results), the timing of detection, and treatment effectiveness are likely to be of overarching importance. Patient and clinician confidence in the new triage test will also be needed to ensure that any potential benefits are realised in practice. An element of ‘trust’ that the new, less invasive test will accurately identify those in need of the more invasive test is required on both parts. Patient preference is not always as intuitive as it might appear, for example, gene testing was favoured over endomyocardial biopsy in the example above, whilst patients at risk of bladder cancer recurrence preferred an immediate result from a more invasive test (cystoscopy) to the result from a less invasive (urine) test a week later.  

Reduce the volume of testing

Sometimes new monitoring strategies aim to reduce the number or frequency of tests without adversely impacting on patient outcome, as is often the goal when monitoring for cancer recurrence. One example is the proposal to reduce the number of CT scans from five over a three year period to two in the first year of follow-up following primary treatment of non-seminomatous testicular cancer. Under the new regime, there could be a change in
spectrum to later stage disease in those detected at the 12 month follow-up point and
detection of those patients whose disease recurs beyond 12 months would rely on clinical
relapse or detection by biochemical markers or chest x-ray. In this situation, the number of
cases that would be missed and stage of disease at detection could be identified by a
longitudinal study design but clinical outcomes between strategies could not be directly
compared without a randomised controlled trial. If however, sufficient evidence exists for
treatment at the various stages of disease, this could be linked to data from a longitudinal
study using a decision analytic type model.\textsuperscript{263}

A trial of less frequent fetal surveillance of small-for-gestational-age fetuses demonstrates the
sometimes complex responses of clinicians and patients to monitoring. Over half of the
experimental group in this trial attended for ultrasound more frequently than scheduled and
underwent additional tests of fetal wellbeing, suggesting that clinicians were not always
comfortable with the planned reduced frequency of fetal surveillance and making it difficult
to assess whether the apparent safety of less frequent monitoring may have been in part
because of this additional surveillance.\textsuperscript{159} At the same time 17\% of women in the twice
weekly surveillance group attended less frequently than requested suggesting a patient
perception of over-frequent monitoring.\textsuperscript{279}

\textbf{Reduce over-treatment}

In some monitoring contexts, a new test can be introduced to replace another simply to better
select the right patients, for example a new DNA-based test to detect human cytomegalovirus
(CMV) infection following stem cell transplant compared to the existing antigenemia test.\textsuperscript{170}
This is more akin to a diagnostic test context where the goal is to use most sensitive and/or
specific test, as the time dimension of monitoring is less relevant than the properties of the
test concerned. Although the new test may detect a different biochemical marker, if there is
no indication of a differential treatment response according to the marker used, and
randomised evidence exists for the effectiveness of treatment in patients identified, then a
study to establish the properties of the tests concerned may be sufficient evidence for the
introduction of the new test.

To fully impact on patient outcomes, however, clinicians must have confidence to act on the
results of the new test. Only half of patients who tested positive on the DNA-based test in the
example above actually underwent treatment, whereas in a trial of a new glactomannan assay
in patients at risk of invasive aspergillosis following stem cell transplant, two thirds of those treated in the experimental arm had a negative monitoring test, perhaps because clinicians had previously relied on clinical assessment as the basis for treatment decisions.\textsuperscript{170, 171}

**Delay or avoid treatment where it is not required**

The final scenario is one in which monitoring is introduced as an alternative to immediate treatment. In this circumstance, patient outcome might be affected by later treatment in the surveillance arm, further delays to necessary treatment due to false negative test results, the effectiveness of treatment at a later stage of disease and by the need for clinician and patient confidence in the monitoring regime. Where immediate treatment is the standard care option, evidence for treatment at a later stage of disease may not be available, so the onus is not only on demonstrating that the monitoring test used is clinically valid and able to detect the point at which treatment is needed, but also on evidencing treatment effectiveness in the surveillance group. In a small trial in infants with mild hip dysplasia, sonographic surveillance allowed abduction treatment to be delayed or avoided with no significant difference in radiological outcomes at one year compared to immediate abduction treatment; with less than 50\% of those in the surveillance arm underwent treatment during the course of the trial.\textsuperscript{159}

**Conclusion**

The impact of a monitoring strategy is driven not only by the properties and timing of testing and the effectiveness of treatment but also by patients’ response to the type and frequency of testing and clinicians trust in, and willingness to comply with, the monitoring protocol. Rutjes and colleagues advocate that what is more important for clinical decision making than the size, level or change in a given marker, is the confidence with which that marker can be used to inform patient management.\textsuperscript{280} A move towards a test validation paradigm is advocated, by which a number of methods (including establishing test properties) are used to determine whether the results of a test are meaningful in practice. In some circumstances, randomised evidence will be needed to fully assess the impact of a test, but this level of evidence will not be needed in every circumstance.

For example, the feasibility of testing and interpretability of test results can be estimated in the development phase of a test, as long as the technical properties of the test are established
in clinically relevant populations rather than laboratory-based studies alone. Patients’ interaction with the testing experience and their likely adherence to monitoring or to subsequent management can be assessed in qualitative studies. Feasibility or pilot studies can help to identify what a new test adds to current clinical practice, particularly in terms of clinicians’ interaction with and likely adherence to a new monitoring strategy, and can help identify potential barriers to implementation, as has been recommended for trials of complex interventions. Estimates of key aspects of test performance can be gained from non-randomised, preferably longitudinal, studies comparing tests against a (delayed) reference standard or clinical outcome. The efficacy of treatment is the only mechanism that requires evaluation in an RCT per se, however, the combined effect of the individual mechanisms that come into play may only be fully assessable in an RCT.

Any decision to undertake an RCT should be informed at a minimum by good evidence of the natural history of disease, the establishment of test properties (in terms of clinical validity and estimation of long-term change in disease status) and evidence (or lack of) treatment efficacy in those patients who are identified by the new monitoring strategy.
Chapter 7 - Simulating monitoring data and evaluating monitoring strategies
Introduction

Tests are used in healthcare to monitor, and subsequently manage, a variety of chronic conditions. The focus of this research is monitoring of progressive or recurrent conditions, where the aim of monitoring is to identify early signs of recurrence or progression prompting a change in management, typically initiation of treatment or further testing.

Monitoring strategies are complex interventions combining a test, a schedule, a decision rule and further diagnostic or therapeutic action. Monitoring strategies stipulate the frequency of testing, and the ‘monitoring rule’ used to identify when a change in patient management is necessary. A monitoring rule indicates the value or values that would trigger a change in management. Monitoring rules can be simple, where a single value above a threshold will prompt a change in management (a ‘snap-shot rule’), or more complex, where a patient requires a series of test results observed at different time points fitting a given criteria (such as relative increase from previous measures) to initiate a change in management (a ‘track-shot rule’).

Although patient monitoring is a fundamental function of healthcare, incurring considerable cost to health care providers, the area of monitoring is under researched and there is an increased need for monitoring strategies to be systematically developed with knowledge of the likely progression of disease and the performance of the monitoring test to be used. Dinnes et al in Chapter 3, reviewed the evidence base for prostate specific antigen (PSA) monitoring to identify recurrence of prostate cancer. The review identified the lack of a systematic approach in developing a monitoring strategy, with monitoring intervals based on standard follow-up schedules and limited evidence of consensus for the thresholds used to initiate treatment.

Stevens et al discussed various statistical models of the transition between the maintenance and re-established control phases of monitoring (the process of detecting when a disease is out of control leading to a change in management, for example treatment or more intensive monitoring) and identified a general statistical model for the evolution of monitoring data over time outlining possible sources of variation. This general statistical model proposes the form of monitoring data based upon the observed values of sequential monitoring tests,
the values of measurement error and other sources of variability, and the true disease state, which can be modelled based on epidemiological evidence but never observed.

This general model along with existing data, and evidence gathered from the literature, can be used to simulate monitoring data and allow the evaluation of strategies for a given target condition. The potential effect of monitoring strategies can then be evaluated and ranked, prior to full-scale investigation.\textsuperscript{282}

The example presented here investigates the use of the Enhanced Liver Fibrosis (ELF) biomarker in monitoring patients with known liver fibrosis, alongside an ongoing prospective multicentre randomised trial (the Enhanced Liver fibrosis (ELF) test to Uncover Cirrhosis as an Indication for Diagnosis and Action for Treatable Events (ELUCIDATE) trial).\textsuperscript{283} The ELUCIDATE trial evaluates ELF for the early detection of progression from liver fibrosis to liver cirrhosis compared to routine care, with the aim of enabling earlier treatment and potentially improved patient outcomes. It is described fully in Chapters 16 – 24.

### Aims and objectives

The aim of the study was to identify the optimal monitoring strategy, from candidate monitoring strategies, for patients known to have liver fibrosis receiving repeated testing using the biomarker ELF.

Candidate strategies were varied and evaluated, with the objectives being to:

- compare the alternative frequencies of monitoring (6 month or 12 month intervals).
- evaluate the benefit of using targeted retesting compared to no retesting.
- compare decision rules (positive results based on crossing a threshold determined by a single value (snapshot simple threshold rule), and track-shot rules based on absolute or relative increases from first test value, absolute or relative increases from last test value and prediction from a linear regression model).

### Methods

Firstly described is the model used to generate the underlying and unobserved disease progression, incorporating estimates of disease progression and the variability of these
estimates, for a cohort of simulated individuals. Then, the process of obtaining observed test result values using the true disease progression values and estimates of test performance is described. Finally, the methods used to evaluate and compare selected monitoring strategies, using both the observed test values and the true disease status, are given. Explanation of the notation used in the model can be seen in Table 13.

Table 13 Model notation

<table>
<thead>
<tr>
<th>Description</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual number</td>
<td>( i )</td>
</tr>
<tr>
<td>Number of initially simulated individuals</td>
<td>( n )</td>
</tr>
<tr>
<td>Number of simulated individuals eligible for randomisation</td>
<td>( N )</td>
</tr>
<tr>
<td>Fibrosis stage</td>
<td>( s )</td>
</tr>
<tr>
<td>Time within fibrosis stage</td>
<td>( x_{is} )</td>
</tr>
<tr>
<td>Time across fibrosis stages</td>
<td>( t )</td>
</tr>
<tr>
<td>Monitoring time points</td>
<td>( T )</td>
</tr>
<tr>
<td>Fibrosis progression</td>
<td>( p )</td>
</tr>
<tr>
<td>Mean fibrosis progression</td>
<td>( \mu )</td>
</tr>
<tr>
<td>Standard deviation of fibrosis progression</td>
<td>( \sigma )</td>
</tr>
<tr>
<td>Starting fibrosis stage</td>
<td>( S_i )</td>
</tr>
<tr>
<td>Probability of starting in each fibrosis stage</td>
<td>( \rho_s )</td>
</tr>
<tr>
<td>ELF score at each stage of fibrosis</td>
<td>( E_s )</td>
</tr>
<tr>
<td>Mean ELF score at each fibrosis stage</td>
<td>( \mu_s )</td>
</tr>
<tr>
<td>Standard deviation of ELF score at each fibrosis stage</td>
<td>( \sigma_s )</td>
</tr>
<tr>
<td>Observed mean ELF score at each fibrosis stage</td>
<td>( \mu_Y )</td>
</tr>
<tr>
<td>Observed standard deviation of ELF score at each fibrosis stage</td>
<td>( \sigma_Y )</td>
</tr>
<tr>
<td>ELF progression between fibrosis stages</td>
<td>( E_{ls} )</td>
</tr>
<tr>
<td>Values from the standard normal distribution</td>
<td>( Z_i )</td>
</tr>
<tr>
<td>Gradient of ELF progression</td>
<td>( B_{ls} )</td>
</tr>
<tr>
<td>Time point when patients progress in fibrosis stage</td>
<td>( \tau_{ls} )</td>
</tr>
<tr>
<td>True ELF by time in fibrosis stage</td>
<td>( U_y )</td>
</tr>
<tr>
<td>True ELF across fibrosis stages</td>
<td>( E_{lt} )</td>
</tr>
<tr>
<td>True ELF over the period of the trial</td>
<td>( U_t )</td>
</tr>
<tr>
<td>Time at registration</td>
<td>( \tau_r )</td>
</tr>
<tr>
<td>Time at randomisation</td>
<td>( \tau_t )</td>
</tr>
<tr>
<td><strong>Total observation error</strong></td>
<td>$\omega_i \omega_k$</td>
</tr>
<tr>
<td>----------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td><strong>Standard deviation of total observation error</strong></td>
<td>$\sigma_\omega$</td>
</tr>
<tr>
<td><strong>Observed ELF</strong></td>
<td>$\gamma_i Y_{it}$</td>
</tr>
<tr>
<td><strong>Observed ELF at monitoring points</strong></td>
<td>$\gamma_i Y_{it}$</td>
</tr>
<tr>
<td><strong>Entry ELF criteria</strong></td>
<td>$\gamma_i' Y_{it}'$</td>
</tr>
<tr>
<td><strong>Frequency of observations</strong></td>
<td>$\theta \theta$</td>
</tr>
<tr>
<td><strong>Range for targeted retesting</strong></td>
<td>$\delta \Delta$</td>
</tr>
<tr>
<td><strong>Time before compensation cirrhosis a patient is considered diseased</strong></td>
<td>$\delta \delta$</td>
</tr>
<tr>
<td><strong>Simple decision rule threshold</strong></td>
<td>$Y_i \cdot Y_i'$</td>
</tr>
<tr>
<td><strong>Simple decision rule threshold at the point of randomisation</strong></td>
<td>$Y_{0i} \cdot Y_{0i}'$</td>
</tr>
<tr>
<td><strong>Absolute increase from start value decision rule threshold</strong></td>
<td>$Y_{Di} \cdot Y_{Di}'$</td>
</tr>
<tr>
<td><strong>Absolute increase from last value decision rule threshold</strong></td>
<td>$Y_{Ei} \cdot Y_{Ei}'$</td>
</tr>
<tr>
<td><strong>Relative increase from start value decision rule threshold</strong></td>
<td>$Y_{Fi} \cdot Y_{Fi}'$</td>
</tr>
<tr>
<td><strong>Relative increase from last value decision rule threshold</strong></td>
<td>$Y_{Gi} \cdot Y_{Gi}'$</td>
</tr>
<tr>
<td><strong>Linear regression decision rule threshold</strong></td>
<td>$Y_{Hi} \cdot Y_{Hi}'$</td>
</tr>
<tr>
<td><strong>Maximum time in cirrhosis before trial entry</strong></td>
<td>$c_i c_1$</td>
</tr>
<tr>
<td><strong>Maximum time in fibrosis before entry to the trial</strong></td>
<td>$c_i c_2$</td>
</tr>
<tr>
<td><strong>Time between registration and randomisation</strong></td>
<td>$c_i c_3$</td>
</tr>
<tr>
<td><strong>Trial duration</strong></td>
<td>$c_i c_4$</td>
</tr>
<tr>
<td><strong>Time between test and retest measures</strong></td>
<td>$\epsilon \epsilon$</td>
</tr>
</tbody>
</table>

**Simulation of true disease progression**

The model simulated true disease progression, generating a random slope and random intercept in terms of fibrosis stage for each patient. The model then converted fibrosis stage to ELF values.

**Fibrosis progression—random slope**

The rate at which patients progress through the fibrosis stages was assumed to be constant throughout the stages of fibrosis and normally distributed, $p_i \sim N(\mu_p, \sigma_p^2)$, where $i = 1, \ldots, n$ and $n$ is the number of simulated individuals; $\mu_p$ is the mean fibrosis progression and $\sigma_p$ is the standard deviation of fibrosis progression. Fibrosis progression was restricted to only positive values, by fixing $p_i$ at 0.01 if $p_i \leq 0$, meaning only increases in fibrosis stage were simulated with no patients having decreasing fibrosis; however, as the increase is just 0.01 fibrosis units per year this effectively means these patients are in a stable fibrosis state.
Fibrosis stage at entry—random intercept

Patients recruited to a trial would be in varying stages of disease at entry. Using data on the likely distribution of fibrosis stage for a population with known liver fibrosis, a multinomial distribution was used to simulate a starting stage for each individual: $S_i \sim \text{Mtn}(0,1,2,3,4)$ ($\rho_0, \rho_1, \rho_2, \rho_3, \rho_4$) where $\rho_s$ is the probability of starting in each stage, $s$ is fibrosis stage and $s = 0, \ldots, 4$.

ELF score link to fibrosis stage

For each stage of fibrosis, the distribution of true ELF scores within fibrosis stage was assumed to follow a normal distribution, $E_s \sim \text{N}(\mu_s, \sigma_s^2)$, where $s$ is the fibrosis stage and $s = 0, \ldots, A$; $\mu_s$ is the mean value of ELF at each fibrosis stage and $\sigma_s$ is the standard deviation of ELF at each fibrosis stage.

ELF progression between fibrosis stages

The model used fibrosis stage on a continuous scale rather than discrete. To generate ELF scores for each patient at all stages of fibrosis, ELF progression between consecutive integer fibrosis stages was assumed to be linear. It was assumed that patients would have ELF values at the same point of the normal distribution for each fibrosis stage (patients would remain a given number of standard deviations from the mean). To randomly select the point of the normal distribution that patients would follow, a value from the standard normal distribution was generated for each patient, $z_i \sim \text{N}(0,1)$. The ELF value for each participant, at each stage of fibrosis was $E_{is} = \mu_s + (z_i \sigma_s)$, see Figure 12 (left).

ELF progression—random slope

The ELF values at the beginning of each fibrosis stage for each individual ($E_{is}$) and the rate each simulated participant progresses through fibrosis ($p_i$), were combined to calculate the increase in ELF per year. The gradient of ELF progression was $\beta_{is} = (E_{is+1} - E_{is})p_i$, for $s = 0, \ldots, 3$. The gradient of ELF progression after stage 4 was assumed to be the same as the gradient between stages 3 and 4, $\beta_{i3} = \beta_{i4}$. The time point signalling when each individual would progress to the next fibrosis stage was $\pi_{is} = \frac{s}{p_i}$. $\beta_{is}$ is the random slope in terms of ELF progression.

The underlying and true ELF progression for each stage and all time points from the onset of fibrosis was then calculated as $E_{js} = E_{is} + \beta_{is}x_{js}$, where $x_{js}$ is time within stage $s$ and $0 \leq x_{js} < \pi_{is+1} - \pi_{is}$ for $s = 0, \ldots, 3$. For stage 4, $s = 4$ and $x_{js} = 0, \ldots, \infty$. The true ELF values for each individual across time could also be expressed as:
\[ E_{it} = \begin{cases} E_{ij0} & \text{for } 0 \leq t < \pi_{i1} \\ E_{ij1} & \text{for } \pi_{i1} \leq t < \pi_{i2} \\ E_{ij2} & \text{for } \pi_{i2} \leq t < \pi_{i3} \\ E_{ij3} & \text{for } \pi_{i3} \leq t < \pi_{i4} \\ E_{ij4} & \text{for } \pi_{i4} \leq t \end{cases} \]

where \( t \) is time across all stages, \( t = 0, \ldots, \infty \). This allowed the simulation of life time progression data for a cohort of patients; see Figure 12 (centre). ELF values were truncated at 0 if a negative value was simulated.

**ELF value at entry—random intercept**

The time at registration for each participant (\( \tau_{ri} \)) was a randomly selected time point from the time period when the individual was in their generated fibrosis stage at study entry (\( S_i \)), a random value from the interval \([\pi_{iS_i}, \pi_{iS_i+1}]\), where \( S_i \) is the starting stage for each individual and \( S_i = 0, \ldots, 3 \). If the participant was in stage 4 of fibrosis at entry, \( S_i = 4 \) and \( \tau_{ri} \) was generated by identifying a random value from the interval \([\pi_{i4}, \pi_{i4} + c_1]\). Where \( c_1 \) is the longest given amount of time a patient can be in stage 4 of liver fibrosis before entering the trial. \( \tau_{ri} \) also has a maximum value of \( c_2 \), where \( c_2 \) is the longest amount of time a patient can have fibrosis at registration to the trial. If for a simulated individual \( \tau_{ri} > \pi_{i4} + c_1 \) or \( \tau_{ri} > c_2 \) the data for that individual was not used in the analysis. This was done to prevent patients being included when they would be confirmed as having cirrhosis, or at a point of fibrosis they would not have reached in their lifetime due to their simulated progression rate. The \( c_1 \) time used means participants registering in the trial who are in stage 4 of liver fibrosis have been in stage 4 for a maximum of \( c_1 \) years. The \( c_2 \) time used means participants have liver fibrosis for a maximum of \( c_2 \) years before being registered in the trial.

In the ELUCIDATE trial, a registration ELF test was given to each patient to assess eligibility. The first ELF test included in the trial data was taken at the point of randomisation. If the start of the trial occurred \( c_3 \) time units after registration, then the time at randomisation was \( \tau_{ri} = \tau_{ri} + c_3 \). The random intercept in terms of ELF is \( \alpha_i = E_{it} \) for \( t = \tau_{ti} \).

**Random slope and random intercept model in terms of ELF**

The underlying disease progression for the simulated individuals over the time of the trial was \( U_{it} = E_{it} \) for \( \tau_{ti} \leq t < \tau_{ti} + c_4 \) where \( c_4 \) is the duration of the trial and \( i \) denotes simulated patients with eligible registration ELF, \( i = 1, \ldots, N \), where \( N \) is the number of simulated patients available for randomisation; see Figure 12 (right).
Figure 12 Fibrosis units linked to ELF score (left), ELF progression through time (centre), and starting stage ELF adjusted ELF progression through time (right)
Simulation of observed values

The true underlying ELF measurements were converted to observed ELF measures by the addition of error.

Error

The error at each observation point ($\omega_{it}$) was formed of within-individual variation and analytical variation. Error was assumed to be normally distributed with a mean of zero, $\omega_{it} \sim \mathcal{N}(0, \sigma^2_{\omega})$. The observed ELF measurement at any given time was $Y_{it} = U_{it} + \omega_{it}$. Values were adjusted to equal 0 if a negative observed ELF value was simulated.

Entry criteria

In order to fulfil trial entry criteria the observed ELF measurement at registration had to be greater than the pre-set value of $Y_r^*$, thus the equation $Y_{it} > Y_r^*$ for $t = \tau_r$ had to be satisfied for each simulated participant to be included in the trial data.

Data sources

Data sources were used to estimate fibrosis progression rate, fibrosis stage at trial entry, measurement error and ELF score link to fibrosis stage. Additional information regarding the data sources used can be seen in Table 14.

Fibrosis progression rate

An estimate of the median rate of fibrosis progression based on data for 1,157 patients was obtained from Poynard et al; the estimate of the median was assumed to be equal to the mean and the 95% confidence interval for the median was used to calculate the standard deviation.\(^{284}\) When consulting clinical experts it was suggested the estimate provided by Poynard et al was identified in a population that was not comparable with that of the ELUCIDATE study (participants in the Poynard study were thought to have less severe disease). The estimate from Poynard et al was used primarily in the simulation model with an adjusted estimate used for sensitivity analyses. Estimates of fibrosis progression are given as Scheuer fibrosis units per year, where Scheuer scores range from 0 to 4 and measure severity of liver disease with stage 0 showing no fibrosis and 4 showing liver cirrhosis.\(^{285}\)
### Table 14 Data used in simulation model

<table>
<thead>
<tr>
<th></th>
<th>Data</th>
<th>Estimates used in model</th>
</tr>
</thead>
</table>
| **Fibrosis progression rate** | Poynard *et al.*: estimate of median fibrosis progression (Scheuer units per year) 0.133 (95% CI 0.125, 0.143). | Estimate calculated from Poynard *et al.*:
  \[ p_i \sim N(0.13, 0.17^2) \]
  Estimate after adjustment (to be used in sensitivity analyses): estimate of fibrosis progression was increased to reflect expert opinion.
  \[ p_i \sim N(0.27, 0.17^2) \] |
| **ELF stage at entry to trial** | Cross sectional data set: estimated proportion of patients in each stage.
  Stage 0- 0.25; stage 1- 0.35; stage 2- 0.13; stage 3- 0.15; stage 4- 0.12. | The cross-sectional data set was used:
  \[ \rho_0 = 0.25, \rho_1 = 0.35, \rho_2 = 0.13, \rho_3 = 0.15, \rho_4 = 0.12 \] |
| **Measurement error** | Longitudinal data set: estimate of the standard deviation of measurement error of 0.81. Siemens: estimate of the standard deviation of total measurement error of 0.11. ELUCIDATE registration and randomisation data: estimate of standard deviation of total measurement error 0.47. | Estimate obtained from ELUCIDATE was used.
  \[ \omega_{it} \sim N(0, 0.47^2) \] |
| **ELF link to fibrosis stage** | Cross sectional data set: estimates of ELF mean (SD) at each fibrosis stage.
  Stage 0- 8.82 (0.87); stage 1- 9.18 (0.96); stage 3- 9.55 (1.00); stage 4- 11.32 (1.47). | After adjustment: measurement error is accounted for to give the true unobserved ELF values and modified to represent stage of stage values.
  \[ E_0 \sim N(8.63, 0.73^2), E_1 \sim N(9.00, 0.84^2), E_2 \sim N(9.36, 0.89^2), E_4 \sim N(10.80, 1.39^2) \] |
**Fibrosis stage at entry**

As the purpose of the study was to simulate and evaluate trial data not all participants would enter the trial at the same stage of fibrosis. A cross-sectional data set with ELF results and Scheuer fibrosis scores following liver biopsy for 921 patients was used to identify the distribution of fibrosis stages in the cohort.\(^{61}\)

**Measurement error**

To estimate the error associated with each observed ELF test value, three data sources were considered. Firstly, a longitudinal data set with repeat ELF measurements (baseline and at 3 months) for 220 patients was subjected to analysis of variance to identify variability at the individual level. The manufacturer of the test also provided information on the performance of the ELF test.\(^{286, 287}\) Due to discrepancies between the estimates from the two sources, data was obtained directly from the ELUCIDATE trial. Registration and randomisation ELF values for 112 eligible participants were again subjected to analysis of variance to identify the variability at the individual level. The estimate obtained from the ELUCIDATE trial data was used in the simulation model.

**ELF score link to fibrosis stage**

The cross-sectional data set above was used to provide an estimate of observed ELF score for patients at each level of fibrosis with a corresponding measure of variability.\(^{288}\) To estimate the true and unobserved standard deviation of ELF scores at each fibrosis stage, the measurement error that would have been included in these observed measures, inflating the variability, was accounted for. As \(\sigma_s^2\) and \(\sigma_\omega^2\) are independent in the simulation, \(\sigma_Y^2 = \sigma_s^2 + \sigma_\omega^2\) so \(\sigma_s = \sqrt{\sigma_Y^2 - \sigma_\omega^2}\), where \(\sigma_Y\) is the observed standard deviation at each stage of fibrosis.

To estimate the true and unobserved mean ELF score at each stage of fibrosis the observed estimates were assumed to give the mean value for the midpoint of the corresponding fibrosis stage thus were altered to reflect an ELF value for the point when a patient initially enters each stage of fibrosis, \(\mu_s = \mu_Y + \frac{\mu_Y + \mu_{Y-1}}{2}\) for \(s = 1, \ldots, 4\), where \(\mu_Y\) is the observed mean ELF value at each stage of fibrosis; when \(s = 0\), \(\mu_s = \mu_0 = \mu_Y + \frac{\mu_Y + \mu_0}{2}\). \(\sigma_s^2\) is the between-individual variability at each fibrosis stage.
Implementation of a monitoring strategy

The effect of implementing different monitoring strategies was predicted using simulated observed values of ELF. The specified monitoring strategy (decision rule, use of retesting and frequency of testing) changed the simulated observed values that would be measured and the how the value or values for each individual would be interpreted.

Monitoring data required

The observed values that would have been measured under each monitoring strategy were extracted; the exact values required were dependent on both the frequency and duration of monitoring. The duration of monitoring was specified by the duration of the trial, $c_4$. If the frequency of monitoring was every $\theta$ time units after randomisation the observed values used to guide the management of participants would be $Y_{it}$ for $t = \tau_{ti} + \theta T$, where $T = 0,1,...,\frac{c_4}{\theta}$, and $\frac{c_4}{\theta}$ is the number of observation points additional to randomisation. The subscript used for time is now simplified to indicate monitoring time points, this will be $T$, where $T = \frac{t-\tau_{ti}}{\theta}$.

To incorporate a targeted retest component an additional test would be carried out $\epsilon$ time units after the scheduled tests, for patients with an observed value at that time point within a specified range ($\Delta$) of the value used to trigger a positive. When a patient required retesting, the mean of the original test and the retest result was calculated and this value was subjected to the decision rules to identify positive participants, $Y_{r\;it} = \frac{Y_{it} + Y_{r\;it} + \epsilon}{2}$. Patients with a value above the upper limit of the range were classed as positive on the initial test without further testing and patients below the limit of the retesting range were classed as negative using just the initial test.

Monitoring strategies

Monitoring strategies are defined by the decision rule for identifying a positive result and the data the decision rule is applied to, which is dependent on the frequency of monitoring and the use of retesting.

Simple decision rule (strategy A)

The simplest decision rule was based on a single value threshold (snap-shot rule). The threshold value, $Y^*$, was specified and any single observed value over this threshold indicated a positive result for that participant at that time point. A result was considered positive when $Y_{it} > Y^*$. 


Retesting (strategy B)

Patients with initial test value within $\Delta$ of the threshold value, $Y^*$, were subjected to retesting. Patients required retesting when $Y^* - \Delta > Y_{IT} < Y^* + \Delta$ and these patients where considered positive when $Y_{IT} > Y^*$. When patients did not require retesting, $Y_{IT} \leq Y^* - \Delta$ or $Y_{IT} \geq Y^* + \Delta$, patients were considered positive when $Y_{IT} > Y^*$. The retesting component could be used with any of the alternative decision rules explained.

Frequency of monitoring (strategy C)

The frequency of monitoring was every $\theta$ time units after the initial test at randomisation. By decreasing or increasing the value of $\theta$ the timing of monitoring tests became more or less frequent respectively. When varying the frequency of monitoring the time points and hence the observations evaluated changed. When changing the frequency of monitoring, patients are considered positive when $Y_{IT} > Y^*$ where the value of $\theta$ changed the data evaluated as $T = \frac{t-\tau_{IT}}{\theta}$. Varying the frequency of monitoring could be used in conjunction with any of the alternative decision rules explained.

Alternative decision rules

Decision rules incorporating previous test results as well as the current result (track-shot rules) to identify positive patients were also considered. Absolute and relative increases from randomisation ELF or from the last recorded ELF measure were investigated. A rule using predictions from a linear regression model fitted using all available observed data points was considered also.

Decisions rules based on absolute and relative increases and the linear regression method required at least two observations to declare a participant as positive. A simple threshold rule was used to identify participants at the point of randomisation using $Y_{i0} > Y_0^*, Y_D^*, Y_E^*, Y_F^*, Y_G^*$ and $Y_H^*$ are specified thresholds for the corresponding decision rule method.

Absolute increase from start value (strategy D)

A result was considered positive when the absolute difference between the test value and the first recorded value for the patient was greater than the threshold, $Y_D^*, Y_{IT} - Y_{i0} > Y_D^*$.

Absolute increase from last observed value (strategy E)

A result was considered positive when the absolute difference between the test value and the last observed test value for that patient was greater than the threshold, $Y_E^*, Y_{IT} - Y_{IT-1} > Y_E^*$.

Relative increase from start value (strategy F)
A result was considered positive when the relative difference between the test value and the first recorded test value for the patient was greater than the threshold, $Y_F^*, \frac{Y_{\text{test}}}{Y_{\text{first}}} > Y_F^*$.

Relative increase from last observed value (strategy G)
A result was considered positive when the relative difference between the test value and the last observed test value for that patient was greater than the threshold, $Y_G^*, \frac{Y_{\text{test}}}{Y_{\text{last}}-1} > Y_G^*$.

Linear regression (strategy H)
The linear regression decision rule involved the fitting of a linear regression model for each participant at each time point, using all available measures for that participant at the time point and the prediction from the model was then used to identify the patient as test positive or negative. A result was considered positive when the prediction from the linear regression model was greater than $Y_H^*$.

Evaluation of a monitoring strategy
To evaluate each strategy, the decision made from implementing that monitoring strategy using the simulated observed values and the corresponding true underlying values was assessed. With knowledge of the true underlying disease state of each participant, the performance of a variety of monitoring strategies was evaluated.

Comparison of observed results with the underlying disease state
Participants had a positive or negative test result based on the simulated observed data and the decision rule employed. The test result was then found to be either true or false depending on the underlying disease state. The purpose of the ELF test was to identify when patients enter compensated cirrhosis, stage 4. From the simulation of underlying true ELF values, the time point when each individual enters stage 4 was $\pi_{i,t_a}$. As it may be beneficial to identify patients prior to time $\pi_{i,t_a}$, participants were classed as ‘diseased’ $\delta$ units of time prior to $\pi_{i,t_a}$ and onwards. If at a testing point a participant was diseased ($\geq \pi_{i,t_a} - \delta$), a positive result would be a true positive and a negative result would be a false negative. If at a testing point the patient was not diseased ($< \pi_{i,t_a} - \delta$), a negative result would be a true negative and a positive result would be a false positive. As a positive result (truly or falsely) caused a change in management and cessation of monitoring, patients that achieved a positive result do not have a test result at subsequent monitoring times. Figure 14 illustrates how a strategy with a simple threshold decision rule can be evaluated.
Figure 13: Observed ELF measures

Figure 14: Results of implementing a monitoring strategy
Measuring the performance of a monitoring strategy

The performance of a monitoring strategy was assessed at each monitoring point by calculating the number of patients at each monitoring test visit, and specifically the number of true positive, false positive, true negative and false negative test results. The number of tests carried out across the duration of the strategy was used to represent resource use. The positive predictive value (PPV) was used to investigate how likely it was for an individual with a positive result to be diseased. To measure patient harm the delay from onset of disease to the point of diagnosis was calculated; this was the time between the onset of compensated cirrhosis ($\pi_{IA}$) and a patient having a positive test result.

When comparing strategies the number of tests per person for the duration of monitoring, PPV (for all tests over the duration of monitoring) and percentage of patients with delayed diagnosis (delay from onset of disease to diagnosis of over 12 months) were used to measure performance. To allow for comparisons to be made between strategies where only two of the three measures of performance ranged, thresholds used by monitoring strategies were varied to obtain a PPV of 25%. A PPV of 25% was chosen as this would be an acceptable PPV in practice.

Evaluation of strategies

The strategy evaluated first was the simple threshold strategy with observations every 6 months and no retest component (reference strategy). Alternative strategies were evaluated where individual components of the reference strategy were varied, the frequency of monitoring ($\theta$), the decision rule and whether a retest value was used. The same simulated data was used when evaluating strategies A-H.

Sensitivity analyses were carried out to estimate the effect of inaccurate information regarding test performance and progression of liver disease. Estimates used in the simulation of data were altered (halved and doubled) and the reference strategy was evaluated with all aspects of the strategy kept constant (including the threshold value). Results for the reference strategy with the threshold varied to give PPV of 25% using data with altered estimates were obtained also. Further sensitivity analyses were undertaken with the fibrosis progression rate adjusted based on expert opinion; the analysis of strategies A-H with PPV held at 25% was replicated using data generated with adjustment and analysis of the reference strategy using varied estimates to generate monitoring data.

To assess the accuracy of the model, the mean and standard deviation of randomisation values were calculated and compared for the ELUCIDATE and simulated data sets. Analysis
of variance was used to assess between-individual and within-individual variability of ELF values recorded for patients in the trial and the simulated results. Multilevel models were fitted using the simulated observed values and observed values from the ELUCIDATE trial (for participants with two or more ELF measures post registration) and the results from these models were compared. In the ELUCIDATE trial ELF measurements were not taken in the majority of cases after the participant had an ELF result of 9.5 or above. To allow for this the ELUCIDATE and simulated data sets were modified so that each patient with an ELF measure of 9.5 or above did not have any subsequent measures.

Simulations were based on a cohort of 20,000 patients to give adequate precision. With 20,000 test results, if one of the performance measures gave an estimate of 15% a corresponding 95% confidence interval would range from 14.5% to 15.5%; for an estimate of 1.5% a 95% confidence interval would range from 1.3% to 1.7%.

**Results**

The estimates from various data sources and details of the data used in the simulation model can be found in Table 14. Additionally the estimates regarding the trial are shown in Table 15. When evaluating strategies patients were considered truly positive if they received a positive result three months prior to entering compensated cirrhosis (\(\delta = 0.25\)). When using targeted retest patients with initial values within ±1 of the threshold value have a retest (\(\Delta = 1\)).

The same simulated data was used when evaluating strategies A-H. For the simulated cohort of 20,000 patients, 5,314 (26.6%) would develop cirrhosis if there were no intervention during the period of the trial.

**Reference monitoring strategy (strategy A)**

Table 16 shows the performance of the reference monitoring strategy at each testing time point. For the reference monitoring strategy (simple threshold, observations 6-monthly and no retest component), the threshold required to maintain the PPV at 25% was 10.715. The sensitivity and PPV calculated for the strategy were highest at the initial observation point and the percentage of tests with a positive result was also larger, due to cases being identified from a prevalent population at the initial testing point. The percentage of false negative results generally rises as the strategy continues in time. Over the duration of the monitoring strategy 7.64 tests per person (152,724 tests in total) were performed and 6.10% of all patients had delay to diagnosis.
**Table 15 Trial consideration estimates used in simulation modelling**

<table>
<thead>
<tr>
<th>Description</th>
<th>Estimate used in simulation modelling</th>
</tr>
</thead>
</table>
| Maximum time in cirrhosis before trial entry                                | To avoid simulating patients that are in advanced cirrhosis, the maximum amount of time a patient has been cirrhotic for is set to 2 years.  
  \[ c_1 = 2 \]                                                                 |
| Maximum time in fibrosis before entry to the trial                          | To avoid patients being simulated at a point of disease they would not have reached given their fibrosis progression rate the maximum amount of time a patient has had fibrosis for at the time of entering the trial is set at 20 years.  
  \[ c_2 = 20 \]                                                                   |
| Time between registration and randomisation                                 | The time between registration and randomisation ELF was estimated to be 3 months.  
  \[ c_3 = 0.25 \]                                                                  |
| Trial duration                                                              | The duration of the trial used in all simulations was 5 years.  
  \[ c_4 = 5 \]                                                                      |
| Time between test and retest measures                                       | The time between a patient having a test and retest (if the original test were in the target range) was estimated to be 1 week.  
  \[ \epsilon = 0.02 \]                                                                |
Table 16 Results by observation point for the reference strategy (strategy A)

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(^a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1162 (5.8)</td>
<td>2236 (11.2)</td>
<td>771 (3.9)</td>
<td>15831 (79.2)</td>
<td>3398 (17.0)</td>
<td>1933 (9.7)</td>
<td>34.2</td>
<td>60.1</td>
</tr>
<tr>
<td>6</td>
<td>16602</td>
<td>207 (1.2)</td>
<td>920 (5.5)</td>
<td>733 (4.4)</td>
<td>14742 (88.8)</td>
<td>1127 (6.8)</td>
<td>940 (5.7)</td>
<td>18.4</td>
<td>22.0</td>
</tr>
<tr>
<td>12</td>
<td>15475</td>
<td>147 (0.9)</td>
<td>661 (4.3)</td>
<td>740 (4.8)</td>
<td>13927 (90.0)</td>
<td>808 (5.2)</td>
<td>887 (5.7)</td>
<td>18.2</td>
<td>16.6</td>
</tr>
<tr>
<td>18</td>
<td>14667</td>
<td>102 (0.7)</td>
<td>558 (3.8)</td>
<td>783 (5.3)</td>
<td>13224 (90.2)</td>
<td>660 (4.5)</td>
<td>885 (6.0)</td>
<td>15.5</td>
<td>11.5</td>
</tr>
<tr>
<td>24</td>
<td>14007</td>
<td>113 (0.8)</td>
<td>495 (3.5)</td>
<td>786 (5.6)</td>
<td>12613 (90.0)</td>
<td>608 (4.3)</td>
<td>899 (6.4)</td>
<td>18.6</td>
<td>12.6</td>
</tr>
<tr>
<td>30</td>
<td>13399</td>
<td>104 (0.8)</td>
<td>446 (3.3)</td>
<td>813 (6.1)</td>
<td>12036 (89.8)</td>
<td>550 (4.1)</td>
<td>917 (6.8)</td>
<td>18.9</td>
<td>11.3</td>
</tr>
<tr>
<td>36</td>
<td>12849</td>
<td>127 (1.0)</td>
<td>458 (3.6)</td>
<td>805 (6.3)</td>
<td>11459 (89.2)</td>
<td>585 (4.6)</td>
<td>932 (7.3)</td>
<td>21.7</td>
<td>13.6</td>
</tr>
<tr>
<td>42</td>
<td>12264</td>
<td>121 (1.0)</td>
<td>429 (3.5)</td>
<td>828 (6.8)</td>
<td>10886 (88.8)</td>
<td>550 (4.5)</td>
<td>949 (7.7)</td>
<td>22.0</td>
<td>12.8</td>
</tr>
<tr>
<td>48</td>
<td>11714</td>
<td>129 (1.1)</td>
<td>449 (3.8)</td>
<td>817 (7.0)</td>
<td>10319 (88.1)</td>
<td>578 (4.9)</td>
<td>946 (8.1)</td>
<td>22.3</td>
<td>13.6</td>
</tr>
<tr>
<td>54</td>
<td>11136</td>
<td>135 (1.2)</td>
<td>390 (3.5)</td>
<td>806 (7.2)</td>
<td>9805 (88.0)</td>
<td>525 (4.7)</td>
<td>941 (8.5)</td>
<td>25.7</td>
<td>14.3</td>
</tr>
<tr>
<td>60</td>
<td>10611</td>
<td>125 (1.2)</td>
<td>369 (3.5)</td>
<td>814 (7.7)</td>
<td>9303 (87.7)</td>
<td>494 (4.7)</td>
<td>939 (8.8)</td>
<td>25.3</td>
<td>13.3</td>
</tr>
<tr>
<td>All</td>
<td>152724</td>
<td>2472 (1.6)</td>
<td>7411 (4.9)</td>
<td>8696 (5.7)</td>
<td>134145 (87.8)</td>
<td>9883 (6.5)</td>
<td>11168 (7.3)</td>
<td>25.0</td>
<td>22.1</td>
</tr>
</tbody>
</table>

\(^a\) Tests performed where the patient was diseased
Comparing strategies with changes to individual components to the reference strategy

Table 17 and Figure 15 show the performance of various monitoring strategies. Results of each strategy by observation point can be found in Tables 18-24.

**Inferior strategies**
The retest strategy (strategy B) and the strategies with decision rules based on absolute and relative increases from the first and last recorded value (strategies D, E, F and G) were inferior to the reference strategy, requiring more tests and causing more patients who had progressed to liver cirrhosis to experience a delay to diagnosis.

The main effect of the retest strategy was to increase the number of tests performed (increase of 3.30 tests per person), with also a small increase in the percentage of patients with delay to diagnosis (absolute increase of 0.40). Whereas, the strategies with decision rules based on absolute and relative increases from initial value showed only small increases in the number of tests required (increases of 0.14 and 0.18 tests per person respectively) but larger increases in the percentage of patients with delay to diagnosis (absolute increases of 1.58% and 2.05% respectively) compared with the reference strategy. The absolute and relative increase from last recorded value decision rules both increased the number of tests required (by 0.98 and 1.18 per person, respectively) and increased the percentage of patients with delay to diagnosis (to 10.42% and 11.09%),

**‘Trade-off’ strategies**
The reduced monitoring frequency strategy (strategy C) showed a ‘trade-off’ between delay to diagnosis and the number of tests required when compared with the results of the reference strategy. The number of tests required decreased by 3.30 tests per person and the percentage of patients with delay to diagnosis increased by 0.15% (absolute increase) compared with the reference strategy.

**Superior strategies**
The reference strategy was found to be inferior to the linear regression strategy. The linear regression strategy used fewer tests (decrease of 0.12 tests per person) and had a lower percentage of patients with delay to diagnosis (absolute decrease of 0.47%) when compared to the reference strategy.
Table 17 Results of strategies A-H

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Monitoring strategy components</th>
<th>PPV</th>
<th>Tests*</th>
<th>Delay to diagnosis†</th>
<th>Test performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Decision rule</td>
<td>Threshold value</td>
<td>Frequency (months)</td>
<td>Retest</td>
<td>Initial threshold</td>
</tr>
<tr>
<td>A</td>
<td>Simple threshold</td>
<td>10.715</td>
<td>0.5</td>
<td>FALSE</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>Simple threshold</td>
<td>10.58</td>
<td>0.5</td>
<td>TRUE</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>Simple threshold</td>
<td>10.55</td>
<td>1</td>
<td>FALSE</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>Absolute increase from initial value</td>
<td>1.295</td>
<td>0.5</td>
<td>FALSE</td>
<td>10.715</td>
</tr>
<tr>
<td>E</td>
<td>Absolute increase from last value</td>
<td>1.46</td>
<td>0.5</td>
<td>FALSE</td>
<td>10.715</td>
</tr>
<tr>
<td>F</td>
<td>Relative increase from initial value</td>
<td>1.144</td>
<td>0.5</td>
<td>FALSE</td>
<td>10.715</td>
</tr>
<tr>
<td>G</td>
<td>Relative increase from last value</td>
<td>1.1795</td>
<td>0.5</td>
<td>FALSE</td>
<td>10.715</td>
</tr>
<tr>
<td>H</td>
<td>Linear regression</td>
<td>10.495</td>
<td>0.5</td>
<td>FALSE</td>
<td>10.715</td>
</tr>
</tbody>
</table>

*Tests over the duration of monitoring.
*b Number of tests per person over the duration of monitoring.
*c Patients with delayed diagnosis (delay from onset of disease to diagnosis of over 12 months).
*d % of all patients with delay to diagnosis.
*e % of patients that would reach cirrhosis within the trial period with delay to diagnosis.
*f TP pp is the number of true positive results per person over the duration of monitoring.
*g FP pp is the number of false positive results per person over the duration of monitoring.
†218974 tests were carried out to generate 153971 results due to retests being used.
Figure 15: Results of various monitoring strategies
### Table 18: Results using retest monitoring strategy (strategy B) by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests* (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>28357/20000</td>
<td>1233 (6.2)</td>
<td>2447 (12.2)</td>
<td>700 (3.5)</td>
<td>15620 (78.1)</td>
<td>3680 (18.4)</td>
<td>1933 (9.7)</td>
<td>33.5</td>
<td>63.8</td>
</tr>
<tr>
<td>6</td>
<td>22924/16320</td>
<td>144 (0.9)</td>
<td>796 (4.9)</td>
<td>711 (4.4)</td>
<td>14669 (89.9)</td>
<td>940 (5.8)</td>
<td>855 (5.2)</td>
<td>15.3</td>
<td>16.8</td>
</tr>
<tr>
<td>12</td>
<td>21549/15380</td>
<td>125 (0.8)</td>
<td>568 (3.7)</td>
<td>734 (4.8)</td>
<td>13953 (90.7)</td>
<td>693 (4.5)</td>
<td>859 (5.6)</td>
<td>18.0</td>
<td>14.6</td>
</tr>
<tr>
<td>18</td>
<td>20617/14687</td>
<td>97 (0.7)</td>
<td>494 (3.4)</td>
<td>775 (5.3)</td>
<td>13321 (90.7)</td>
<td>591 (4.0)</td>
<td>872 (5.9)</td>
<td>16.4</td>
<td>11.1</td>
</tr>
<tr>
<td>24</td>
<td>19913/14096</td>
<td>117 (0.8)</td>
<td>438 (3.1)</td>
<td>776 (5.5)</td>
<td>12765 (90.6)</td>
<td>555 (3.9)</td>
<td>893 (6.3)</td>
<td>21.1</td>
<td>13.1</td>
</tr>
<tr>
<td>30</td>
<td>19167/13541</td>
<td>103 (0.8)</td>
<td>408 (3.0)</td>
<td>803 (5.9)</td>
<td>12227 (90.3)</td>
<td>511 (3.8)</td>
<td>906 (6.7)</td>
<td>20.2</td>
<td>11.4</td>
</tr>
<tr>
<td>36</td>
<td>18555/13030</td>
<td>105 (0.8)</td>
<td>413 (3.2)</td>
<td>816 (6.3)</td>
<td>11696 (89.8)</td>
<td>518 (4.0)</td>
<td>921 (7.1)</td>
<td>20.3</td>
<td>11.4</td>
</tr>
<tr>
<td>42</td>
<td>17927/12512</td>
<td>105 (0.8)</td>
<td>422 (3.4)</td>
<td>849 (6.8)</td>
<td>11136 (89.0)</td>
<td>527 (4.2)</td>
<td>954 (7.6)</td>
<td>19.9</td>
<td>11.0</td>
</tr>
<tr>
<td>48</td>
<td>17272/11985</td>
<td>108 (0.9)</td>
<td>408 (3.4)</td>
<td>858 (7.2)</td>
<td>10611 (88.5)</td>
<td>516 (4.3)</td>
<td>966 (8.1)</td>
<td>20.9</td>
<td>11.2</td>
</tr>
<tr>
<td>54</td>
<td>16674/11469</td>
<td>126 (1.1)</td>
<td>392 (3.4)</td>
<td>855 (7.5)</td>
<td>10096 (88.0)</td>
<td>518 (4.5)</td>
<td>981 (8.6)</td>
<td>24.3</td>
<td>12.8</td>
</tr>
<tr>
<td>60</td>
<td>16019/10951</td>
<td>88 (0.8)</td>
<td>269 (2.5)</td>
<td>887 (8.1)</td>
<td>9707 (88.6)</td>
<td>357 (3.3)</td>
<td>975 (8.9)</td>
<td>24.6</td>
<td>9.0</td>
</tr>
<tr>
<td>All</td>
<td>218974/153971</td>
<td>2351 (1.5)</td>
<td>7055 (4.6)</td>
<td>8764 (5.7)</td>
<td>135801 (88.2)</td>
<td>9406 (6.1)</td>
<td>11115 (7.2)</td>
<td>25.0</td>
<td>21.2</td>
</tr>
</tbody>
</table>

*Tests performed/number of people tests performed on (number of results generated)

Table 19: Results using reduced frequency of monitoring strategy (strategy C) by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1247 (6.2)</td>
<td>2735 (13.7)</td>
<td>686 (3.4)</td>
<td>15332 (76.7)</td>
<td>3982 (19.9)</td>
<td>1933 (9.7)</td>
<td>31.3</td>
<td>64.5</td>
</tr>
<tr>
<td>12</td>
<td>16018</td>
<td>277 (1.7)</td>
<td>1253 (7.8)</td>
<td>739 (4.6)</td>
<td>13749 (85.8)</td>
<td>1530 (9.6)</td>
<td>1016 (6.3)</td>
<td>18.1</td>
<td>27.3</td>
</tr>
<tr>
<td>24</td>
<td>14488</td>
<td>243 (1.7)</td>
<td>983 (6.8)</td>
<td>775 (5.3)</td>
<td>12487 (86.2)</td>
<td>1226 (8.5)</td>
<td>1018 (7.0)</td>
<td>19.8</td>
<td>23.9</td>
</tr>
<tr>
<td>36</td>
<td>13262</td>
<td>247 (1.9)</td>
<td>952 (7.2)</td>
<td>782 (5.9)</td>
<td>11281 (85.1)</td>
<td>1199 (9.0)</td>
<td>1029 (7.8)</td>
<td>20.6</td>
<td>24.0</td>
</tr>
<tr>
<td>48</td>
<td>12063</td>
<td>246 (2.0)</td>
<td>861 (7.1)</td>
<td>806 (6.7)</td>
<td>10150 (84.1)</td>
<td>1107 (9.2)</td>
<td>1052 (8.7)</td>
<td>22.2</td>
<td>23.4</td>
</tr>
<tr>
<td>60</td>
<td>10956</td>
<td>252 (2.3)</td>
<td>757 (6.9)</td>
<td>809 (7.4)</td>
<td>9138 (83.4)</td>
<td>1009 (9.2)</td>
<td>1061 (9.7)</td>
<td>25.0</td>
<td>23.8</td>
</tr>
<tr>
<td>All</td>
<td>86787</td>
<td>2512 (2.9)</td>
<td>7541 (8.7)</td>
<td>4597 (5.3)</td>
<td>72137 (83.1)</td>
<td>10053 (11.6)</td>
<td>7109 (8.2)</td>
<td>25.0</td>
<td>35.3</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased
Table 20: Results using absolute increase from start value monitoring strategy (strategy D) by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1162 (5.8)</td>
<td>2236 (11.2)</td>
<td>771 (3.9)</td>
<td>15831 (79.2)</td>
<td>3398 (17.0)</td>
<td>1933 (9.7)</td>
<td>34.2</td>
<td>60.1</td>
</tr>
<tr>
<td>6</td>
<td>16602</td>
<td>564 (3.4)</td>
<td>890 (5.4)</td>
<td>15098 (90.9)</td>
<td>614 (3.7)</td>
<td>940 (5.7)</td>
<td>11.3</td>
<td>6.1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>15988</td>
<td>517 (3.2)</td>
<td>1014 (6.3)</td>
<td>14391 (90.0)</td>
<td>583 (3.6)</td>
<td>1080 (6.8)</td>
<td>11.3</td>
<td>6.1</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>15405</td>
<td>550 (3.6)</td>
<td>1085 (7.0)</td>
<td>13675 (88.8)</td>
<td>645 (4.2)</td>
<td>1180 (7.7)</td>
<td>14.7</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>14760</td>
<td>582 (3.9)</td>
<td>1132 (7.7)</td>
<td>12930 (87.6)</td>
<td>698 (4.7)</td>
<td>1248 (8.5)</td>
<td>16.6</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>14062</td>
<td>1123 (8.0)</td>
<td>12211 (86.8)</td>
<td>728 (5.2)</td>
<td>1292 (9.2)</td>
<td>8.1</td>
<td>14.7</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>14334</td>
<td>603 (4.5)</td>
<td>11478 (86.1)</td>
<td>790 (5.9)</td>
<td>1253 (9.4)</td>
<td>11.3</td>
<td>13.1</td>
<td>14.9</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>14244</td>
<td>1048 (8.4)</td>
<td>10736 (85.6)</td>
<td>760 (6.1)</td>
<td>1222 (9.7)</td>
<td>8.1</td>
<td>14.7</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>11784</td>
<td>583 (4.9)</td>
<td>10034 (85.1)</td>
<td>795 (6.7)</td>
<td>1167 (9.9)</td>
<td>11.3</td>
<td>13.1</td>
<td>14.9</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>10989</td>
<td>865 (7.9)</td>
<td>9315 (84.8)</td>
<td>809 (7.4)</td>
<td>1083 (9.9)</td>
<td>11.3</td>
<td>13.1</td>
<td>14.9</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>10180</td>
<td>574 (5.6)</td>
<td>8637 (84.8)</td>
<td>778 (7.6)</td>
<td>969 (9.5)</td>
<td>11.3</td>
<td>13.1</td>
<td>14.9</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>155648</td>
<td>2653 (1.7)</td>
<td>7945 (5.1)</td>
<td>10714 (6.9)</td>
<td>134336 (86.3)</td>
<td>10598 (6.8)</td>
<td>26.2</td>
<td>21.1</td>
<td></td>
</tr>
</tbody>
</table>

\* Tests performed where the patient was diseased

Table 21: Results using absolute increase from last value monitoring strategy (strategy E) by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>2236 (11.2)</td>
<td>771 (3.9)</td>
<td>15831 (79.2)</td>
<td>3398 (17.0)</td>
<td>1933 (9.7)</td>
<td>34.2</td>
<td>60.1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>16602</td>
<td>890 (5.4)</td>
<td>15098 (90.9)</td>
<td>614 (3.7)</td>
<td>940 (5.7)</td>
<td>11.3</td>
<td>6.1</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>15988</td>
<td>1014 (6.3)</td>
<td>14391 (90.0)</td>
<td>583 (3.6)</td>
<td>1080 (6.8)</td>
<td>11.3</td>
<td>6.1</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>15405</td>
<td>1085 (7.0)</td>
<td>13675 (88.8)</td>
<td>645 (4.2)</td>
<td>1180 (7.7)</td>
<td>14.7</td>
<td>8.1</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>14760</td>
<td>1132 (7.7)</td>
<td>12930 (87.6)</td>
<td>698 (4.7)</td>
<td>1248 (8.5)</td>
<td>16.6</td>
<td>9.3</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>14062</td>
<td>12211 (86.8)</td>
<td>728 (5.2)</td>
<td>1292 (9.2)</td>
<td>8.1</td>
<td>14.7</td>
<td>8.1</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>14334</td>
<td>11478 (86.1)</td>
<td>790 (5.9)</td>
<td>1253 (9.4)</td>
<td>11.3</td>
<td>13.1</td>
<td>14.9</td>
<td>14.9</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>14244</td>
<td>10736 (85.6)</td>
<td>760 (6.1)</td>
<td>1222 (9.7)</td>
<td>8.1</td>
<td>14.7</td>
<td>8.1</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>11784</td>
<td>10034 (85.1)</td>
<td>795 (6.7)</td>
<td>1167 (9.9)</td>
<td>11.3</td>
<td>13.1</td>
<td>14.9</td>
<td>14.9</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>10989</td>
<td>9315 (84.8)</td>
<td>809 (7.4)</td>
<td>1083 (9.9)</td>
<td>11.3</td>
<td>13.1</td>
<td>14.9</td>
<td>14.9</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>10180</td>
<td>8637 (84.8)</td>
<td>778 (7.6)</td>
<td>969 (9.5)</td>
<td>11.3</td>
<td>13.1</td>
<td>14.9</td>
<td>14.9</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>172363</td>
<td>134336 (86.3)</td>
<td>10598 (6.8)</td>
<td>13367 (8.6)</td>
<td>25.0</td>
<td>21.1</td>
<td>21.1</td>
<td>21.1</td>
<td></td>
</tr>
</tbody>
</table>

\* Tests performed where the patient was diseased
### Table 22: Results using relative increase from start value monitoring strategy (strategy F) by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1162 (5.8)</td>
<td>2236 (11.2)</td>
<td>771 (3.9)</td>
<td>15831 (79.2)</td>
<td>3398 (17.0)</td>
<td>1933 (9.7)</td>
<td>34.2</td>
<td>60.1</td>
</tr>
<tr>
<td>6</td>
<td>16602</td>
<td>44 (0.3)</td>
<td>641 (3.9)</td>
<td>896 (5.4)</td>
<td>15021 (90.5)</td>
<td>685 (4.1)</td>
<td>940 (5.7)</td>
<td>6.4</td>
<td>4.7</td>
</tr>
<tr>
<td>12</td>
<td>15917</td>
<td>48 (0.3)</td>
<td>539 (3.4)</td>
<td>1041 (6.5)</td>
<td>14289 (89.8)</td>
<td>587 (3.7)</td>
<td>1089 (6.8)</td>
<td>8.2</td>
<td>4.4</td>
</tr>
<tr>
<td>18</td>
<td>15330</td>
<td>84 (0.5)</td>
<td>525 (3.4)</td>
<td>1125 (7.3)</td>
<td>13596 (88.7)</td>
<td>609 (4.0)</td>
<td>1209 (7.9)</td>
<td>13.8</td>
<td>6.9</td>
</tr>
<tr>
<td>24</td>
<td>14721</td>
<td>100 (0.7)</td>
<td>558 (3.8)</td>
<td>1193 (8.1)</td>
<td>12870 (87.4)</td>
<td>658 (4.5)</td>
<td>1293 (8.8)</td>
<td>15.2</td>
<td>7.7</td>
</tr>
<tr>
<td>30</td>
<td>14063</td>
<td>143 (1.0)</td>
<td>524 (3.7)</td>
<td>1217 (8.7)</td>
<td>12179 (86.6)</td>
<td>667 (4.7)</td>
<td>1360 (9.7)</td>
<td>21.4</td>
<td>10.5</td>
</tr>
<tr>
<td>36</td>
<td>13396</td>
<td>169 (1.3)</td>
<td>519 (3.9)</td>
<td>1184 (8.8)</td>
<td>11524 (86.0)</td>
<td>688 (5.1)</td>
<td>1353 (10.1)</td>
<td>24.6</td>
<td>12.5</td>
</tr>
<tr>
<td>42</td>
<td>12708</td>
<td>179 (1.4)</td>
<td>528 (4.2)</td>
<td>1164 (9.2)</td>
<td>10837 (85.3)</td>
<td>707 (5.6)</td>
<td>1343 (10.6)</td>
<td>25.3</td>
<td>13.3</td>
</tr>
<tr>
<td>48</td>
<td>12001</td>
<td>195 (1.6)</td>
<td>568 (4.7)</td>
<td>1099 (9.2)</td>
<td>10139 (84.5)</td>
<td>763 (6.4)</td>
<td>1294 (10.8)</td>
<td>25.6</td>
<td>15.1</td>
</tr>
<tr>
<td>54</td>
<td>11238</td>
<td>226 (2.0)</td>
<td>528 (4.7)</td>
<td>1009 (9.0)</td>
<td>9475 (84.3)</td>
<td>754 (6.7)</td>
<td>1235 (11.0)</td>
<td>30.0</td>
<td>18.3</td>
</tr>
<tr>
<td>60</td>
<td>10484</td>
<td>220 (2.1)</td>
<td>530 (5.1)</td>
<td>918 (8.8)</td>
<td>8816 (84.1)</td>
<td>750 (7.2)</td>
<td>1138 (10.9)</td>
<td>29.3</td>
<td>19.3</td>
</tr>
<tr>
<td>All</td>
<td>156460</td>
<td>2570 (1.6)</td>
<td>7696 (4.9)</td>
<td>11617 (7.4)</td>
<td>134577 (86.0)</td>
<td>10266 (6.6)</td>
<td>14187 (9.1)</td>
<td>25.0</td>
<td>18.1</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased

### Table 23: Results using relative increase from last value monitoring strategy (strategy G) by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1162 (5.8)</td>
<td>2236 (11.2)</td>
<td>771 (3.9)</td>
<td>15831 (79.2)</td>
<td>3398 (17.0)</td>
<td>1933 (9.7)</td>
<td>34.2</td>
<td>60.1</td>
</tr>
<tr>
<td>6</td>
<td>16602</td>
<td>44 (0.3)</td>
<td>641 (3.9)</td>
<td>896 (5.4)</td>
<td>15021 (90.5)</td>
<td>685 (4.1)</td>
<td>940 (5.7)</td>
<td>6.4</td>
<td>4.7</td>
</tr>
<tr>
<td>12</td>
<td>15917</td>
<td>48 (0.3)</td>
<td>539 (3.4)</td>
<td>1041 (6.5)</td>
<td>14289 (89.8)</td>
<td>587 (3.7)</td>
<td>1089 (6.8)</td>
<td>8.2</td>
<td>4.4</td>
</tr>
<tr>
<td>18</td>
<td>15330</td>
<td>84 (0.5)</td>
<td>525 (3.4)</td>
<td>1125 (7.3)</td>
<td>13596 (88.7)</td>
<td>609 (4.0)</td>
<td>1209 (7.9)</td>
<td>13.8</td>
<td>6.9</td>
</tr>
<tr>
<td>24</td>
<td>14721</td>
<td>100 (0.7)</td>
<td>558 (3.8)</td>
<td>1193 (8.1)</td>
<td>12870 (87.4)</td>
<td>658 (4.5)</td>
<td>1293 (8.8)</td>
<td>15.2</td>
<td>7.7</td>
</tr>
<tr>
<td>30</td>
<td>14063</td>
<td>143 (1.0)</td>
<td>524 (3.7)</td>
<td>1217 (8.7)</td>
<td>12179 (86.6)</td>
<td>667 (4.7)</td>
<td>1360 (9.7)</td>
<td>21.4</td>
<td>10.5</td>
</tr>
<tr>
<td>36</td>
<td>13396</td>
<td>169 (1.3)</td>
<td>519 (3.9)</td>
<td>1184 (8.8)</td>
<td>11524 (86.0)</td>
<td>688 (5.1)</td>
<td>1353 (10.1)</td>
<td>24.6</td>
<td>12.5</td>
</tr>
<tr>
<td>42</td>
<td>12708</td>
<td>179 (1.4)</td>
<td>528 (4.2)</td>
<td>1164 (9.2)</td>
<td>10837 (85.3)</td>
<td>707 (5.6)</td>
<td>1343 (10.6)</td>
<td>25.3</td>
<td>13.3</td>
</tr>
<tr>
<td>48</td>
<td>12001</td>
<td>195 (1.6)</td>
<td>568 (4.7)</td>
<td>1099 (9.2)</td>
<td>10139 (84.5)</td>
<td>763 (6.4)</td>
<td>1294 (10.8)</td>
<td>25.6</td>
<td>15.1</td>
</tr>
<tr>
<td>54</td>
<td>11238</td>
<td>226 (2.0)</td>
<td>528 (4.7)</td>
<td>1009 (9.0)</td>
<td>9475 (84.3)</td>
<td>754 (6.7)</td>
<td>1235 (11.0)</td>
<td>30.0</td>
<td>18.3</td>
</tr>
<tr>
<td>60</td>
<td>10484</td>
<td>220 (2.1)</td>
<td>530 (5.1)</td>
<td>918 (8.8)</td>
<td>8816 (84.1)</td>
<td>750 (7.2)</td>
<td>1138 (10.9)</td>
<td>29.3</td>
<td>19.3</td>
</tr>
<tr>
<td>All</td>
<td>176385</td>
<td>2570 (1.6)</td>
<td>7696 (4.9)</td>
<td>11617 (7.4)</td>
<td>134577 (86.0)</td>
<td>10266 (6.6)</td>
<td>14187 (9.1)</td>
<td>25.0</td>
<td>18.1</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased
Table 24: Results using linear regression monitoring strategy (strategy H) by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(^a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1162 (5.8)</td>
<td>2236 (11.2)</td>
<td>771 (3.9)</td>
<td>15831 (79.2)</td>
<td>3398 (17.0)</td>
<td>1933 (9.7)</td>
<td>34.2</td>
<td>60.1</td>
</tr>
<tr>
<td>6</td>
<td>16602</td>
<td>280 (1.7)</td>
<td>1488 (9.0)</td>
<td>660 (4.0)</td>
<td>14174 (85.4)</td>
<td>1768 (10.6)</td>
<td>940 (5.7)</td>
<td>15.8</td>
<td>29.8</td>
</tr>
<tr>
<td>12</td>
<td>14834</td>
<td>146 (1.0)</td>
<td>741 (5.0)</td>
<td>650 (4.4)</td>
<td>13297 (89.6)</td>
<td>887 (6.0)</td>
<td>796 (5.4)</td>
<td>16.5</td>
<td>18.3</td>
</tr>
<tr>
<td>18</td>
<td>13947</td>
<td>92 (0.7)</td>
<td>492 (3.5)</td>
<td>685 (4.9)</td>
<td>12678 (90.9)</td>
<td>584 (4.2)</td>
<td>777 (5.6)</td>
<td>15.8</td>
<td>11.8</td>
</tr>
<tr>
<td>24</td>
<td>13363</td>
<td>87 (0.7)</td>
<td>385 (2.9)</td>
<td>706 (5.3)</td>
<td>12185 (91.2)</td>
<td>472 (3.5)</td>
<td>793 (5.9)</td>
<td>18.4</td>
<td>11.0</td>
</tr>
<tr>
<td>30</td>
<td>12891</td>
<td>93 (0.7)</td>
<td>307 (2.4)</td>
<td>742 (5.8)</td>
<td>11749 (91.1)</td>
<td>400 (3.1)</td>
<td>835 (6.5)</td>
<td>23.2</td>
<td>11.1</td>
</tr>
<tr>
<td>36</td>
<td>12491</td>
<td>85 (0.7)</td>
<td>284 (2.3)</td>
<td>772 (6.2)</td>
<td>11350 (90.9)</td>
<td>369 (3.0)</td>
<td>857 (6.9)</td>
<td>23.0</td>
<td>9.9</td>
</tr>
<tr>
<td>42</td>
<td>12122</td>
<td>68 (0.6)</td>
<td>268 (2.2)</td>
<td>838 (6.9)</td>
<td>10948 (90.3)</td>
<td>336 (2.8)</td>
<td>906 (7.5)</td>
<td>20.2</td>
<td>7.5</td>
</tr>
<tr>
<td>48</td>
<td>11786</td>
<td>100 (0.8)</td>
<td>273 (2.3)</td>
<td>859 (7.3)</td>
<td>10554 (89.5)</td>
<td>373 (3.2)</td>
<td>959 (8.1)</td>
<td>26.8</td>
<td>10.4</td>
</tr>
<tr>
<td>54</td>
<td>11413</td>
<td>107 (0.9)</td>
<td>277 (2.4)</td>
<td>880 (7.7)</td>
<td>10149 (88.9)</td>
<td>384 (3.4)</td>
<td>987 (8.6)</td>
<td>27.9</td>
<td>10.8</td>
</tr>
<tr>
<td>60</td>
<td>11029</td>
<td>111 (1.0)</td>
<td>260 (2.4)</td>
<td>906 (8.2)</td>
<td>9752 (88.4)</td>
<td>371 (3.4)</td>
<td>1017 (9.2)</td>
<td>29.9</td>
<td>10.9</td>
</tr>
<tr>
<td>All</td>
<td>150478</td>
<td>2331 (1.5)</td>
<td>7011 (4.7)</td>
<td>8469 (5.6)</td>
<td>132667 (88.2)</td>
<td>9342 (6.2)</td>
<td>10800 (7.2)</td>
<td>25.0</td>
<td>21.6</td>
</tr>
</tbody>
</table>

\(^a\) Tests performed where the patient was diseased
Sensitivity analyses

Comparing results from the reference strategy when varying estimates of test performance and disease progression

Table 25 demonstrates the effect on the reference strategy of increasing (doubling) or decreasing (halving) various parameter estimates. Results for each monitoring time point using these alternate estimates can be seen in Tables 26-37.

Improved estimates of test performance (decreased measurement error and decreased between-individual variability) both improved PPV (absolute increases of 4.6% and 8.7% respectively) and increased the number of tests required (increases of 0.73 and 0.91 tests per person) with decreased measurement error also increasing the percentage of patients with delay to diagnosis (absolute increase of 1.30%). Both increased and decreased between-individual variability reduced the percentage of patients with delay to diagnosis (absolute decrease of 0.72% and 2.12% respectively). An increased rate of fibrosis progression led to both increased PPV (absolute increase of 4.2%) and percentage of patients with delay to diagnosis (absolute increase of 1.52%) but decreased the number of tests required (decrease of 0.64 tests per person).

The largest difference in PPV was achieved by increasing the between-individual variability (absolute decrease of 8.8%); the largest difference in number of tests required was achieved by increasing measurement error (decrease of 1.84 tests per person); and the largest difference in the percentage of patients with delay to diagnosis was achieved by decreasing between-individual variability (absolute decrease of 2.12%).

Adjusted fibrosis progression rate

Results of evaluating strategies based on data with the adjusted estimate of fibrosis progression can be seen in Tables 38-59 and Figure 16. Strategies appeared to behave in a similar way in comparison to the reference strategy with threshold selected to obtain PPV of 25% when using the unadjusted estimate.
<table>
<thead>
<tr>
<th>Change in data simulation</th>
<th>Threshold</th>
<th>PPV (%)</th>
<th>Number of tests per person&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Delay&lt;sup&gt;b&lt;/sup&gt; (%)</th>
<th>Develop cirrhosis&lt;sup&gt;c&lt;/sup&gt; n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>10.715</td>
<td>25.0</td>
<td>7.64</td>
<td>6.10</td>
<td>5314 (26.6)</td>
</tr>
<tr>
<td>Decreased† measurement error</td>
<td>10.715</td>
<td>29.6</td>
<td>8.37</td>
<td>7.40</td>
<td>5248 (26.2)</td>
</tr>
<tr>
<td></td>
<td>10.450</td>
<td>25.0</td>
<td>7.70</td>
<td>5.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increased‡ measurement error</td>
<td>10.715</td>
<td>17.6</td>
<td>5.84</td>
<td>3.85</td>
<td>5421 (27.1)</td>
</tr>
<tr>
<td></td>
<td>11.365</td>
<td>25.0</td>
<td>7.65</td>
<td>7.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decrease† between-individual variability</td>
<td>10.715</td>
<td>33.6</td>
<td>8.55</td>
<td>3.98</td>
<td>5139 (25.7)</td>
</tr>
<tr>
<td></td>
<td>10.463</td>
<td>25.0</td>
<td>7.84</td>
<td>2.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increased‡ between-individual variability</td>
<td>10.715</td>
<td>16.5</td>
<td>5.80</td>
<td>5.38</td>
<td>5272 (26.4)</td>
</tr>
<tr>
<td></td>
<td>11.905</td>
<td>25.0</td>
<td>8.26</td>
<td>9.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decreased† fibrosis progression rate</td>
<td>10.715</td>
<td>22.7</td>
<td>7.90</td>
<td>4.95</td>
<td>4440 (22.2)</td>
</tr>
<tr>
<td></td>
<td>10.860</td>
<td>25.0</td>
<td>8.29</td>
<td>5.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increased‡ fibrosis progression rate</td>
<td>10.715</td>
<td>29.2</td>
<td>7.00</td>
<td>7.62</td>
<td>7689 (38.4)</td>
</tr>
<tr>
<td></td>
<td>10.460</td>
<td>25.0</td>
<td>6.21</td>
<td>5.63</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of tests per person over the duration of monitoring.

<sup>b</sup> % of all patients with delayed diagnosis (delay from onset of disease to diagnosis of over 12 months).

<sup>c</sup> Patients that would go on to develop cirrhosis in the monitoring duration if no intervention were received.

†Decrease is halving the estimate used in the original simulation.
Increase is doubling the estimate used in the original simulation.

Table 26: Results using reference strategy with decreased measurement error by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1153 (5.8)</td>
<td>1931 (9.7)</td>
<td>750 (3.8)</td>
<td>16166 (80.8)</td>
<td>3084 (15.4)</td>
<td>1903 (9.5)</td>
<td>37.4</td>
<td>60.6</td>
</tr>
<tr>
<td>6</td>
<td>16916</td>
<td>113 (0.7)</td>
<td>478 (2.8)</td>
<td>810 (4.8)</td>
<td>15515 (91.7)</td>
<td>591 (3.5)</td>
<td>923 (5.5)</td>
<td>19.1</td>
<td>12.2</td>
</tr>
<tr>
<td>12</td>
<td>16325</td>
<td>92 (0.6)</td>
<td>371 (2.3)</td>
<td>881 (5.4)</td>
<td>14981 (91.8)</td>
<td>463 (2.8)</td>
<td>973 (6.0)</td>
<td>19.9</td>
<td>9.5</td>
</tr>
<tr>
<td>18</td>
<td>15862</td>
<td>109 (0.7)</td>
<td>341 (2.1)</td>
<td>934 (5.9)</td>
<td>14478 (91.3)</td>
<td>450 (2.8)</td>
<td>1043 (6.6)</td>
<td>24.2</td>
<td>10.5</td>
</tr>
<tr>
<td>24</td>
<td>15412</td>
<td>103 (0.7)</td>
<td>331 (2.1)</td>
<td>967 (6.3)</td>
<td>14011 (90.9)</td>
<td>434 (2.8)</td>
<td>1070 (6.9)</td>
<td>23.7</td>
<td>9.6</td>
</tr>
<tr>
<td>30</td>
<td>14978</td>
<td>93 (0.6)</td>
<td>356 (2.4)</td>
<td>1025 (6.8)</td>
<td>13504 (90.2)</td>
<td>449 (3.0)</td>
<td>1118 (7.5)</td>
<td>20.7</td>
<td>8.3</td>
</tr>
<tr>
<td>36</td>
<td>14529</td>
<td>129 (0.9)</td>
<td>353 (2.4)</td>
<td>1032 (7.1)</td>
<td>13015 (89.6)</td>
<td>482 (3.3)</td>
<td>1161 (8.0)</td>
<td>26.8</td>
<td>11.1</td>
</tr>
<tr>
<td>42</td>
<td>14047</td>
<td>128 (0.9)</td>
<td>356 (2.5)</td>
<td>1061 (7.6)</td>
<td>12502 (89.0)</td>
<td>484 (3.4)</td>
<td>1189 (8.5)</td>
<td>26.4</td>
<td>10.8</td>
</tr>
<tr>
<td>48</td>
<td>13563</td>
<td>112 (0.8)</td>
<td>341 (2.5)</td>
<td>1086 (8.0)</td>
<td>12024 (88.7)</td>
<td>453 (3.3)</td>
<td>1198 (8.8)</td>
<td>24.7</td>
<td>9.3</td>
</tr>
<tr>
<td>54</td>
<td>13110</td>
<td>137 (1.0)</td>
<td>334 (2.5)</td>
<td>1105 (8.4)</td>
<td>11534 (88.0)</td>
<td>471 (3.6)</td>
<td>1242 (9.5)</td>
<td>29.1</td>
<td>11.0</td>
</tr>
<tr>
<td>60</td>
<td>12639</td>
<td>164 (1.3)</td>
<td>370 (2.9)</td>
<td>1095 (8.7)</td>
<td>11010 (87.1)</td>
<td>534 (4.2)</td>
<td>1259 (10.0)</td>
<td>30.7</td>
<td>13.0</td>
</tr>
<tr>
<td>All</td>
<td>167381</td>
<td>2333 (1.4)</td>
<td>5562 (3.3)</td>
<td>10746 (6.4)</td>
<td>148740 (88.9)</td>
<td>7895 (4.7)</td>
<td>13079 (7.8)</td>
<td>29.6</td>
<td>17.8</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased.
### Table 27: Results using reference strategy with decreased measurement error by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1289 (6.4)</td>
<td>2755 (13.8)</td>
<td>614 (3.1)</td>
<td>15342 (76.7)</td>
<td>4044 (20.2)</td>
<td>1903 (9.5)</td>
<td>31.9</td>
<td>67.7</td>
</tr>
<tr>
<td>6</td>
<td>15956</td>
<td>114 (0.7)</td>
<td>665 (4.2)</td>
<td>645 (4.0)</td>
<td>14532 (91.1)</td>
<td>779 (4.9)</td>
<td>759 (4.8)</td>
<td>14.6</td>
<td>15.0</td>
</tr>
<tr>
<td>12</td>
<td>15177</td>
<td>93 (0.6)</td>
<td>441 (2.9)</td>
<td>686 (4.5)</td>
<td>13957 (92.0)</td>
<td>534 (3.5)</td>
<td>779 (5.1)</td>
<td>17.4</td>
<td>11.9</td>
</tr>
<tr>
<td>18</td>
<td>14643</td>
<td>97 (0.7)</td>
<td>431 (2.9)</td>
<td>726 (5.0)</td>
<td>13389 (91.4)</td>
<td>528 (3.6)</td>
<td>823 (5.6)</td>
<td>18.4</td>
<td>11.8</td>
</tr>
<tr>
<td>24</td>
<td>14115</td>
<td>83 (0.6)</td>
<td>438 (3.1)</td>
<td>761 (5.4)</td>
<td>12833 (90.9)</td>
<td>521 (3.7)</td>
<td>844 (6.0)</td>
<td>15.9</td>
<td>9.8</td>
</tr>
<tr>
<td>30</td>
<td>13594</td>
<td>111 (0.8)</td>
<td>410 (3.0)</td>
<td>772 (5.7)</td>
<td>12301 (90.5)</td>
<td>521 (3.8)</td>
<td>883 (6.5)</td>
<td>21.3</td>
<td>12.6</td>
</tr>
<tr>
<td>36</td>
<td>13074</td>
<td>100 (0.8)</td>
<td>399 (3.1)</td>
<td>785 (6.0)</td>
<td>11789 (90.2)</td>
<td>499 (3.8)</td>
<td>885 (6.8)</td>
<td>20.0</td>
<td>11.3</td>
</tr>
<tr>
<td>42</td>
<td>12574</td>
<td>108 (0.9)</td>
<td>361 (2.9)</td>
<td>815 (6.5)</td>
<td>11290 (89.8)</td>
<td>469 (3.7)</td>
<td>923 (7.3)</td>
<td>23.0</td>
<td>11.7</td>
</tr>
<tr>
<td>48</td>
<td>12105</td>
<td>108 (0.9)</td>
<td>379 (3.1)</td>
<td>818 (6.8)</td>
<td>10800 (89.2)</td>
<td>487 (4.0)</td>
<td>926 (7.6)</td>
<td>22.2</td>
<td>11.7</td>
</tr>
<tr>
<td>54</td>
<td>11618</td>
<td>133 (1.1)</td>
<td>383 (3.3)</td>
<td>817 (7.0)</td>
<td>10285 (88.5)</td>
<td>516 (4.4)</td>
<td>950 (8.2)</td>
<td>25.8</td>
<td>14.0</td>
</tr>
<tr>
<td>60</td>
<td>11102</td>
<td>118 (1.1)</td>
<td>393 (3.5)</td>
<td>815 (7.3)</td>
<td>9776 (88.1)</td>
<td>511 (4.6)</td>
<td>933 (8.4)</td>
<td>23.1</td>
<td>12.6</td>
</tr>
<tr>
<td>All</td>
<td>153957</td>
<td>2354 (1.5)</td>
<td>7055 (4.6)</td>
<td>8254 (5.4)</td>
<td>136294 (88.5)</td>
<td>9409 (6.1)</td>
<td>10608 (6.9)</td>
<td>25.0</td>
<td>22.2</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased

### Table 28: Results using reference strategy with increased measurement error by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1169 (5.8)</td>
<td>3257 (16.3)</td>
<td>812 (4.1)</td>
<td>14762 (73.8)</td>
<td>4426 (22.1)</td>
<td>1981 (9.9)</td>
<td>26.4</td>
<td>59.0</td>
</tr>
<tr>
<td>6</td>
<td>15574</td>
<td>369 (2.4)</td>
<td>1897 (12.2)</td>
<td>621 (4.0)</td>
<td>12687 (81.5)</td>
<td>2266 (14.5)</td>
<td>990 (6.4)</td>
<td>16.3</td>
<td>37.3</td>
</tr>
<tr>
<td>12</td>
<td>13308</td>
<td>213 (1.6)</td>
<td>1360 (10.2)</td>
<td>540 (4.1)</td>
<td>11195 (84.1)</td>
<td>1573 (11.8)</td>
<td>753 (5.7)</td>
<td>13.5</td>
<td>28.3</td>
</tr>
<tr>
<td>18</td>
<td>11735</td>
<td>141 (1.2)</td>
<td>1130 (9.6)</td>
<td>495 (4.2)</td>
<td>9969 (85.0)</td>
<td>1271 (10.8)</td>
<td>636 (5.4)</td>
<td>11.1</td>
<td>22.2</td>
</tr>
<tr>
<td>24</td>
<td>10464</td>
<td>123 (1.2)</td>
<td>881 (8.4)</td>
<td>455 (4.3)</td>
<td>9005 (86.1)</td>
<td>1004 (9.6)</td>
<td>578 (5.5)</td>
<td>12.3</td>
<td>21.3</td>
</tr>
<tr>
<td>30</td>
<td>9460</td>
<td>99 (1.0)</td>
<td>753 (8.0)</td>
<td>430 (4.5)</td>
<td>8178 (86.4)</td>
<td>852 (9.0)</td>
<td>529 (5.6)</td>
<td>11.6</td>
<td>18.7</td>
</tr>
<tr>
<td>36</td>
<td>8608</td>
<td>93 (1.1)</td>
<td>682 (7.9)</td>
<td>408 (4.7)</td>
<td>7425 (86.3)</td>
<td>775 (9.0)</td>
<td>501 (5.8)</td>
<td>12.0</td>
<td>18.6</td>
</tr>
<tr>
<td>42</td>
<td>7833</td>
<td>88 (1.1)</td>
<td>534 (6.8)</td>
<td>404 (5.2)</td>
<td>6807 (86.9)</td>
<td>622 (7.9)</td>
<td>492 (6.3)</td>
<td>14.1</td>
<td>17.9</td>
</tr>
<tr>
<td>48</td>
<td>7211</td>
<td>86 (1.2)</td>
<td>514 (7.1)</td>
<td>391 (5.4)</td>
<td>6220 (86.3)</td>
<td>600 (8.3)</td>
<td>477 (6.6)</td>
<td>14.3</td>
<td>18.0</td>
</tr>
<tr>
<td>54</td>
<td>6611</td>
<td>78 (1.2)</td>
<td>454 (6.9)</td>
<td>388 (5.9)</td>
<td>5691 (86.1)</td>
<td>532 (8.0)</td>
<td>466 (7.0)</td>
<td>14.7</td>
<td>16.7</td>
</tr>
<tr>
<td>60</td>
<td>6079</td>
<td>77 (1.3)</td>
<td>409 (6.7)</td>
<td>375 (6.2)</td>
<td>5218 (85.8)</td>
<td>486 (8.0)</td>
<td>452 (7.4)</td>
<td>15.8</td>
<td>17.0</td>
</tr>
<tr>
<td>All</td>
<td>116883</td>
<td>2536 (2.2)</td>
<td>11871 (10.2)</td>
<td>5319 (4.6)</td>
<td>97157 (83.1)</td>
<td>14407 (12.3)</td>
<td>7855 (6.7)</td>
<td>17.6</td>
<td>32.3</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased
Table 29: Results using reference strategy with increased measurement error by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>17590</td>
<td>368 (2.1)</td>
<td>1081 (6.1)</td>
<td>1001 (5.7)</td>
<td>15140 (86.1)</td>
<td>1449 (8.2)</td>
<td>1369 (7.8)</td>
<td>25.4</td>
<td>26.9</td>
</tr>
<tr>
<td>12</td>
<td>16141</td>
<td>274 (1.7)</td>
<td>827 (5.1)</td>
<td>942 (5.8)</td>
<td>14098 (87.3)</td>
<td>1101 (6.8)</td>
<td>1216 (7.5)</td>
<td>24.9</td>
<td>22.5</td>
</tr>
<tr>
<td>18</td>
<td>15040</td>
<td>189 (1.3)</td>
<td>726 (4.8)</td>
<td>913 (6.1)</td>
<td>13212 (87.8)</td>
<td>915 (6.1)</td>
<td>1102 (7.3)</td>
<td>20.7</td>
<td>17.2</td>
</tr>
<tr>
<td>24</td>
<td>14125</td>
<td>160 (1.1)</td>
<td>583 (4.1)</td>
<td>889 (6.3)</td>
<td>12493 (88.4)</td>
<td>743 (5.3)</td>
<td>1049 (7.4)</td>
<td>21.5</td>
<td>15.3</td>
</tr>
<tr>
<td>30</td>
<td>13382</td>
<td>139 (1.0)</td>
<td>580 (4.3)</td>
<td>881 (6.6)</td>
<td>11782 (88.0)</td>
<td>719 (5.4)</td>
<td>1020 (7.6)</td>
<td>19.3</td>
<td>13.6</td>
</tr>
<tr>
<td>36</td>
<td>12663</td>
<td>130 (1.0)</td>
<td>596 (4.7)</td>
<td>861 (6.8)</td>
<td>11076 (87.5)</td>
<td>726 (5.7)</td>
<td>991 (7.8)</td>
<td>17.9</td>
<td>13.1</td>
</tr>
<tr>
<td>42</td>
<td>11937</td>
<td>121 (1.0)</td>
<td>473 (4.0)</td>
<td>883 (7.4)</td>
<td>10460 (87.6)</td>
<td>594 (5.0)</td>
<td>1004 (8.4)</td>
<td>20.4</td>
<td>12.1</td>
</tr>
<tr>
<td>48</td>
<td>11343</td>
<td>129 (1.1)</td>
<td>497 (4.4)</td>
<td>877 (7.7)</td>
<td>9840 (86.7)</td>
<td>626 (5.5)</td>
<td>1006 (8.9)</td>
<td>20.6</td>
<td>12.8</td>
</tr>
<tr>
<td>54</td>
<td>10717</td>
<td>119 (1.1)</td>
<td>441 (4.1)</td>
<td>883 (8.2)</td>
<td>9274 (86.5)</td>
<td>560 (5.2)</td>
<td>1002 (9.3)</td>
<td>21.2</td>
<td>11.9</td>
</tr>
<tr>
<td>60</td>
<td>10157</td>
<td>125 (1.2)</td>
<td>433 (4.3)</td>
<td>882 (8.7)</td>
<td>8717 (85.8)</td>
<td>558 (5.5)</td>
<td>1007 (9.9)</td>
<td>22.4</td>
<td>12.4</td>
</tr>
<tr>
<td>All</td>
<td>153095</td>
<td>2602 (1.7)</td>
<td>7799 (5.1)</td>
<td>10145 (6.6)</td>
<td>132549 (86.6)</td>
<td>10401 (6.8)</td>
<td>12747 (8.3)</td>
<td>25.0</td>
<td>20.4</td>
</tr>
</tbody>
</table>

\(a\) Tests performed where the patient was diseased

Table 30: Results using reference strategy with decreased between-individual variability by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1171 (5.9)</td>
<td>963 (4.8)</td>
<td>639 (3.2)</td>
<td>17227 (86.1)</td>
<td>2134 (10.7)</td>
<td>1810 (9.0)</td>
<td>54.9</td>
<td>64.7</td>
</tr>
<tr>
<td>6</td>
<td>17866</td>
<td>290 (1.6)</td>
<td>548 (3.1)</td>
<td>514 (2.9)</td>
<td>16514 (92.4)</td>
<td>838 (4.7)</td>
<td>804 (4.5)</td>
<td>34.6</td>
<td>36.1</td>
</tr>
<tr>
<td>12</td>
<td>17028</td>
<td>185 (1.1)</td>
<td>421 (2.5)</td>
<td>500 (2.9)</td>
<td>15922 (93.5)</td>
<td>606 (3.6)</td>
<td>685 (4.0)</td>
<td>30.5</td>
<td>27</td>
</tr>
<tr>
<td>18</td>
<td>16422</td>
<td>181 (1.1)</td>
<td>406 (2.5)</td>
<td>464 (2.8)</td>
<td>15371 (93.6)</td>
<td>587 (3.6)</td>
<td>645 (3.9)</td>
<td>30.8</td>
<td>28.1</td>
</tr>
<tr>
<td>24</td>
<td>15835</td>
<td>128 (0.8)</td>
<td>391 (2.5)</td>
<td>447 (2.8)</td>
<td>14869 (93.9)</td>
<td>519 (3.3)</td>
<td>575 (3.6)</td>
<td>24.7</td>
<td>22.3</td>
</tr>
<tr>
<td>30</td>
<td>15316</td>
<td>117 (0.8)</td>
<td>373 (2.4)</td>
<td>449 (2.9)</td>
<td>14377 (93.9)</td>
<td>490 (3.2)</td>
<td>566 (3.7)</td>
<td>23.9</td>
<td>20.7</td>
</tr>
<tr>
<td>36</td>
<td>14826</td>
<td>127 (0.9)</td>
<td>445 (3.0)</td>
<td>434 (2.9)</td>
<td>13820 (93.2)</td>
<td>572 (3.9)</td>
<td>561 (3.8)</td>
<td>22.2</td>
<td>22.6</td>
</tr>
<tr>
<td>42</td>
<td>14254</td>
<td>117 (0.8)</td>
<td>424 (3.0)</td>
<td>436 (3.1)</td>
<td>13277 (93.1)</td>
<td>541 (3.8)</td>
<td>553 (3.9)</td>
<td>21.6</td>
<td>21.2</td>
</tr>
<tr>
<td>48</td>
<td>13713</td>
<td>127 (0.9)</td>
<td>471 (3.4)</td>
<td>431 (3.1)</td>
<td>12684 (92.5)</td>
<td>598 (4.4)</td>
<td>558 (4.1)</td>
<td>21.2</td>
<td>22.8</td>
</tr>
<tr>
<td>54</td>
<td>13115</td>
<td>136 (1.0)</td>
<td>424 (3.2)</td>
<td>435 (3.3)</td>
<td>12120 (92.4)</td>
<td>560 (4.3)</td>
<td>571 (4.4)</td>
<td>24.3</td>
<td>23.8</td>
</tr>
<tr>
<td>60</td>
<td>12555</td>
<td>124 (1.0)</td>
<td>474 (3.8)</td>
<td>442 (3.5)</td>
<td>11515 (91.7)</td>
<td>598 (4.8)</td>
<td>566 (4.5)</td>
<td>20.7</td>
<td>21.9</td>
</tr>
<tr>
<td>All</td>
<td>170930</td>
<td>2703 (1.6)</td>
<td>5340 (3.1)</td>
<td>5191 (3.0)</td>
<td>157696 (92.3)</td>
<td>8043 (4.7)</td>
<td>7894 (4.6)</td>
<td>33.6</td>
<td>34.2</td>
</tr>
</tbody>
</table>

\(a\) Tests performed where the patient was diseased
Table 31: Results using reference strategy with decreased between-individual variability by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased*n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1333 (6.7)</td>
<td>1517 (7.6)</td>
<td>477 (2.4)</td>
<td>16673 (83.4)</td>
<td>2850 (14.2)</td>
<td>1810 (9.0)</td>
<td>46.8</td>
<td>73.6</td>
</tr>
<tr>
<td>6</td>
<td>17150</td>
<td>259 (1.5)</td>
<td>826 (4.8)</td>
<td>333 (1.9)</td>
<td>15732 (91.7)</td>
<td>1085 (6.3)</td>
<td>592 (3.5)</td>
<td>23.9</td>
<td>43.8</td>
</tr>
<tr>
<td>12</td>
<td>16065</td>
<td>156 (1.0)</td>
<td>655 (4.1)</td>
<td>298 (1.9)</td>
<td>14956 (93.1)</td>
<td>811 (5.0)</td>
<td>454 (2.8)</td>
<td>19.2</td>
<td>34.4</td>
</tr>
<tr>
<td>18</td>
<td>15254</td>
<td>131 (0.9)</td>
<td>558 (3.7)</td>
<td>271 (1.8)</td>
<td>14294 (93.7)</td>
<td>689 (4.5)</td>
<td>402 (2.6)</td>
<td>19.0</td>
<td>32.6</td>
</tr>
<tr>
<td>24</td>
<td>14565</td>
<td>86 (0.6)</td>
<td>562 (3.9)</td>
<td>253 (1.7)</td>
<td>13664 (93.8)</td>
<td>648 (4.4)</td>
<td>339 (2.3)</td>
<td>13.3</td>
<td>25.4</td>
</tr>
<tr>
<td>30</td>
<td>13917</td>
<td>79 (0.6)</td>
<td>552 (4.0)</td>
<td>256 (1.8)</td>
<td>13030 (93.6)</td>
<td>631 (4.5)</td>
<td>335 (2.4)</td>
<td>12.5</td>
<td>23.6</td>
</tr>
<tr>
<td>36</td>
<td>13286</td>
<td>96 (0.7)</td>
<td>576 (4.3)</td>
<td>241 (1.8)</td>
<td>12373 (93.1)</td>
<td>672 (5.1)</td>
<td>337 (2.5)</td>
<td>14.3</td>
<td>28.5</td>
</tr>
<tr>
<td>42</td>
<td>12614</td>
<td>72 (0.6)</td>
<td>529 (4.2)</td>
<td>238 (1.9)</td>
<td>11775 (93.3)</td>
<td>601 (4.8)</td>
<td>310 (2.5)</td>
<td>12.0</td>
<td>23.2</td>
</tr>
<tr>
<td>48</td>
<td>12013</td>
<td>85 (0.7)</td>
<td>575 (4.8)</td>
<td>250 (2.1)</td>
<td>11103 (92.4)</td>
<td>660 (5.5)</td>
<td>335 (2.8)</td>
<td>12.9</td>
<td>25.4</td>
</tr>
<tr>
<td>54</td>
<td>11353</td>
<td>107 (0.9)</td>
<td>565 (5.0)</td>
<td>236 (2.1)</td>
<td>10445 (92.0)</td>
<td>672 (5.9)</td>
<td>343 (3.0)</td>
<td>15.9</td>
<td>31.2</td>
</tr>
<tr>
<td>60</td>
<td>10681</td>
<td>84 (0.8)</td>
<td>535 (5.0)</td>
<td>236 (2.2)</td>
<td>9826 (92.0)</td>
<td>619 (5.8)</td>
<td>320 (3.0)</td>
<td>13.6</td>
<td>26.2</td>
</tr>
<tr>
<td>All</td>
<td>156898</td>
<td>2488 (1.6)</td>
<td>7450 (4.7)</td>
<td>3089 (2.0)</td>
<td>143871 (91.7)</td>
<td>9938 (6.3)</td>
<td>5577 (3.6)</td>
<td>25.0</td>
<td>44.6</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased

Table 32: Results using reference strategy with increased between-individual variability by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased*n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1279 (6.4)</td>
<td>5806 (29.0)</td>
<td>656 (3.3)</td>
<td>12259 (61.3)</td>
<td>7085 (35.4)</td>
<td>1935 (9.7)</td>
<td>18.1</td>
<td>66.1</td>
</tr>
<tr>
<td>6</td>
<td>12915</td>
<td>104 (0.8)</td>
<td>1193 (9.2)</td>
<td>685 (5.3)</td>
<td>10933 (84.7)</td>
<td>1297 (10.0)</td>
<td>789 (6.1)</td>
<td>8.0</td>
<td>13.2</td>
</tr>
<tr>
<td>12</td>
<td>11618</td>
<td>78 (0.7)</td>
<td>769 (6.6)</td>
<td>707 (6.1)</td>
<td>10065 (86.6)</td>
<td>846 (7.3)</td>
<td>785 (6.8)</td>
<td>9.2</td>
<td>9.9</td>
</tr>
<tr>
<td>18</td>
<td>10772</td>
<td>81 (0.8)</td>
<td>598 (5.5)</td>
<td>734 (6.8)</td>
<td>9368 (87.0)</td>
<td>670 (6.2)</td>
<td>815 (7.6)</td>
<td>12.1</td>
<td>9.9</td>
</tr>
<tr>
<td>24</td>
<td>10102</td>
<td>85 (0.8)</td>
<td>468 (4.6)</td>
<td>757 (7.5)</td>
<td>8792 (87.0)</td>
<td>553 (5.5)</td>
<td>842 (8.3)</td>
<td>15.4</td>
<td>10.1</td>
</tr>
<tr>
<td>30</td>
<td>9549</td>
<td>84 (0.9)</td>
<td>392 (4.1)</td>
<td>745 (7.8)</td>
<td>8328 (87.2)</td>
<td>476 (5.0)</td>
<td>829 (8.7)</td>
<td>17.6</td>
<td>10.1</td>
</tr>
<tr>
<td>36</td>
<td>9073</td>
<td>78 (0.9)</td>
<td>368 (4.1)</td>
<td>758 (8.4)</td>
<td>7869 (86.7)</td>
<td>446 (4.9)</td>
<td>836 (9.2)</td>
<td>17.5</td>
<td>9.3</td>
</tr>
<tr>
<td>42</td>
<td>8627</td>
<td>85 (1.0)</td>
<td>356 (4.1)</td>
<td>774 (9.0)</td>
<td>7412 (85.9)</td>
<td>441 (5.1)</td>
<td>839 (10.0)</td>
<td>19.3</td>
<td>9.9</td>
</tr>
<tr>
<td>48</td>
<td>8186</td>
<td>90 (1.1)</td>
<td>302 (3.7)</td>
<td>768 (9.4)</td>
<td>7026 (85.8)</td>
<td>392 (4.8)</td>
<td>838 (10.5)</td>
<td>23.0</td>
<td>10.5</td>
</tr>
<tr>
<td>54</td>
<td>7794</td>
<td>99 (1.3)</td>
<td>285 (3.7)</td>
<td>779 (10.0)</td>
<td>6631 (85.1)</td>
<td>384 (4.9)</td>
<td>878 (11.3)</td>
<td>25.8</td>
<td>11.3</td>
</tr>
<tr>
<td>60</td>
<td>7410</td>
<td>79 (1.1)</td>
<td>305 (4.1)</td>
<td>794 (10.7)</td>
<td>6232 (84.1)</td>
<td>384 (5.2)</td>
<td>873 (11.8)</td>
<td>20.6</td>
<td>9.0</td>
</tr>
<tr>
<td>All</td>
<td>116046</td>
<td>2142 (1.8)</td>
<td>10832 (9.3)</td>
<td>8157 (7.0)</td>
<td>94915 (81.8)</td>
<td>12974 (11.2)</td>
<td>10299 (8.9)</td>
<td>16.5</td>
<td>20.8</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased
Table 33: Results using reference strategy with increased between-individual variability by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>890 (4.4)</td>
<td>2368 (11.8)</td>
<td>1045 (5.2)</td>
<td>15697 (78.5)</td>
<td>3258 (16.3)</td>
<td>1935 (9.7)</td>
<td>27.3</td>
<td>46.0</td>
</tr>
<tr>
<td>6</td>
<td>16742</td>
<td>117 (0.7)</td>
<td>610 (3.6)</td>
<td>1127 (6.7)</td>
<td>14888 (88.9)</td>
<td>727 (4.3)</td>
<td>1244 (7.4)</td>
<td>16.1</td>
<td>9.4</td>
</tr>
<tr>
<td>12</td>
<td>16015</td>
<td>98 (0.6)</td>
<td>413 (2.6)</td>
<td>1208 (7.5)</td>
<td>14296 (89.3)</td>
<td>511 (3.2)</td>
<td>1306 (8.2)</td>
<td>19.2</td>
<td>7.5</td>
</tr>
<tr>
<td>18</td>
<td>15504</td>
<td>102 (0.7)</td>
<td>339 (2.2)</td>
<td>1286 (8.3)</td>
<td>13777 (88.9)</td>
<td>441 (2.8)</td>
<td>1388 (9.0)</td>
<td>23.1</td>
<td>7.3</td>
</tr>
<tr>
<td>24</td>
<td>15063</td>
<td>81 (0.5)</td>
<td>322 (2.1)</td>
<td>1399 (9.3)</td>
<td>13251 (88.0)</td>
<td>413 (2.7)</td>
<td>1480 (9.8)</td>
<td>19.6</td>
<td>5.5</td>
</tr>
<tr>
<td>30</td>
<td>14650</td>
<td>78 (0.5)</td>
<td>308 (2.1)</td>
<td>1490 (10.2)</td>
<td>12774 (87.2)</td>
<td>386 (2.6)</td>
<td>1568 (10.7)</td>
<td>20.2</td>
<td>5.0</td>
</tr>
<tr>
<td>36</td>
<td>14264</td>
<td>114 (0.8)</td>
<td>295 (2.1)</td>
<td>1535 (10.8)</td>
<td>12320 (86.4)</td>
<td>409 (2.9)</td>
<td>1649 (11.6)</td>
<td>27.9</td>
<td>6.9</td>
</tr>
<tr>
<td>42</td>
<td>13855</td>
<td>103 (0.7)</td>
<td>276 (2.0)</td>
<td>1606 (11.6)</td>
<td>11870 (85.7)</td>
<td>379 (2.7)</td>
<td>1709 (12.3)</td>
<td>27.2</td>
<td>6.0</td>
</tr>
<tr>
<td>48</td>
<td>13476</td>
<td>122 (0.9)</td>
<td>318 (2.4)</td>
<td>1641 (12.2)</td>
<td>11395 (84.6)</td>
<td>440 (3.3)</td>
<td>1763 (13.1)</td>
<td>27.7</td>
<td>6.9</td>
</tr>
<tr>
<td>54</td>
<td>13036</td>
<td>103 (0.8)</td>
<td>271 (2.1)</td>
<td>1742 (13.4)</td>
<td>10920 (83.8)</td>
<td>374 (2.9)</td>
<td>1845 (14.2)</td>
<td>27.5</td>
<td>5.6</td>
</tr>
<tr>
<td>60</td>
<td>12662</td>
<td>130 (1.0)</td>
<td>286 (2.3)</td>
<td>1817 (14.4)</td>
<td>10429 (82.4)</td>
<td>416 (3.3)</td>
<td>1947 (15.4)</td>
<td>31.2</td>
<td>6.7</td>
</tr>
<tr>
<td>All</td>
<td>165267</td>
<td>1938 (1.2)</td>
<td>5816 (3.5)</td>
<td>15896 (9.6)</td>
<td>141617 (85.7)</td>
<td>7754 (4.7)</td>
<td>17834 (10.8)</td>
<td>25.0</td>
<td>10.9</td>
</tr>
</tbody>
</table>

* Tests performed where the patient was diseased

Table 34: Results using reference strategy with decreased fibrosis progression rate by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1038 (5.2)</td>
<td>2122 (10.6)</td>
<td>698 (3.5)</td>
<td>16142 (80.7)</td>
<td>3160 (15.8)</td>
<td>1736 (8.7)</td>
<td>32.8</td>
<td>59.8</td>
</tr>
<tr>
<td>6</td>
<td>16840</td>
<td>172 (1.0)</td>
<td>848 (5.0)</td>
<td>631 (3.7)</td>
<td>15189 (90.2)</td>
<td>1020 (6.1)</td>
<td>803 (4.8)</td>
<td>16.9</td>
<td>21.4</td>
</tr>
<tr>
<td>12</td>
<td>15820</td>
<td>110 (0.7)</td>
<td>603 (3.8)</td>
<td>617 (3.9)</td>
<td>14490 (91.6)</td>
<td>713 (4.5)</td>
<td>727 (4.6)</td>
<td>15.4</td>
<td>15.1</td>
</tr>
<tr>
<td>18</td>
<td>15107</td>
<td>97 (0.6)</td>
<td>580 (3.8)</td>
<td>624 (4.1)</td>
<td>13806 (91.4)</td>
<td>677 (4.5)</td>
<td>721 (4.8)</td>
<td>14.3</td>
<td>13.5</td>
</tr>
<tr>
<td>24</td>
<td>14430</td>
<td>86 (0.6)</td>
<td>448 (3.1)</td>
<td>648 (4.5)</td>
<td>13248 (91.8)</td>
<td>534 (3.7)</td>
<td>734 (5.1)</td>
<td>16.1</td>
<td>11.7</td>
</tr>
<tr>
<td>30</td>
<td>13896</td>
<td>84 (0.6)</td>
<td>402 (2.9)</td>
<td>655 (4.7)</td>
<td>12755 (91.8)</td>
<td>486 (3.5)</td>
<td>739 (5.3)</td>
<td>17.3</td>
<td>11.4</td>
</tr>
<tr>
<td>36</td>
<td>13410</td>
<td>82 (0.6)</td>
<td>459 (3.4)</td>
<td>677 (5.0)</td>
<td>12192 (90.9)</td>
<td>541 (4.0)</td>
<td>759 (5.7)</td>
<td>15.2</td>
<td>10.8</td>
</tr>
<tr>
<td>42</td>
<td>12869</td>
<td>102 (0.8)</td>
<td>413 (3.2)</td>
<td>667 (5.2)</td>
<td>11687 (90.8)</td>
<td>515 (4.0)</td>
<td>769 (6.0)</td>
<td>19.8</td>
<td>13.3</td>
</tr>
<tr>
<td>48</td>
<td>12354</td>
<td>108 (0.9)</td>
<td>416 (3.4)</td>
<td>646 (5.2)</td>
<td>11184 (90.5)</td>
<td>524 (4.2)</td>
<td>754 (6.1)</td>
<td>20.6</td>
<td>14.3</td>
</tr>
<tr>
<td>54</td>
<td>11830</td>
<td>84 (0.7)</td>
<td>371 (3.1)</td>
<td>663 (5.6)</td>
<td>10712 (90.5)</td>
<td>455 (3.8)</td>
<td>747 (6.3)</td>
<td>18.5</td>
<td>11.2</td>
</tr>
<tr>
<td>60</td>
<td>11375</td>
<td>96 (0.8)</td>
<td>351 (3.1)</td>
<td>678 (6.0)</td>
<td>10250 (90.1)</td>
<td>447 (3.9)</td>
<td>774 (6.8)</td>
<td>21.5</td>
<td>12.4</td>
</tr>
<tr>
<td>All</td>
<td>157931</td>
<td>2059 (1.3)</td>
<td>7013 (4.4)</td>
<td>7204 (4.6)</td>
<td>141655 (89.7)</td>
<td>9072 (5.7)</td>
<td>9263 (5.9)</td>
<td>22.7</td>
<td>22.2</td>
</tr>
</tbody>
</table>

* Tests performed where the patient was diseased
### Table 35: Results using reference strategy with decreased fibrosis progression rate by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(^a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>960 (4.8)</td>
<td>1707 (8.5)</td>
<td>776 (3.9)</td>
<td>16557 (82.8)</td>
<td>2667 (13.3)</td>
<td>1736 (8.7)</td>
<td>36.0</td>
<td>55.3</td>
</tr>
<tr>
<td>6</td>
<td>17333</td>
<td>184 (1.1)</td>
<td>730 (4.2)</td>
<td>707 (4.1)</td>
<td>15712 (90.6)</td>
<td>914 (5.3)</td>
<td>891 (5.1)</td>
<td>20.1</td>
<td>20.7</td>
</tr>
<tr>
<td>12</td>
<td>16419</td>
<td>125 (0.8)</td>
<td>516 (3.1)</td>
<td>692 (4.2)</td>
<td>15086 (91.9)</td>
<td>641 (3.9)</td>
<td>817 (5.0)</td>
<td>19.5</td>
<td>15.3</td>
</tr>
<tr>
<td>18</td>
<td>15778</td>
<td>101 (0.6)</td>
<td>466 (3.0)</td>
<td>706 (4.5)</td>
<td>14505 (91.9)</td>
<td>567 (3.6)</td>
<td>807 (5.1)</td>
<td>17.8</td>
<td>12.5</td>
</tr>
<tr>
<td>24</td>
<td>15211</td>
<td>96 (0.6)</td>
<td>395 (2.6)</td>
<td>735 (4.8)</td>
<td>13985 (91.9)</td>
<td>491 (3.2)</td>
<td>831 (5.5)</td>
<td>19.6</td>
<td>11.6</td>
</tr>
<tr>
<td>30</td>
<td>14720</td>
<td>77 (0.5)</td>
<td>393 (2.7)</td>
<td>760 (5.2)</td>
<td>13490 (91.6)</td>
<td>470 (3.2)</td>
<td>837 (5.7)</td>
<td>16.4</td>
<td>9.2</td>
</tr>
<tr>
<td>36</td>
<td>14250</td>
<td>83 (0.6)</td>
<td>427 (3.0)</td>
<td>792 (5.6)</td>
<td>12948 (90.9)</td>
<td>510 (3.6)</td>
<td>875 (6.1)</td>
<td>16.3</td>
<td>9.5</td>
</tr>
<tr>
<td>42</td>
<td>13740</td>
<td>101 (0.7)</td>
<td>374 (2.7)</td>
<td>793 (5.8)</td>
<td>12472 (90.8)</td>
<td>475 (3.5)</td>
<td>894 (6.5)</td>
<td>21.3</td>
<td>11.3</td>
</tr>
<tr>
<td>48</td>
<td>13265</td>
<td>110 (0.8)</td>
<td>381 (2.9)</td>
<td>783 (5.9)</td>
<td>11991 (90.4)</td>
<td>491 (3.7)</td>
<td>893 (6.7)</td>
<td>22.4</td>
<td>12.3</td>
</tr>
<tr>
<td>54</td>
<td>12774</td>
<td>100 (0.8)</td>
<td>379 (3.0)</td>
<td>795 (6.2)</td>
<td>11500 (90.0)</td>
<td>479 (3.7)</td>
<td>895 (7.0)</td>
<td>20.9</td>
<td>11.2</td>
</tr>
<tr>
<td>60</td>
<td>12295</td>
<td>105 (0.9)</td>
<td>342 (2.8)</td>
<td>813 (6.6)</td>
<td>11035 (89.8)</td>
<td>447 (3.6)</td>
<td>918 (7.5)</td>
<td>23.5</td>
<td>11.4</td>
</tr>
<tr>
<td>All</td>
<td>165785</td>
<td>2042 (1.2)</td>
<td>6110 (3.7)</td>
<td>8352 (5.0)</td>
<td>149281 (90.0)</td>
<td>8152 (4.9)</td>
<td>10394 (6.3)</td>
<td>25.0</td>
<td>19.6</td>
</tr>
</tbody>
</table>

\(^a\)Tests performed where the patient was diseased

### Table 36: Results using reference strategy with increased fibrosis progression rate by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(^a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1535 (7.7)</td>
<td>2428 (12.1)</td>
<td>947 (4.7)</td>
<td>15090 (75.4)</td>
<td>3963 (19.8)</td>
<td>2482 (12.4)</td>
<td>38.7</td>
<td>61.8</td>
</tr>
<tr>
<td>6</td>
<td>16037</td>
<td>269 (1.7)</td>
<td>990 (6.2)</td>
<td>909 (5.7)</td>
<td>13869 (86.5)</td>
<td>1259 (7.9)</td>
<td>1178 (7.3)</td>
<td>21.4</td>
<td>22.8</td>
</tr>
<tr>
<td>12</td>
<td>14778</td>
<td>200 (1.4)</td>
<td>772 (5.2)</td>
<td>930 (6.3)</td>
<td>12876 (87.1)</td>
<td>972 (6.6)</td>
<td>1130 (7.6)</td>
<td>20.6</td>
<td>17.7</td>
</tr>
<tr>
<td>18</td>
<td>13806</td>
<td>186 (1.3)</td>
<td>649 (4.7)</td>
<td>945 (6.8)</td>
<td>12026 (87.1)</td>
<td>835 (6.0)</td>
<td>1131 (8.2)</td>
<td>22.3</td>
<td>16.4</td>
</tr>
<tr>
<td>24</td>
<td>12971</td>
<td>163 (1.3)</td>
<td>607 (4.7)</td>
<td>938 (7.2)</td>
<td>11263 (86.8)</td>
<td>770 (5.9)</td>
<td>1101 (8.5)</td>
<td>21.2</td>
<td>14.8</td>
</tr>
<tr>
<td>30</td>
<td>12201</td>
<td>165 (1.4)</td>
<td>557 (4.6)</td>
<td>975 (8.0)</td>
<td>10504 (86.1)</td>
<td>722 (5.9)</td>
<td>1140 (9.3)</td>
<td>22.9</td>
<td>14.5</td>
</tr>
<tr>
<td>36</td>
<td>11479</td>
<td>196 (1.7)</td>
<td>596 (5.2)</td>
<td>969 (8.4)</td>
<td>9718 (84.7)</td>
<td>792 (6.9)</td>
<td>1165 (10.1)</td>
<td>24.7</td>
<td>16.8</td>
</tr>
<tr>
<td>42</td>
<td>10687</td>
<td>171 (1.6)</td>
<td>506 (4.7)</td>
<td>974 (9.1)</td>
<td>9036 (84.6)</td>
<td>677 (6.3)</td>
<td>1145 (10.7)</td>
<td>25.3</td>
<td>14.9</td>
</tr>
<tr>
<td>48</td>
<td>10010</td>
<td>204 (2.0)</td>
<td>512 (5.1)</td>
<td>979 (9.8)</td>
<td>8315 (83.1)</td>
<td>716 (7.2)</td>
<td>1183 (11.8)</td>
<td>28.5</td>
<td>17.2</td>
</tr>
<tr>
<td>54</td>
<td>9294</td>
<td>197 (2.1)</td>
<td>440 (4.7)</td>
<td>1009 (10.9)</td>
<td>7648 (82.3)</td>
<td>637 (6.9)</td>
<td>1206 (13.0)</td>
<td>30.9</td>
<td>16.3</td>
</tr>
<tr>
<td>60</td>
<td>8657</td>
<td>205 (2.4)</td>
<td>413 (4.8)</td>
<td>1009 (11.7)</td>
<td>7030 (81.2)</td>
<td>618 (7.1)</td>
<td>1214 (14.0)</td>
<td>33.2</td>
<td>16.9</td>
</tr>
<tr>
<td>All</td>
<td>139920</td>
<td>3491 (2.5)</td>
<td>8470 (6.1)</td>
<td>10584 (7.6)</td>
<td>117375 (83.9)</td>
<td>11961 (8.5)</td>
<td>14075 (10.1)</td>
<td>29.2</td>
<td>24.8</td>
</tr>
</tbody>
</table>

\(^a\)Tests performed where the patient was diseased
Table 37: Results using reference strategy with increased fibrosis progression rate by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(^a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1704 (8.5)</td>
<td>3332 (16.7)</td>
<td>778 (3.9)</td>
<td>14186 (70.9)</td>
<td>5036 (25.2)</td>
<td>2482 (12.4)</td>
<td>33.8</td>
<td>68.7</td>
</tr>
<tr>
<td>6</td>
<td>14964</td>
<td>271 (1.8)</td>
<td>1308 (8.7)</td>
<td>709 (4.7)</td>
<td>12676 (84.7)</td>
<td>1579 (10.6)</td>
<td>980 (6.5)</td>
<td>17.2</td>
<td>27.7</td>
</tr>
<tr>
<td>12</td>
<td>13385</td>
<td>183 (1.4)</td>
<td>979 (7.3)</td>
<td>699 (5.2)</td>
<td>11524 (86.1)</td>
<td>1162 (8.7)</td>
<td>882 (6.6)</td>
<td>15.7</td>
<td>20.7</td>
</tr>
<tr>
<td>18</td>
<td>12223</td>
<td>174 (1.4)</td>
<td>743 (6.1)</td>
<td>705 (5.8)</td>
<td>10601 (86.7)</td>
<td>917 (7.5)</td>
<td>879 (7.2)</td>
<td>19.0</td>
<td>19.8</td>
</tr>
<tr>
<td>24</td>
<td>11306</td>
<td>147 (1.3)</td>
<td>691 (6.1)</td>
<td>680 (6.0)</td>
<td>9788 (86.6)</td>
<td>838 (7.4)</td>
<td>827 (7.3)</td>
<td>17.5</td>
<td>17.8</td>
</tr>
<tr>
<td>30</td>
<td>10468</td>
<td>139 (1.3)</td>
<td>566 (5.4)</td>
<td>693 (6.6)</td>
<td>9070 (86.6)</td>
<td>705 (6.7)</td>
<td>832 (7.9)</td>
<td>19.7</td>
<td>16.7</td>
</tr>
<tr>
<td>36</td>
<td>9763</td>
<td>139 (1.4)</td>
<td>591 (6.1)</td>
<td>699 (7.2)</td>
<td>8334 (85.4)</td>
<td>730 (7.5)</td>
<td>838 (8.6)</td>
<td>19.0</td>
<td>16.6</td>
</tr>
<tr>
<td>42</td>
<td>9033</td>
<td>146 (1.6)</td>
<td>553 (6.1)</td>
<td>684 (7.6)</td>
<td>7650 (84.7)</td>
<td>699 (7.7)</td>
<td>830 (9.2)</td>
<td>20.9</td>
<td>17.6</td>
</tr>
<tr>
<td>48</td>
<td>8334</td>
<td>159 (1.9)</td>
<td>472 (5.7)</td>
<td>689 (8.3)</td>
<td>7014 (84.2)</td>
<td>631 (7.6)</td>
<td>848 (10.2)</td>
<td>25.2</td>
<td>18.8</td>
</tr>
<tr>
<td>54</td>
<td>7703</td>
<td>162 (2.1)</td>
<td>465 (6.0)</td>
<td>706 (9.2)</td>
<td>6370 (82.7)</td>
<td>627 (8.1)</td>
<td>868 (11.3)</td>
<td>25.8</td>
<td>18.7</td>
</tr>
<tr>
<td>60</td>
<td>7076</td>
<td>163 (2.3)</td>
<td>450 (6.4)</td>
<td>712 (10.1)</td>
<td>5751 (81.3)</td>
<td>613 (8.7)</td>
<td>875 (12.4)</td>
<td>26.6</td>
<td>18.6</td>
</tr>
<tr>
<td>All</td>
<td>124255</td>
<td>3387 (2.7)</td>
<td>10150 (8.2)</td>
<td>7754 (6.2)</td>
<td>102964 (82.9)</td>
<td>13537 (10.9)</td>
<td>11141 (9.0)</td>
<td>25.0</td>
<td>30.4</td>
</tr>
</tbody>
</table>

\(^a\) Tests performed where the patient was diseased
<table>
<thead>
<tr>
<th>Strategy</th>
<th>Monitoring strategy components</th>
<th>PPV</th>
<th>Tests*</th>
<th>Delay to diagnosis*</th>
<th>Test performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Decision rule</td>
<td>Threshold value</td>
<td>Frequency (months)</td>
<td>Retest</td>
<td>Initial threshold</td>
</tr>
<tr>
<td>A</td>
<td>Simple threshold</td>
<td>10.46</td>
<td>0.5</td>
<td>FALSE</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>Simple threshold</td>
<td>10.325</td>
<td>0.5</td>
<td>TRUE</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>Simple threshold</td>
<td>10.265</td>
<td>1</td>
<td>FALSE</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>Absolute increase from initial value</td>
<td>1.138</td>
<td>0.5</td>
<td>FALSE</td>
<td>10.46</td>
</tr>
<tr>
<td>E</td>
<td>Absolute increase from last value</td>
<td>1.245</td>
<td>0.5</td>
<td>FALSE</td>
<td>10.46</td>
</tr>
<tr>
<td>F</td>
<td>Relative increase from initial value</td>
<td>1.122</td>
<td>0.5</td>
<td>FALSE</td>
<td>10.46</td>
</tr>
<tr>
<td>G</td>
<td>Relative increase from last value</td>
<td>1.154</td>
<td>0.5</td>
<td>FALSE</td>
<td>10.46</td>
</tr>
<tr>
<td>H</td>
<td>Linear regression</td>
<td>10.235</td>
<td>0.5</td>
<td>FALSE</td>
<td>10.46</td>
</tr>
</tbody>
</table>

*Tests over the duration of monitoring.

b Number of tests per person over the duration of monitoring.

c Patients with delayed diagnosis (delay from onset of disease to diagnosis of over 12 months).

d % of all patients with delay to diagnosis.

e % of patients that would reach cirrhosis within the trial period with delay to diagnosis.

f TP pp is the number of true positive results per person over the duration of monitoring.

g FP pp is the number of false positive results per person over the duration of monitoring.

192590 tests were carried out to generate 125091 results due to retests being used.
Table 39: Results by observation point for the reference strategy (strategy A) for adjusted fibrosis progression estimate data

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1704 (8.5)</td>
<td>3332 (16.7)</td>
<td>778 (3.9)</td>
<td>14186 (70.9)</td>
<td>5036 (25.2)</td>
<td>2482 (12.4)</td>
<td>33.8</td>
<td>68.7</td>
</tr>
<tr>
<td>6</td>
<td>14964</td>
<td>271 (1.8)</td>
<td>1308 (8.7)</td>
<td>709 (4.7)</td>
<td>12676 (84.7)</td>
<td>1579 (10.6)</td>
<td>980 (6.5)</td>
<td>17.2</td>
<td>27.7</td>
</tr>
<tr>
<td>12</td>
<td>13385</td>
<td>183 (1.4)</td>
<td>979 (7.3)</td>
<td>699 (5.2)</td>
<td>11524 (86.1)</td>
<td>1162 (8.7)</td>
<td>882 (6.6)</td>
<td>15.7</td>
<td>20.7</td>
</tr>
<tr>
<td>18</td>
<td>12223</td>
<td>174 (1.4)</td>
<td>743 (6.1)</td>
<td>705 (5.8)</td>
<td>10601 (86.7)</td>
<td>917 (7.5)</td>
<td>879 (7.2)</td>
<td>19.0</td>
<td>19.8</td>
</tr>
<tr>
<td>24</td>
<td>11306</td>
<td>147 (1.3)</td>
<td>691 (6.1)</td>
<td>680 (6.0)</td>
<td>9788 (86.6)</td>
<td>838 (7.4)</td>
<td>827 (7.3)</td>
<td>17.5</td>
<td>17.8</td>
</tr>
<tr>
<td>30</td>
<td>10468</td>
<td>139 (1.3)</td>
<td>566 (5.4)</td>
<td>693 (6.6)</td>
<td>9070 (86.6)</td>
<td>705 (6.7)</td>
<td>832 (7.9)</td>
<td>19.7</td>
<td>16.7</td>
</tr>
<tr>
<td>36</td>
<td>9763</td>
<td>139 (1.4)</td>
<td>591 (6.1)</td>
<td>699 (7.2)</td>
<td>8334 (85.4)</td>
<td>730 (7.5)</td>
<td>838 (8.6)</td>
<td>19.0</td>
<td>16.6</td>
</tr>
<tr>
<td>42</td>
<td>9033</td>
<td>146 (1.6)</td>
<td>553 (6.1)</td>
<td>684 (7.6)</td>
<td>7650 (84.7)</td>
<td>699 (7.7)</td>
<td>830 (9.2)</td>
<td>20.9</td>
<td>17.6</td>
</tr>
<tr>
<td>48</td>
<td>8334</td>
<td>159 (1.9)</td>
<td>472 (5.7)</td>
<td>689 (8.3)</td>
<td>7014 (84.2)</td>
<td>631 (7.6)</td>
<td>848 (10.2)</td>
<td>25.2</td>
<td>18.8</td>
</tr>
<tr>
<td>54</td>
<td>7703</td>
<td>162 (2.1)</td>
<td>465 (6.0)</td>
<td>706 (9.2)</td>
<td>6370 (82.7)</td>
<td>627 (8.1)</td>
<td>868 (11.3)</td>
<td>25.8</td>
<td>18.7</td>
</tr>
<tr>
<td>60</td>
<td>7076</td>
<td>163 (2.3)</td>
<td>450 (6.4)</td>
<td>712 (10.1)</td>
<td>5751 (81.3)</td>
<td>613 (8.7)</td>
<td>875 (12.4)</td>
<td>26.6</td>
<td>18.6</td>
</tr>
</tbody>
</table>

All 124255 3387 (2.7) 10150 (8.2) 7754 (6.2) 102964 (82.9) 13537 (10.9) 11141 (9.0) 25.0 30.4

*Tests performed where the patient was diseased

Table 40: Results using retest monitoring strategy (strategy B) by observation point for adjusted fibrosis progression estimate data

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests^a(n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>30364/20000</td>
<td>1805 (9.0)</td>
<td>3694 (18.5)</td>
<td>677 (3.4)</td>
<td>13824 (69.1)</td>
<td>5499 (27.5)</td>
<td>2482 (12.4)</td>
<td>32.8</td>
<td>72.7</td>
</tr>
<tr>
<td>6</td>
<td>22078/14501</td>
<td>196 (1.4)</td>
<td>1108 (7.6)</td>
<td>656 (4.5)</td>
<td>12541 (86.5)</td>
<td>1304 (9.0)</td>
<td>852 (5.9)</td>
<td>15.0</td>
<td>23.0</td>
</tr>
<tr>
<td>12</td>
<td>20017/13197</td>
<td>141 (1.1)</td>
<td>827 (6.3)</td>
<td>682 (5.2)</td>
<td>11547 (87.5)</td>
<td>968 (7.3)</td>
<td>823 (6.2)</td>
<td>14.6</td>
<td>17.1</td>
</tr>
<tr>
<td>18</td>
<td>18707/12229</td>
<td>139 (1.1)</td>
<td>691 (5.7)</td>
<td>709 (5.8)</td>
<td>10690 (87.4)</td>
<td>830 (6.8)</td>
<td>848 (6.9)</td>
<td>16.7</td>
<td>16.4</td>
</tr>
<tr>
<td>24</td>
<td>17432/11399</td>
<td>146 (1.3)</td>
<td>585 (5.1)</td>
<td>682 (6.0)</td>
<td>9986 (87.6)</td>
<td>731 (6.4)</td>
<td>828 (7.3)</td>
<td>20.0</td>
<td>17.6</td>
</tr>
<tr>
<td>30</td>
<td>16494/10668</td>
<td>132 (1.2)</td>
<td>564 (5.3)</td>
<td>706 (6.6)</td>
<td>9266 (86.9)</td>
<td>696 (6.5)</td>
<td>838 (7.9)</td>
<td>19.0</td>
<td>15.8</td>
</tr>
<tr>
<td>36</td>
<td>15442/9972</td>
<td>153 (1.5)</td>
<td>562 (5.6)</td>
<td>698 (7.0)</td>
<td>8559 (85.8)</td>
<td>715 (7.2)</td>
<td>851 (8.5)</td>
<td>21.4</td>
<td>18.0</td>
</tr>
<tr>
<td>42</td>
<td>14378/8257</td>
<td>145 (1.6)</td>
<td>537 (5.8)</td>
<td>686 (7.4)</td>
<td>7889 (85.2)</td>
<td>682 (7.4)</td>
<td>831 (9.0)</td>
<td>21.3</td>
<td>17.4</td>
</tr>
<tr>
<td>48</td>
<td>13460/8575</td>
<td>161 (1.9)</td>
<td>466 (5.4)</td>
<td>683 (8.0)</td>
<td>7265 (84.7)</td>
<td>627 (7.3)</td>
<td>844 (9.8)</td>
<td>25.7</td>
<td>19.1</td>
</tr>
<tr>
<td>54</td>
<td>12558/7948</td>
<td>141 (1.8)</td>
<td>462 (5.8)</td>
<td>703 (8.8)</td>
<td>6642 (83.6)</td>
<td>603 (7.6)</td>
<td>844 (10.6)</td>
<td>23.4</td>
<td>16.7</td>
</tr>
<tr>
<td>60</td>
<td>11660/7345</td>
<td>100 (1.4)</td>
<td>293 (4.0)</td>
<td>766 (10.4)</td>
<td>6186 (84.2)</td>
<td>393 (5.4)</td>
<td>866 (11.8)</td>
<td>25.4</td>
<td>11.5</td>
</tr>
</tbody>
</table>

All 192590/125091 3259 (2.6) 9789 (7.8) 7648 (6.1) 104395 (83.5) 13048 (10.4) 10907 (8.7) 25.0 29.9

^Tests performed/number of people tests performed on (number of results generated)

*Tests performed where the patient was diseased
Table 41: Results using reduced frequency of monitoring strategy (strategy C) by observation point for adjusted fibrosis progression estimate data

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1818 (9.1)</td>
<td>4227 (21.1)</td>
<td>664 (3.3)</td>
<td>13291 (66.5)</td>
<td>6045 (30.2)</td>
<td>2482 (12.4)</td>
<td>30.1</td>
<td>73.2</td>
</tr>
<tr>
<td>12</td>
<td>13955</td>
<td>383 (2.7)</td>
<td>1789 (12.8)</td>
<td>650 (4.7)</td>
<td>11133 (79.8)</td>
<td>2172 (15.6)</td>
<td>1033 (7.4)</td>
<td>17.6</td>
<td>37.1</td>
</tr>
<tr>
<td>24</td>
<td>11783</td>
<td>314 (2.7)</td>
<td>1327 (11.2)</td>
<td>634 (5.4)</td>
<td>9508 (80.7)</td>
<td>1641 (13.9)</td>
<td>948 (8.0)</td>
<td>19.1</td>
<td>33.1</td>
</tr>
<tr>
<td>36</td>
<td>10142</td>
<td>292 (2.9)</td>
<td>1133 (11.2)</td>
<td>654 (6.4)</td>
<td>8063 (79.5)</td>
<td>1425 (14.1)</td>
<td>946 (9.3)</td>
<td>20.5</td>
<td>30.9</td>
</tr>
<tr>
<td>48</td>
<td>8717</td>
<td>318 (3.6)</td>
<td>995 (11.4)</td>
<td>641 (7.4)</td>
<td>6763 (77.6)</td>
<td>1313 (15.1)</td>
<td>959 (11.0)</td>
<td>24.2</td>
<td>33.2</td>
</tr>
<tr>
<td>60</td>
<td>7404</td>
<td>331 (4.5)</td>
<td>887 (12.0)</td>
<td>651 (8.8)</td>
<td>5535 (74.8)</td>
<td>1218 (16.5)</td>
<td>982 (13.3)</td>
<td>27.2</td>
<td>33.7</td>
</tr>
<tr>
<td>All</td>
<td>72001</td>
<td>3456 (4.8)</td>
<td>10358 (14.4)</td>
<td>3894 (5.4)</td>
<td>54293 (75.4)</td>
<td>13814 (19.2)</td>
<td>7350 (10.2)</td>
<td>25.0</td>
<td>47.0</td>
</tr>
</tbody>
</table>

* Tests performed where the patient was diseased

Table 42: Results using absolute increase from start value monitoring strategy (strategy D) by observation point for adjusted fibrosis progression estimate data

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1704 (8.5)</td>
<td>3332 (16.7)</td>
<td>778 (3.9)</td>
<td>14186 (70.9)</td>
<td>5036 (25.2)</td>
<td>2482 (12.4)</td>
<td>33.8</td>
<td>68.7</td>
</tr>
<tr>
<td>6</td>
<td>14964</td>
<td>90 (0.6)</td>
<td>867 (5.8)</td>
<td>890 (5.9)</td>
<td>13117 (87.7)</td>
<td>957 (6.4)</td>
<td>980 (6.5)</td>
<td>9.4</td>
<td>9.2</td>
</tr>
<tr>
<td>12</td>
<td>14007</td>
<td>132 (0.9)</td>
<td>816 (5.8)</td>
<td>968 (6.9)</td>
<td>12091 (86.3)</td>
<td>948 (6.8)</td>
<td>1100 (7.9)</td>
<td>13.9</td>
<td>12.0</td>
</tr>
<tr>
<td>18</td>
<td>13059</td>
<td>172 (1.3)</td>
<td>852 (6.5)</td>
<td>1008 (7.7)</td>
<td>11027 (84.4)</td>
<td>1024 (7.8)</td>
<td>1180 (9.0)</td>
<td>16.8</td>
<td>14.6</td>
</tr>
<tr>
<td>24</td>
<td>12035</td>
<td>185 (1.5)</td>
<td>754 (6.3)</td>
<td>991 (8.2)</td>
<td>10105 (84.0)</td>
<td>939 (7.8)</td>
<td>1176 (9.8)</td>
<td>19.7</td>
<td>15.7</td>
</tr>
<tr>
<td>30</td>
<td>11096</td>
<td>222 (2.0)</td>
<td>780 (7.0)</td>
<td>966 (8.7)</td>
<td>9128 (82.3)</td>
<td>1002 (9.0)</td>
<td>1188 (10.7)</td>
<td>22.2</td>
<td>18.7</td>
</tr>
<tr>
<td>36</td>
<td>10094</td>
<td>233 (2.3)</td>
<td>785 (7.8)</td>
<td>882 (8.7)</td>
<td>8194 (81.2)</td>
<td>1018 (10.1)</td>
<td>1115 (11.0)</td>
<td>22.9</td>
<td>20.9</td>
</tr>
<tr>
<td>42</td>
<td>9076</td>
<td>231 (2.5)</td>
<td>794 (8.7)</td>
<td>789 (8.7)</td>
<td>7262 (80.0)</td>
<td>1025 (11.3)</td>
<td>1020 (11.2)</td>
<td>22.5</td>
<td>22.6</td>
</tr>
<tr>
<td>48</td>
<td>8051</td>
<td>249 (3.1)</td>
<td>732 (9.1)</td>
<td>688 (8.5)</td>
<td>6382 (79.3)</td>
<td>981 (12.2)</td>
<td>937 (11.6)</td>
<td>25.4</td>
<td>26.6</td>
</tr>
<tr>
<td>54</td>
<td>7070</td>
<td>224 (3.2)</td>
<td>599 (8.5)</td>
<td>594 (8.4)</td>
<td>5653 (80.0)</td>
<td>823 (11.6)</td>
<td>818 (11.6)</td>
<td>27.2</td>
<td>27.4</td>
</tr>
<tr>
<td>60</td>
<td>6247</td>
<td>190 (3.0)</td>
<td>566 (9.1)</td>
<td>527 (8.4)</td>
<td>4964 (79.5)</td>
<td>756 (12.1)</td>
<td>717 (11.5)</td>
<td>25.1</td>
<td>26.5</td>
</tr>
<tr>
<td>All</td>
<td>125699</td>
<td>3632 (2.9)</td>
<td>10877 (8.7)</td>
<td>9081 (7.2)</td>
<td>102109 (81.2)</td>
<td>14509 (11.5)</td>
<td>12713 (10.1)</td>
<td>25.0</td>
<td>28.6</td>
</tr>
</tbody>
</table>

* Tests performed where the patient was diseased
Table 43: Results using absolute increase from last value monitoring strategy (strategy E) by observation point for adjusted fibrosis progression estimate data

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1704 (8.5)</td>
<td>3332 (16.7)</td>
<td>778 (3.9)</td>
<td>14186 (70.9)</td>
<td>5036 (25.2)</td>
<td>2482 (12.4)</td>
<td>33.8</td>
<td>68.7</td>
</tr>
<tr>
<td>6</td>
<td>14964</td>
<td>63 (0.4)</td>
<td>629 (4.2)</td>
<td>917 (6.1)</td>
<td>13355 (89.2)</td>
<td>692 (4.6)</td>
<td>980 (6.5)</td>
<td>9.1</td>
<td>6.4</td>
</tr>
<tr>
<td>12</td>
<td>14272</td>
<td>61 (0.4)</td>
<td>600 (4.2)</td>
<td>1078 (7.6)</td>
<td>12533 (87.8)</td>
<td>661 (4.6)</td>
<td>1139 (8.0)</td>
<td>9.2</td>
<td>5.4</td>
</tr>
<tr>
<td>18</td>
<td>13611</td>
<td>69 (0.5)</td>
<td>534 (3.9)</td>
<td>1231 (9.0)</td>
<td>11777 (86.5)</td>
<td>603 (4.4)</td>
<td>1300 (9.6)</td>
<td>11.4</td>
<td>5.3</td>
</tr>
<tr>
<td>24</td>
<td>13008</td>
<td>76 (0.6)</td>
<td>451 (3.5)</td>
<td>1360 (10.5)</td>
<td>11121 (85.5)</td>
<td>527 (4.1)</td>
<td>1436 (11.0)</td>
<td>14.4</td>
<td>5.3</td>
</tr>
<tr>
<td>30</td>
<td>12481</td>
<td>89 (0.7)</td>
<td>445 (3.6)</td>
<td>1520 (12.2)</td>
<td>10427 (83.5)</td>
<td>534 (4.3)</td>
<td>1609 (12.9)</td>
<td>16.7</td>
<td>5.5</td>
</tr>
<tr>
<td>36</td>
<td>11947</td>
<td>83 (0.7)</td>
<td>425 (3.6)</td>
<td>1687 (14.1)</td>
<td>9752 (81.6)</td>
<td>508 (4.3)</td>
<td>1770 (14.8)</td>
<td>16.3</td>
<td>4.7</td>
</tr>
<tr>
<td>42</td>
<td>11439</td>
<td>101 (0.9)</td>
<td>431 (3.8)</td>
<td>1852 (16.2)</td>
<td>9055 (79.2)</td>
<td>532 (4.7)</td>
<td>1953 (17.1)</td>
<td>19.0</td>
<td>5.2</td>
</tr>
<tr>
<td>48</td>
<td>10907</td>
<td>135 (1.2)</td>
<td>428 (3.9)</td>
<td>2034 (18.6)</td>
<td>12192 (97.2)</td>
<td>563 (4.3)</td>
<td>2169 (19.9)</td>
<td>24.0</td>
<td>6.2</td>
</tr>
<tr>
<td>54</td>
<td>10344</td>
<td>123 (1.2)</td>
<td>345 (3.3)</td>
<td>2256 (21.8)</td>
<td>7620 (73.7)</td>
<td>468 (4.5)</td>
<td>2379 (23.0)</td>
<td>26.3</td>
<td>5.2</td>
</tr>
<tr>
<td>60</td>
<td>9876</td>
<td>147 (1.5)</td>
<td>316 (3.2)</td>
<td>2423 (24.5)</td>
<td>6990 (70.8)</td>
<td>463 (4.7)</td>
<td>2570 (26.0)</td>
<td>31.7</td>
<td>5.7</td>
</tr>
<tr>
<td>All</td>
<td>142849</td>
<td>2651 (1.9)</td>
<td>7936 (5.6)</td>
<td>17136 (12.0)</td>
<td>115126 (80.6)</td>
<td>10587 (7.4)</td>
<td>19787 (13.9)</td>
<td>25.0</td>
<td>13.4</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased

Table 44: Results using relative increase from start value monitoring strategy (strategy F) by observation point for adjusted fibrosis progression estimate data

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1704 (8.5)</td>
<td>3332 (16.7)</td>
<td>778 (3.9)</td>
<td>14186 (70.9)</td>
<td>5036 (25.2)</td>
<td>2482 (12.4)</td>
<td>33.8</td>
<td>68.7</td>
</tr>
<tr>
<td>6</td>
<td>14964</td>
<td>87 (0.6)</td>
<td>1020 (6.8)</td>
<td>893 (6.0)</td>
<td>12964 (86.6)</td>
<td>1107 (7.4)</td>
<td>980 (6.5)</td>
<td>7.9</td>
<td>8.9</td>
</tr>
<tr>
<td>12</td>
<td>13857</td>
<td>133 (1.0)</td>
<td>877 (6.3)</td>
<td>975 (7.0)</td>
<td>11872 (85.7)</td>
<td>1010 (7.3)</td>
<td>1108 (8.0)</td>
<td>13.2</td>
<td>12.0</td>
</tr>
<tr>
<td>18</td>
<td>12847</td>
<td>159 (1.2)</td>
<td>862 (6.7)</td>
<td>1028 (8.0)</td>
<td>10798 (84.1)</td>
<td>1021 (7.9)</td>
<td>1187 (9.2)</td>
<td>15.6</td>
<td>13.4</td>
</tr>
<tr>
<td>24</td>
<td>11826</td>
<td>182 (1.5)</td>
<td>784 (6.6)</td>
<td>1019 (8.6)</td>
<td>9841 (83.2)</td>
<td>966 (8.2)</td>
<td>1201 (10.2)</td>
<td>18.8</td>
<td>15.2</td>
</tr>
<tr>
<td>30</td>
<td>10860</td>
<td>227 (2.1)</td>
<td>737 (6.8)</td>
<td>994 (9.2)</td>
<td>8902 (82.0)</td>
<td>964 (8.9)</td>
<td>1221 (11.2)</td>
<td>23.5</td>
<td>18.6</td>
</tr>
<tr>
<td>36</td>
<td>9896</td>
<td>243 (2.5)</td>
<td>759 (7.7)</td>
<td>911 (9.2)</td>
<td>7983 (80.7)</td>
<td>1002 (10.1)</td>
<td>1154 (11.7)</td>
<td>24.3</td>
<td>21.1</td>
</tr>
<tr>
<td>42</td>
<td>8894</td>
<td>215 (2.4)</td>
<td>731 (8.2)</td>
<td>843 (9.5)</td>
<td>7105 (79.9)</td>
<td>946 (10.6)</td>
<td>1058 (11.9)</td>
<td>22.7</td>
<td>20.3</td>
</tr>
<tr>
<td>48</td>
<td>7948</td>
<td>255 (3.2)</td>
<td>701 (8.8)</td>
<td>737 (9.3)</td>
<td>6255 (78.7)</td>
<td>956 (12.0)</td>
<td>992 (12.5)</td>
<td>26.7</td>
<td>25.7</td>
</tr>
<tr>
<td>54</td>
<td>6992</td>
<td>241 (3.4)</td>
<td>598 (8.6)</td>
<td>637 (9.1)</td>
<td>5516 (78.9)</td>
<td>839 (12.0)</td>
<td>878 (12.6)</td>
<td>28.7</td>
<td>27.4</td>
</tr>
<tr>
<td>60</td>
<td>6153</td>
<td>196 (3.2)</td>
<td>522 (8.5)</td>
<td>558 (9.1)</td>
<td>4877 (79.3)</td>
<td>718 (11.7)</td>
<td>754 (12.3)</td>
<td>27.3</td>
<td>26.0</td>
</tr>
<tr>
<td>All</td>
<td>124237</td>
<td>3642 (2.9)</td>
<td>10923 (8.8)</td>
<td>9373 (7.5)</td>
<td>100299 (80.7)</td>
<td>14565 (11.7)</td>
<td>13015 (10.5)</td>
<td>25.0</td>
<td>28.0</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased
Table 45: Results using relative increase from last value monitoring strategy (strategy G) by observation point for adjusted fibrosis progression estimate data

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1704 (8.5)</td>
<td>3332 (16.7)</td>
<td>778 (3.9)</td>
<td>14186 (70.9)</td>
<td>5036 (25.2)</td>
<td>2482 (12.4)</td>
<td>33.8</td>
<td>68.7</td>
</tr>
<tr>
<td>6</td>
<td>14964</td>
<td>33 (0.2)</td>
<td>455 (3.0)</td>
<td>947 (6.3)</td>
<td>13529 (90.4)</td>
<td>488 (3.3)</td>
<td>980 (6.5)</td>
<td>6.8</td>
<td>3.4</td>
</tr>
<tr>
<td>12</td>
<td>14476</td>
<td>41 (0.3)</td>
<td>429 (3.0)</td>
<td>1135 (7.8)</td>
<td>12871 (88.9)</td>
<td>470 (3.2)</td>
<td>1176 (8.1)</td>
<td>8.7</td>
<td>3.5</td>
</tr>
<tr>
<td>18</td>
<td>14006</td>
<td>40 (0.3)</td>
<td>379 (2.7)</td>
<td>1321 (9.4)</td>
<td>12266 (87.6)</td>
<td>419 (3.0)</td>
<td>1361 (9.7)</td>
<td>9.5</td>
<td>2.9</td>
</tr>
<tr>
<td>24</td>
<td>13587</td>
<td>44 (0.3)</td>
<td>327 (2.4)</td>
<td>1508 (11.1)</td>
<td>11708 (86.2)</td>
<td>371 (2.7)</td>
<td>1552 (11.4)</td>
<td>11.9</td>
<td>2.8</td>
</tr>
<tr>
<td>30</td>
<td>13216</td>
<td>45 (0.3)</td>
<td>311 (2.4)</td>
<td>1741 (13.2)</td>
<td>11119 (84.1)</td>
<td>356 (2.7)</td>
<td>1786 (13.5)</td>
<td>12.6</td>
<td>2.5</td>
</tr>
<tr>
<td>36</td>
<td>12860</td>
<td>44 (0.3)</td>
<td>290 (2.3)</td>
<td>1978 (15.4)</td>
<td>10548 (82.0)</td>
<td>334 (2.6)</td>
<td>2022 (15.7)</td>
<td>13.2</td>
<td>2.2</td>
</tr>
<tr>
<td>42</td>
<td>12526</td>
<td>47 (0.4)</td>
<td>282 (2.3)</td>
<td>2254 (18.0)</td>
<td>9943 (79.4)</td>
<td>329 (2.6)</td>
<td>2301 (18.4)</td>
<td>14.3</td>
<td>2.0</td>
</tr>
<tr>
<td>48</td>
<td>12197</td>
<td>63 (0.5)</td>
<td>283 (2.3)</td>
<td>2565 (21.0)</td>
<td>9286 (76.1)</td>
<td>346 (2.8)</td>
<td>2628 (21.5)</td>
<td>18.2</td>
<td>2.4</td>
</tr>
<tr>
<td>54</td>
<td>11851</td>
<td>56 (0.5)</td>
<td>212 (1.8)</td>
<td>2923 (24.7)</td>
<td>8660 (73.1)</td>
<td>268 (2.3)</td>
<td>2979 (25.1)</td>
<td>20.9</td>
<td>1.9</td>
</tr>
<tr>
<td>60</td>
<td>11583</td>
<td>52 (0.4)</td>
<td>204 (1.8)</td>
<td>3267 (28.2)</td>
<td>8060 (69.6)</td>
<td>256 (2.2)</td>
<td>3319 (28.7)</td>
<td>20.3</td>
<td>1.6</td>
</tr>
<tr>
<td>All</td>
<td>151266</td>
<td>2169 (1.4)</td>
<td>6504 (4.3)</td>
<td>20417 (13.5)</td>
<td>122176 (80.8)</td>
<td>8673 (5.7)</td>
<td>22586 (14.9)</td>
<td>25.0</td>
<td>9.6</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased

Table 46: Results using linear regression monitoring strategy (strategy H) by observation point for adjusted fibrosis progression estimate data

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1704 (8.5)</td>
<td>3332 (16.7)</td>
<td>778 (3.9)</td>
<td>14186 (70.9)</td>
<td>5036 (25.2)</td>
<td>2482 (12.4)</td>
<td>33.8</td>
<td>68.7</td>
</tr>
<tr>
<td>6</td>
<td>14964</td>
<td>367 (2.5)</td>
<td>2070 (13.8)</td>
<td>613 (4.1)</td>
<td>11914 (79.6)</td>
<td>2437 (16.3)</td>
<td>980 (6.5)</td>
<td>15.1</td>
<td>37.4</td>
</tr>
<tr>
<td>12</td>
<td>12527</td>
<td>156 (1.2)</td>
<td>1032 (8.2)</td>
<td>607 (4.8)</td>
<td>10732 (85.7)</td>
<td>1188 (9.5)</td>
<td>763 (6.1)</td>
<td>13.1</td>
<td>20.4</td>
</tr>
<tr>
<td>18</td>
<td>11339</td>
<td>135 (1.2)</td>
<td>703 (6.2)</td>
<td>619 (5.5)</td>
<td>9882 (82.7)</td>
<td>838 (7.4)</td>
<td>754 (6.6)</td>
<td>16.1</td>
<td>17.9</td>
</tr>
<tr>
<td>24</td>
<td>10501</td>
<td>126 (1.2)</td>
<td>513 (4.9)</td>
<td>597 (5.7)</td>
<td>9265 (88.2)</td>
<td>639 (6.1)</td>
<td>723 (6.9)</td>
<td>19.7</td>
<td>17.4</td>
</tr>
<tr>
<td>30</td>
<td>9862</td>
<td>99 (1.0)</td>
<td>397 (4.0)</td>
<td>649 (6.6)</td>
<td>8717 (88.4)</td>
<td>496 (5.0)</td>
<td>748 (7.6)</td>
<td>20.0</td>
<td>13.2</td>
</tr>
<tr>
<td>36</td>
<td>9366</td>
<td>97 (1.0)</td>
<td>413 (4.4)</td>
<td>684 (7.3)</td>
<td>8172 (87.3)</td>
<td>510 (5.4)</td>
<td>781 (8.3)</td>
<td>19.0</td>
<td>12.4</td>
</tr>
<tr>
<td>42</td>
<td>8856</td>
<td>125 (1.4)</td>
<td>362 (4.1)</td>
<td>702 (7.9)</td>
<td>7667 (86.6)</td>
<td>487 (5.5)</td>
<td>827 (9.3)</td>
<td>25.7</td>
<td>15.1</td>
</tr>
<tr>
<td>48</td>
<td>8369</td>
<td>130 (1.6)</td>
<td>297 (3.5)</td>
<td>737 (8.8)</td>
<td>7205 (86.1)</td>
<td>427 (5.1)</td>
<td>867 (10.4)</td>
<td>30.4</td>
<td>15.0</td>
</tr>
<tr>
<td>54</td>
<td>7942</td>
<td>151 (1.9)</td>
<td>303 (3.8)</td>
<td>764 (9.6)</td>
<td>6724 (84.7)</td>
<td>454 (5.7)</td>
<td>915 (11.5)</td>
<td>33.3</td>
<td>16.5</td>
</tr>
<tr>
<td>60</td>
<td>7488</td>
<td>147 (2.0)</td>
<td>271 (3.6)</td>
<td>801 (10.7)</td>
<td>6269 (83.7)</td>
<td>418 (5.6)</td>
<td>948 (12.7)</td>
<td>35.2</td>
<td>15.5</td>
</tr>
<tr>
<td>All</td>
<td>121214</td>
<td>3237 (2.7)</td>
<td>9693 (8.0)</td>
<td>7551 (6.2)</td>
<td>100733 (83.1)</td>
<td>12930 (10.7)</td>
<td>10788 (8.9)</td>
<td>25.0</td>
<td>30.0</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased
<table>
<thead>
<tr>
<th>Change in data simulation</th>
<th>Threshold</th>
<th>PPV (%)</th>
<th>Number of tests per person(^a)</th>
<th>Delay(^b) (%)</th>
<th>Develop cirrhosis(^c) n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>10.460</td>
<td>25.0</td>
<td>6.21</td>
<td>5.63</td>
<td>7689 (38.45)</td>
</tr>
<tr>
<td>Decreased† measurement error</td>
<td>10.460</td>
<td>29.7</td>
<td>7.04</td>
<td>7.18</td>
<td>7664 (38.32)</td>
</tr>
<tr>
<td></td>
<td>10.205</td>
<td>25.0</td>
<td>6.22</td>
<td>5.16</td>
<td>(25 (0.13))</td>
</tr>
<tr>
<td>Increased‡ measurement error</td>
<td>10.460</td>
<td>19.9</td>
<td>4.64</td>
<td>3.86</td>
<td>7808 (39.04)</td>
</tr>
<tr>
<td></td>
<td>10.97</td>
<td>25.0</td>
<td>6.01</td>
<td>6.60</td>
<td>(119 (0.60))</td>
</tr>
<tr>
<td>Decrease† between-individual variability</td>
<td>10.460</td>
<td>28.5</td>
<td>7.10</td>
<td>2.97</td>
<td>7531 (37.66)</td>
</tr>
<tr>
<td></td>
<td>10.35</td>
<td>25.0</td>
<td>6.66</td>
<td>2.23</td>
<td>(158 (0.79))</td>
</tr>
<tr>
<td>Increased‡ between-individual variability</td>
<td>10.460</td>
<td>19.7</td>
<td>4.78</td>
<td>5.63</td>
<td>7659 (37.85)</td>
</tr>
<tr>
<td></td>
<td>11.32</td>
<td>25.0</td>
<td>6.68</td>
<td>10.08</td>
<td>(120 (0.60))</td>
</tr>
<tr>
<td>Decreased† fibrosis progression rate</td>
<td>10.460</td>
<td>20.7</td>
<td>6.85</td>
<td>4.72</td>
<td>5314 (26.57)</td>
</tr>
<tr>
<td></td>
<td>10.725</td>
<td>25.0</td>
<td>7.64</td>
<td>6.10</td>
<td>(2375 (11.88))</td>
</tr>
<tr>
<td>Increased‡ fibrosis progression rate</td>
<td>10.460</td>
<td>36.3</td>
<td>5.13</td>
<td>9.29</td>
<td>13967 (69.84)</td>
</tr>
<tr>
<td></td>
<td>9.765</td>
<td>25.0</td>
<td>3.14</td>
<td>2.81</td>
<td>(6278 (31.39))</td>
</tr>
</tbody>
</table>

\(^a\) Number of tests per person over the duration of monitoring.

\(^b\) % of all patients with delayed diagnosis (delay from onset of disease to diagnosis of over 12 months).

\(^c\) Patients that would go on to develop cirrhosis in the monitoring duration if no intervention were received.

† Decrease is halving the estimate used in the original simulation.

‡ Increase is doubling the estimate used in the original simulation.
### Table 48: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with decreased measurement error by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1705 (8.5)</td>
<td>2899 (14.5)</td>
<td>752 (3.8)</td>
<td>14644 (73.2)</td>
<td>4604 (23.0)</td>
<td>2457 (12.3)</td>
<td>37.0</td>
<td>69.4</td>
</tr>
<tr>
<td>6</td>
<td>15396</td>
<td>152 (1.0)</td>
<td>773 (5.0)</td>
<td>802 (5.2)</td>
<td>13669 (88.8)</td>
<td>925 (6.0)</td>
<td>954 (6.2)</td>
<td>16.4</td>
<td>15.9</td>
</tr>
<tr>
<td>12</td>
<td>14471</td>
<td>157 (1.1)</td>
<td>563 (3.9)</td>
<td>845 (5.8)</td>
<td>12906 (89.2)</td>
<td>720 (5.0)</td>
<td>1002 (6.9)</td>
<td>21.8</td>
<td>15.7</td>
</tr>
<tr>
<td>18</td>
<td>13751</td>
<td>150 (1.1)</td>
<td>543 (3.9)</td>
<td>899 (6.5)</td>
<td>12159 (88.4)</td>
<td>693 (5.0)</td>
<td>1049 (7.6)</td>
<td>21.6</td>
<td>14.3</td>
</tr>
<tr>
<td>24</td>
<td>13058</td>
<td>161 (1.2)</td>
<td>521 (4.0)</td>
<td>895 (6.9)</td>
<td>11481 (87.9)</td>
<td>682 (5.2)</td>
<td>1056 (8.1)</td>
<td>23.6</td>
<td>15.2</td>
</tr>
<tr>
<td>30</td>
<td>12376</td>
<td>168 (1.4)</td>
<td>490 (4.0)</td>
<td>928 (7.5)</td>
<td>10790 (87.2)</td>
<td>658 (5.3)</td>
<td>1096 (8.9)</td>
<td>25.5</td>
<td>15.3</td>
</tr>
<tr>
<td>36</td>
<td>11718</td>
<td>165 (1.4)</td>
<td>519 (4.4)</td>
<td>941 (8.0)</td>
<td>10093 (86.1)</td>
<td>684 (5.8)</td>
<td>1106 (9.4)</td>
<td>24.1</td>
<td>14.9</td>
</tr>
<tr>
<td>42</td>
<td>11034</td>
<td>166 (1.5)</td>
<td>502 (4.5)</td>
<td>959 (8.7)</td>
<td>9407 (85.3)</td>
<td>668 (6.1)</td>
<td>1125 (10.2)</td>
<td>24.9</td>
<td>14.8</td>
</tr>
<tr>
<td>48</td>
<td>10366</td>
<td>218 (2.1)</td>
<td>490 (4.7)</td>
<td>958 (9.2)</td>
<td>8700 (83.9)</td>
<td>708 (6.8)</td>
<td>1176 (11.3)</td>
<td>30.8</td>
<td>18.5</td>
</tr>
<tr>
<td>54</td>
<td>9658</td>
<td>191 (2.0)</td>
<td>462 (4.8)</td>
<td>978 (10.1)</td>
<td>8027 (83.1)</td>
<td>653 (6.8)</td>
<td>1169 (12.1)</td>
<td>29.2</td>
<td>16.3</td>
</tr>
<tr>
<td>60</td>
<td>9005</td>
<td>230 (2.6)</td>
<td>427 (4.7)</td>
<td>963 (10.7)</td>
<td>7385 (82.0)</td>
<td>657 (7.3)</td>
<td>1193 (13.2)</td>
<td>35.0</td>
<td>19.3</td>
</tr>
<tr>
<td>All</td>
<td>140833</td>
<td>3463 (2.5)</td>
<td>8189 (5.8)</td>
<td>9920 (7.0)</td>
<td>119261 (84.7)</td>
<td>11652 (8.3)</td>
<td>13383 (9.5)</td>
<td>29.7</td>
<td>25.9</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased

### Table 49: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with decreased measurement error by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1869 (9.3)</td>
<td>4113 (20.6)</td>
<td>588 (2.9)</td>
<td>13430 (67.2)</td>
<td>5982 (29.9)</td>
<td>2457 (12.3)</td>
<td>31.2</td>
<td>76.1</td>
</tr>
<tr>
<td>6</td>
<td>14018</td>
<td>142 (1.0)</td>
<td>926 (6.6)</td>
<td>607 (4.3)</td>
<td>12343 (88.1)</td>
<td>1068 (7.6)</td>
<td>749 (5.3)</td>
<td>13.3</td>
<td>19.0</td>
</tr>
<tr>
<td>12</td>
<td>12950</td>
<td>135 (1.0)</td>
<td>675 (5.2)</td>
<td>632 (4.9)</td>
<td>11508 (88.9)</td>
<td>810 (6.3)</td>
<td>767 (5.9)</td>
<td>16.7</td>
<td>17.6</td>
</tr>
<tr>
<td>18</td>
<td>12140</td>
<td>133 (1.1)</td>
<td>661 (5.4)</td>
<td>657 (5.4)</td>
<td>10689 (88.0)</td>
<td>794 (6.5)</td>
<td>790 (6.5)</td>
<td>16.8</td>
<td>16.8</td>
</tr>
<tr>
<td>24</td>
<td>11346</td>
<td>119 (1.0)</td>
<td>555 (4.9)</td>
<td>655 (5.8)</td>
<td>10017 (88.3)</td>
<td>674 (5.9)</td>
<td>774 (6.8)</td>
<td>17.7</td>
<td>15.4</td>
</tr>
<tr>
<td>30</td>
<td>10672</td>
<td>139 (1.3)</td>
<td>598 (5.6)</td>
<td>664 (6.2)</td>
<td>9271 (86.9)</td>
<td>737 (6.9)</td>
<td>803 (7.5)</td>
<td>18.9</td>
<td>17.3</td>
</tr>
<tr>
<td>36</td>
<td>9935</td>
<td>138 (1.4)</td>
<td>525 (5.3)</td>
<td>660 (6.6)</td>
<td>8612 (86.7)</td>
<td>663 (6.7)</td>
<td>798 (8.0)</td>
<td>20.8</td>
<td>17.3</td>
</tr>
<tr>
<td>42</td>
<td>9272</td>
<td>121 (1.3)</td>
<td>537 (5.8)</td>
<td>673 (7.3)</td>
<td>7941 (85.6)</td>
<td>658 (7.1)</td>
<td>794 (8.6)</td>
<td>18.4</td>
<td>15.2</td>
</tr>
<tr>
<td>48</td>
<td>8614</td>
<td>170 (2.0)</td>
<td>451 (5.2)</td>
<td>647 (7.5)</td>
<td>7346 (85.3)</td>
<td>621 (7.2)</td>
<td>817 (9.5)</td>
<td>27.4</td>
<td>20.8</td>
</tr>
<tr>
<td>54</td>
<td>7993</td>
<td>165 (2.1)</td>
<td>464 (5.8)</td>
<td>653 (8.2)</td>
<td>6711 (84.0)</td>
<td>629 (7.9)</td>
<td>818 (10.2)</td>
<td>26.2</td>
<td>20.2</td>
</tr>
<tr>
<td>60</td>
<td>7364</td>
<td>190 (2.6)</td>
<td>439 (6.0)</td>
<td>624 (8.5)</td>
<td>6111 (83.0)</td>
<td>629 (8.5)</td>
<td>814 (11.1)</td>
<td>30.2</td>
<td>23.3</td>
</tr>
<tr>
<td>All</td>
<td>124304</td>
<td>3321 (2.7)</td>
<td>9944 (8.0)</td>
<td>7060 (5.7)</td>
<td>103979 (83.6)</td>
<td>13265 (10.7)</td>
<td>10381 (8.4)</td>
<td>25.0</td>
<td>32.0</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased
Table 50: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with increased measurement error by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1790 (8.9)</td>
<td>4412 (22.1)</td>
<td>773 (3.9)</td>
<td>13025 (65.1)</td>
<td>6202 (31.0)</td>
<td>2563 (12.8)</td>
<td>28.9</td>
<td>69.8</td>
</tr>
<tr>
<td>6</td>
<td>13798</td>
<td>385 (2.8)</td>
<td>2254 (16.3)</td>
<td>608 (4.4)</td>
<td>10551 (76.5)</td>
<td>2639 (19.1)</td>
<td>993 (7.2)</td>
<td>14.6</td>
<td>38.8</td>
</tr>
<tr>
<td>12</td>
<td>11159</td>
<td>243 (2.2)</td>
<td>1572 (14.1)</td>
<td>532 (4.8)</td>
<td>8812 (79.0)</td>
<td>1815 (16.3)</td>
<td>775 (6.9)</td>
<td>13.4</td>
<td>31.4</td>
</tr>
<tr>
<td>18</td>
<td>9344</td>
<td>169 (1.8)</td>
<td>1255 (13.4)</td>
<td>484 (5.2)</td>
<td>7436 (79.6)</td>
<td>1424 (15.2)</td>
<td>653 (7.0)</td>
<td>11.9</td>
<td>25.9</td>
</tr>
<tr>
<td>24</td>
<td>7920</td>
<td>123 (1.6)</td>
<td>892 (11.3)</td>
<td>452 (5.7)</td>
<td>6453 (81.5)</td>
<td>1015 (12.8)</td>
<td>575 (7.3)</td>
<td>12.1</td>
<td>21.4</td>
</tr>
<tr>
<td>30</td>
<td>6905</td>
<td>130 (1.9)</td>
<td>738 (10.7)</td>
<td>432 (6.3)</td>
<td>5605 (81.2)</td>
<td>868 (12.6)</td>
<td>562 (8.1)</td>
<td>15.0</td>
<td>23.1</td>
</tr>
<tr>
<td>36</td>
<td>6037</td>
<td>105 (1.7)</td>
<td>636 (10.5)</td>
<td>406 (6.7)</td>
<td>4890 (81.0)</td>
<td>741 (12.3)</td>
<td>511 (8.5)</td>
<td>14.2</td>
<td>20.5</td>
</tr>
<tr>
<td>42</td>
<td>5296</td>
<td>95 (1.8)</td>
<td>528 (10.0)</td>
<td>380 (7.2)</td>
<td>4293 (81.1)</td>
<td>623 (11.8)</td>
<td>475 (9.0)</td>
<td>15.2</td>
<td>20.0</td>
</tr>
<tr>
<td>48</td>
<td>4673</td>
<td>99 (2.1)</td>
<td>464 (9.9)</td>
<td>372 (8.0)</td>
<td>3738 (80.0)</td>
<td>563 (12.0)</td>
<td>471 (10.1)</td>
<td>17.6</td>
<td>21.0</td>
</tr>
<tr>
<td>54</td>
<td>4110</td>
<td>89 (2.2)</td>
<td>378 (9.2)</td>
<td>374 (9.1)</td>
<td>3269 (79.5)</td>
<td>467 (11.4)</td>
<td>463 (11.3)</td>
<td>19.1</td>
<td>19.2</td>
</tr>
<tr>
<td>60</td>
<td>3643</td>
<td>101 (2.8)</td>
<td>312 (8.6)</td>
<td>363 (10.0)</td>
<td>2867 (78.7)</td>
<td>413 (11.3)</td>
<td>464 (12.7)</td>
<td>24.5</td>
<td>21.8</td>
</tr>
<tr>
<td>All</td>
<td>92885</td>
<td>3329 (3.6)</td>
<td>13441 (14.5)</td>
<td>5176 (5.6)</td>
<td>70939 (76.4)</td>
<td>16770 (18.1)</td>
<td>8505 (9.2)</td>
<td>19.9</td>
<td>39.1</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased.

Table 51: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with increased measurement error by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1468 (7.3)</td>
<td>2728 (13.6)</td>
<td>1095 (5.5)</td>
<td>14709 (73.5)</td>
<td>4196 (21.0)</td>
<td>2563 (12.8)</td>
<td>35.0</td>
<td>57.3</td>
</tr>
<tr>
<td>6</td>
<td>15804</td>
<td>464 (2.9)</td>
<td>1604 (10.1)</td>
<td>916 (5.8)</td>
<td>12820 (81.1)</td>
<td>2068 (13.1)</td>
<td>1380 (8.7)</td>
<td>22.4</td>
<td>33.6</td>
</tr>
<tr>
<td>12</td>
<td>13736</td>
<td>278 (2.0)</td>
<td>1252 (9.1)</td>
<td>871 (6.3)</td>
<td>11335 (82.5)</td>
<td>1530 (11.1)</td>
<td>1149 (8.4)</td>
<td>18.2</td>
<td>24.2</td>
</tr>
<tr>
<td>18</td>
<td>12206</td>
<td>222 (1.8)</td>
<td>1015 (8.3)</td>
<td>826 (6.8)</td>
<td>10143 (83.1)</td>
<td>1237 (10.1)</td>
<td>1048 (8.6)</td>
<td>17.9</td>
<td>21.2</td>
</tr>
<tr>
<td>24</td>
<td>10969</td>
<td>206 (1.9)</td>
<td>811 (7.4)</td>
<td>768 (7.0)</td>
<td>9184 (83.7)</td>
<td>1017 (9.3)</td>
<td>974 (8.9)</td>
<td>20.3</td>
<td>21.1</td>
</tr>
<tr>
<td>30</td>
<td>9952</td>
<td>177 (1.8)</td>
<td>750 (7.5)</td>
<td>760 (7.6)</td>
<td>8265 (83.0)</td>
<td>927 (9.3)</td>
<td>937 (9.4)</td>
<td>19.1</td>
<td>18.9</td>
</tr>
<tr>
<td>36</td>
<td>9025</td>
<td>168 (1.9)</td>
<td>651 (7.2)</td>
<td>735 (8.1)</td>
<td>7471 (82.8)</td>
<td>819 (9.1)</td>
<td>903 (10.0)</td>
<td>20.5</td>
<td>18.6</td>
</tr>
<tr>
<td>42</td>
<td>8206</td>
<td>145 (1.8)</td>
<td>582 (7.1)</td>
<td>723 (8.8)</td>
<td>6756 (82.3)</td>
<td>727 (8.9)</td>
<td>868 (10.6)</td>
<td>19.9</td>
<td>16.7</td>
</tr>
<tr>
<td>48</td>
<td>7479</td>
<td>173 (2.3)</td>
<td>543 (7.3)</td>
<td>709 (9.5)</td>
<td>6054 (80.9)</td>
<td>716 (9.6)</td>
<td>882 (11.8)</td>
<td>24.2</td>
<td>19.6</td>
</tr>
<tr>
<td>54</td>
<td>6763</td>
<td>159 (2.4)</td>
<td>460 (6.8)</td>
<td>721 (10.7)</td>
<td>5423 (80.2)</td>
<td>619 (9.2)</td>
<td>880 (13.0)</td>
<td>25.7</td>
<td>18.1</td>
</tr>
<tr>
<td>60</td>
<td>6144</td>
<td>139 (2.3)</td>
<td>388 (6.3)</td>
<td>731 (11.9)</td>
<td>4886 (79.5)</td>
<td>527 (8.6)</td>
<td>870 (14.2)</td>
<td>26.4</td>
<td>16.0</td>
</tr>
<tr>
<td>All</td>
<td>120284</td>
<td>3599 (3.0)</td>
<td>10784 (9.0)</td>
<td>8855 (7.4)</td>
<td>97046 (80.7)</td>
<td>14383 (12.0)</td>
<td>12454 (10.4)</td>
<td>25.0</td>
<td>28.9</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased.
Table 52: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with decreased between-individual variability by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1809 (9.0)</td>
<td>1745 (8.7)</td>
<td>546 (2.7)</td>
<td>15900 (79.5)</td>
<td>3554 (17.8)</td>
<td>2355 (11.8)</td>
<td>50.9</td>
<td>76.8</td>
</tr>
<tr>
<td>6</td>
<td>16446</td>
<td>349 (2.1)</td>
<td>925 (5.6)</td>
<td>397 (2.4)</td>
<td>14775 (89.8)</td>
<td>1274 (7.7)</td>
<td>746 (4.5)</td>
<td>27.4</td>
<td>46.8</td>
</tr>
<tr>
<td>12</td>
<td>15172</td>
<td>189 (1.2)</td>
<td>771 (5.1)</td>
<td>368 (2.4)</td>
<td>13844 (91.2)</td>
<td>960 (6.3)</td>
<td>557 (3.7)</td>
<td>19.7</td>
<td>33.9</td>
</tr>
<tr>
<td>18</td>
<td>14212</td>
<td>143 (1.0)</td>
<td>731 (5.1)</td>
<td>367 (2.6)</td>
<td>12971 (91.3)</td>
<td>874 (6.1)</td>
<td>510 (3.6)</td>
<td>16.4</td>
<td>28.0</td>
</tr>
<tr>
<td>24</td>
<td>13338</td>
<td>139 (1.0)</td>
<td>668 (5.0)</td>
<td>335 (2.5)</td>
<td>12196 (91.4)</td>
<td>807 (6.1)</td>
<td>474 (3.6)</td>
<td>17.2</td>
<td>29.3</td>
</tr>
<tr>
<td>30</td>
<td>12531</td>
<td>141 (1.1)</td>
<td>660 (5.3)</td>
<td>334 (2.7)</td>
<td>11396 (90.9)</td>
<td>801 (6.4)</td>
<td>475 (3.8)</td>
<td>17.6</td>
<td>29.7</td>
</tr>
<tr>
<td>36</td>
<td>11730</td>
<td>140 (1.2)</td>
<td>695 (5.9)</td>
<td>334 (2.8)</td>
<td>10561 (90.0)</td>
<td>835 (7.1)</td>
<td>474 (4.0)</td>
<td>16.8</td>
<td>29.5</td>
</tr>
<tr>
<td>42</td>
<td>10895</td>
<td>153 (1.4)</td>
<td>665 (6.1)</td>
<td>304 (2.8)</td>
<td>9773 (89.7)</td>
<td>818 (7.5)</td>
<td>457 (4.2)</td>
<td>18.7</td>
<td>33.5</td>
</tr>
<tr>
<td>48</td>
<td>10077</td>
<td>132 (1.3)</td>
<td>698 (6.9)</td>
<td>332 (3.3)</td>
<td>8915 (88.5)</td>
<td>830 (8.2)</td>
<td>464 (4.6)</td>
<td>15.9</td>
<td>28.4</td>
</tr>
<tr>
<td>54</td>
<td>9247</td>
<td>153 (1.7)</td>
<td>649 (7.0)</td>
<td>326 (3.5)</td>
<td>8119 (87.8)</td>
<td>802 (8.7)</td>
<td>479 (5.2)</td>
<td>19.1</td>
<td>31.9</td>
</tr>
<tr>
<td>60</td>
<td>8445</td>
<td>143 (1.7)</td>
<td>560 (6.6)</td>
<td>341 (4.0)</td>
<td>7401 (87.6)</td>
<td>703 (8.3)</td>
<td>484 (5.7)</td>
<td>20.3</td>
<td>29.5</td>
</tr>
<tr>
<td>All</td>
<td>142093</td>
<td>3491 (2.5)</td>
<td>8767 (6.2)</td>
<td>3984 (2.8)</td>
<td>125851 (88.6)</td>
<td>12258 (8.6)</td>
<td>7475 (5.3)</td>
<td>28.5</td>
<td>46.7</td>
</tr>
</tbody>
</table>

* Tests performed where the patient was diseased

Table 53: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with decreased between-individual variability by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1917 (9.6)</td>
<td>2120 (10.6)</td>
<td>438 (2.2)</td>
<td>15525 (77.6)</td>
<td>4037 (20.2)</td>
<td>2355 (11.8)</td>
<td>47.5</td>
<td>81.4</td>
</tr>
<tr>
<td>6</td>
<td>15963</td>
<td>297 (1.9)</td>
<td>1119 (7.0)</td>
<td>321 (2.0)</td>
<td>14226 (89.1)</td>
<td>1416 (8.9)</td>
<td>618 (3.9)</td>
<td>21.0</td>
<td>48.1</td>
</tr>
<tr>
<td>12</td>
<td>14547</td>
<td>161 (1.1)</td>
<td>951 (6.5)</td>
<td>290 (2.0)</td>
<td>13145 (90.4)</td>
<td>1112 (7.6)</td>
<td>451 (3.1)</td>
<td>14.5</td>
<td>35.7</td>
</tr>
<tr>
<td>18</td>
<td>13435</td>
<td>127 (0.9)</td>
<td>849 (6.3)</td>
<td>287 (2.1)</td>
<td>12172 (90.6)</td>
<td>976 (7.3)</td>
<td>414 (3.1)</td>
<td>13.0</td>
<td>30.7</td>
</tr>
<tr>
<td>24</td>
<td>12459</td>
<td>121 (1.0)</td>
<td>757 (6.1)</td>
<td>247 (2.0)</td>
<td>11334 (91.0)</td>
<td>878 (7.0)</td>
<td>368 (3.0)</td>
<td>13.8</td>
<td>32.9</td>
</tr>
<tr>
<td>30</td>
<td>11581</td>
<td>122 (1.1)</td>
<td>722 (6.2)</td>
<td>235 (2.0)</td>
<td>10502 (90.7)</td>
<td>844 (7.3)</td>
<td>357 (3.1)</td>
<td>14.5</td>
<td>34.2</td>
</tr>
<tr>
<td>36</td>
<td>10737</td>
<td>119 (1.1)</td>
<td>757 (7.1)</td>
<td>222 (2.1)</td>
<td>9639 (89.8)</td>
<td>876 (8.2)</td>
<td>341 (3.2)</td>
<td>13.6</td>
<td>34.9</td>
</tr>
<tr>
<td>42</td>
<td>9861</td>
<td>109 (1.1)</td>
<td>721 (7.3)</td>
<td>211 (2.1)</td>
<td>8820 (89.4)</td>
<td>830 (8.4)</td>
<td>320 (3.2)</td>
<td>13.1</td>
<td>34.1</td>
</tr>
<tr>
<td>48</td>
<td>9031</td>
<td>95 (1.1)</td>
<td>696 (7.7)</td>
<td>239 (2.6)</td>
<td>8001 (88.6)</td>
<td>791 (8.8)</td>
<td>334 (3.7)</td>
<td>12.0</td>
<td>28.4</td>
</tr>
<tr>
<td>54</td>
<td>8240</td>
<td>126 (1.5)</td>
<td>672 (8.2)</td>
<td>222 (2.7)</td>
<td>7220 (87.6)</td>
<td>798 (9.7)</td>
<td>348 (4.2)</td>
<td>15.8</td>
<td>36.2</td>
</tr>
<tr>
<td>60</td>
<td>7442</td>
<td>129 (1.7)</td>
<td>605 (8.1)</td>
<td>221 (3.0)</td>
<td>6487 (87.2)</td>
<td>734 (9.9)</td>
<td>350 (4.7)</td>
<td>17.6</td>
<td>36.9</td>
</tr>
</tbody>
</table>

* Tests performed where the patient was diseased
Table 54: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with increased between-individual variability by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1779 (8.9)</td>
<td>6916 (34.6)</td>
<td>679 (3.4)</td>
<td>10626 (53.1)</td>
<td>8695 (43.5)</td>
<td>2458 (12.3)</td>
<td>20.5</td>
<td>72.4</td>
</tr>
<tr>
<td>6</td>
<td>11305</td>
<td>171 (1.5)</td>
<td>1299 (11.5)</td>
<td>683 (6.0)</td>
<td>9152 (81.0)</td>
<td>1470 (13.0)</td>
<td>854 (7.6)</td>
<td>11.6</td>
<td>20.0</td>
</tr>
<tr>
<td>12</td>
<td>9835</td>
<td>109 (1.1)</td>
<td>808 (8.2)</td>
<td>741 (7.5)</td>
<td>8177 (83.1)</td>
<td>917 (9.3)</td>
<td>850 (8.6)</td>
<td>11.9</td>
<td>12.8</td>
</tr>
<tr>
<td>18</td>
<td>8918</td>
<td>119 (1.3)</td>
<td>631 (7.1)</td>
<td>746 (8.4)</td>
<td>7422 (83.2)</td>
<td>750 (8.4)</td>
<td>865 (9.7)</td>
<td>15.9</td>
<td>13.8</td>
</tr>
<tr>
<td>24</td>
<td>8168</td>
<td>99 (1.2)</td>
<td>544 (6.7)</td>
<td>757 (9.3)</td>
<td>6768 (82.9)</td>
<td>643 (7.9)</td>
<td>856 (10.5)</td>
<td>15.4</td>
<td>11.6</td>
</tr>
<tr>
<td>30</td>
<td>7525</td>
<td>111 (1.5)</td>
<td>446 (5.9)</td>
<td>765 (10.2)</td>
<td>6203 (82.4)</td>
<td>557 (7.4)</td>
<td>876 (11.6)</td>
<td>19.9</td>
<td>12.7</td>
</tr>
<tr>
<td>36</td>
<td>6968</td>
<td>124 (1.8)</td>
<td>390 (5.6)</td>
<td>752 (10.8)</td>
<td>5702 (81.8)</td>
<td>514 (7.4)</td>
<td>876 (12.6)</td>
<td>24.1</td>
<td>14.2</td>
</tr>
<tr>
<td>42</td>
<td>6454</td>
<td>129 (2.0)</td>
<td>412 (6.4)</td>
<td>747 (11.6)</td>
<td>5166 (80.0)</td>
<td>541 (8.4)</td>
<td>876 (13.6)</td>
<td>23.8</td>
<td>14.7</td>
</tr>
<tr>
<td>48</td>
<td>5913</td>
<td>143 (2.4)</td>
<td>346 (5.9)</td>
<td>718 (12.1)</td>
<td>4706 (79.6)</td>
<td>489 (8.3)</td>
<td>861 (14.6)</td>
<td>29.2</td>
<td>16.6</td>
</tr>
<tr>
<td>54</td>
<td>5424</td>
<td>122 (2.2)</td>
<td>305 (5.6)</td>
<td>727 (13.4)</td>
<td>4270 (78.7)</td>
<td>427 (7.9)</td>
<td>849 (15.7)</td>
<td>28.6</td>
<td>14.4</td>
</tr>
<tr>
<td>60</td>
<td>4997</td>
<td>130 (2.6)</td>
<td>286 (5.7)</td>
<td>717 (14.3)</td>
<td>3864 (77.3)</td>
<td>416 (8.3)</td>
<td>847 (17.0)</td>
<td>31.2</td>
<td>15.3</td>
</tr>
<tr>
<td>All</td>
<td>95507</td>
<td>3036 (3.2)</td>
<td>12383 (13.0)</td>
<td>8032 (8.4)</td>
<td>72056 (75.4)</td>
<td>15419 (16.1)</td>
<td>11068 (11.6)</td>
<td>19.7</td>
<td>27.4</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased

Table 55: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with increased between-individual variability by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1406 (7.0)</td>
<td>3876 (19.4)</td>
<td>1052 (5.3)</td>
<td>13666 (68.3)</td>
<td>5282 (26.4)</td>
<td>2458 (12.3)</td>
<td>26.6</td>
<td>57.2</td>
</tr>
<tr>
<td>6</td>
<td>14718</td>
<td>190 (1.3)</td>
<td>1002 (6.8)</td>
<td>1109 (7.5)</td>
<td>12417 (84.4)</td>
<td>1192 (8.1)</td>
<td>1299 (8.8)</td>
<td>15.9</td>
<td>14.6</td>
</tr>
<tr>
<td>12</td>
<td>13526</td>
<td>141 (1.0)</td>
<td>667 (4.9)</td>
<td>1215 (9.0)</td>
<td>11503 (85.0)</td>
<td>808 (6.0)</td>
<td>1356 (10.0)</td>
<td>17.5</td>
<td>10.4</td>
</tr>
<tr>
<td>18</td>
<td>12718</td>
<td>125 (1.0)</td>
<td>523 (4.1)</td>
<td>1272 (10.0)</td>
<td>10798 (84.9)</td>
<td>648 (5.1)</td>
<td>1397 (11.0)</td>
<td>19.3</td>
<td>8.9</td>
</tr>
<tr>
<td>24</td>
<td>12070</td>
<td>117 (1.0)</td>
<td>461 (3.8)</td>
<td>1349 (11.2)</td>
<td>10143 (84.0)</td>
<td>578 (4.8)</td>
<td>1466 (12.1)</td>
<td>20.2</td>
<td>8.0</td>
</tr>
<tr>
<td>30</td>
<td>11492</td>
<td>132 (1.1)</td>
<td>423 (3.7)</td>
<td>1418 (12.3)</td>
<td>9519 (82.8)</td>
<td>555 (4.8)</td>
<td>1550 (13.5)</td>
<td>23.8</td>
<td>8.5</td>
</tr>
<tr>
<td>36</td>
<td>10937</td>
<td>159 (1.5)</td>
<td>441 (4.0)</td>
<td>1437 (13.1)</td>
<td>8900 (81.4)</td>
<td>600 (5.5)</td>
<td>1596 (14.6)</td>
<td>26.5</td>
<td>10.0</td>
</tr>
<tr>
<td>42</td>
<td>10337</td>
<td>130 (1.3)</td>
<td>392 (3.8)</td>
<td>1509 (14.6)</td>
<td>8306 (80.4)</td>
<td>522 (5.0)</td>
<td>1639 (15.9)</td>
<td>24.9</td>
<td>7.9</td>
</tr>
<tr>
<td>48</td>
<td>9815</td>
<td>161 (1.6)</td>
<td>363 (3.7)</td>
<td>1587 (16.2)</td>
<td>7704 (78.5)</td>
<td>524 (5.3)</td>
<td>1748 (17.8)</td>
<td>30.7</td>
<td>9.2</td>
</tr>
<tr>
<td>54</td>
<td>9291</td>
<td>196 (2.1)</td>
<td>342 (3.7)</td>
<td>1627 (17.5)</td>
<td>7126 (76.7)</td>
<td>538 (5.8)</td>
<td>1823 (19.6)</td>
<td>36.4</td>
<td>10.8</td>
</tr>
<tr>
<td>60</td>
<td>8753</td>
<td>183 (2.1)</td>
<td>317 (3.6)</td>
<td>1651 (18.9)</td>
<td>6602 (75.4)</td>
<td>500 (5.7)</td>
<td>1834 (21.0)</td>
<td>36.6</td>
<td>10.0</td>
</tr>
<tr>
<td>All</td>
<td>133657</td>
<td>2940 (2.2)</td>
<td>8807 (6.6)</td>
<td>15226 (11.4)</td>
<td>106684 (79.8)</td>
<td>11747 (8.8)</td>
<td>18166 (13.6)</td>
<td>25.0</td>
<td>16.2</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased
## Table 56: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with decreased fibrosis progression rate by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(^a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1290 (6.4)</td>
<td>3094 (15.5)</td>
<td>643 (3.2)</td>
<td>14973 (74.9)</td>
<td>4384 (21.9)</td>
<td>1933 (9.7)</td>
<td>29.4</td>
<td>66.7</td>
</tr>
<tr>
<td>6</td>
<td>15616</td>
<td>174 (1.1)</td>
<td>1204 (7.7)</td>
<td>606 (3.9)</td>
<td>13632 (87.3)</td>
<td>1378 (8.8)</td>
<td>780 (5.0)</td>
<td>12.6</td>
<td>23.3</td>
</tr>
<tr>
<td>12</td>
<td>14238</td>
<td>128 (0.9)</td>
<td>833 (5.9)</td>
<td>601 (4.2)</td>
<td>12676 (89.0)</td>
<td>961 (6.7)</td>
<td>729 (5.1)</td>
<td>13.3</td>
<td>17.6</td>
</tr>
<tr>
<td>18</td>
<td>13277</td>
<td>116 (0.9)</td>
<td>693 (5.2)</td>
<td>602 (4.5)</td>
<td>11866 (89.4)</td>
<td>809 (6.1)</td>
<td>718 (5.4)</td>
<td>14.3</td>
<td>16.2</td>
</tr>
<tr>
<td>24</td>
<td>12468</td>
<td>99 (0.8)</td>
<td>608 (4.9)</td>
<td>597 (4.8)</td>
<td>11164 (89.5)</td>
<td>707 (5.7)</td>
<td>696 (5.6)</td>
<td>14.0</td>
<td>14.2</td>
</tr>
<tr>
<td>30</td>
<td>11761</td>
<td>108 (0.9)</td>
<td>532 (4.5)</td>
<td>594 (5.1)</td>
<td>10527 (89.5)</td>
<td>640 (5.4)</td>
<td>702 (6.0)</td>
<td>16.9</td>
<td>15.4</td>
</tr>
<tr>
<td>36</td>
<td>11121</td>
<td>98 (0.9)</td>
<td>540 (4.9)</td>
<td>596 (5.4)</td>
<td>9887 (88.9)</td>
<td>638 (5.7)</td>
<td>694 (6.2)</td>
<td>15.4</td>
<td>14.1</td>
</tr>
<tr>
<td>42</td>
<td>10483</td>
<td>110 (1.0)</td>
<td>463 (4.4)</td>
<td>593 (5.7)</td>
<td>9317 (89.9)</td>
<td>573 (5.5)</td>
<td>703 (6.7)</td>
<td>19.2</td>
<td>15.6</td>
</tr>
<tr>
<td>48</td>
<td>9910</td>
<td>106 (1.1)</td>
<td>460 (4.6)</td>
<td>572 (5.8)</td>
<td>8772 (88.5)</td>
<td>566 (5.7)</td>
<td>678 (6.8)</td>
<td>18.7</td>
<td>15.6</td>
</tr>
<tr>
<td>54</td>
<td>9344</td>
<td>90 (1.0)</td>
<td>415 (4.4)</td>
<td>583 (6.2)</td>
<td>8256 (88.4)</td>
<td>505 (5.4)</td>
<td>673 (7.2)</td>
<td>17.8</td>
<td>13.4</td>
</tr>
<tr>
<td>60</td>
<td>8839</td>
<td>90 (1.0)</td>
<td>401 (4.5)</td>
<td>587 (6.6)</td>
<td>7761 (87.8)</td>
<td>491 (5.6)</td>
<td>677 (7.7)</td>
<td>18.3</td>
<td>13.3</td>
</tr>
<tr>
<td>All</td>
<td>137057</td>
<td>2409 (1.8)</td>
<td>9243 (6.7)</td>
<td>6574 (4.8)</td>
<td>118831 (86.7)</td>
<td>11652 (8.5)</td>
<td>8983 (6.6)</td>
<td>20.7</td>
<td>26.8</td>
</tr>
</tbody>
</table>

\(^a\)Tests performed where the patient was diseased

## Table 57: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with decreased fibrosis progression rate by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased(^a) n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>1162 (5.8)</td>
<td>2236 (11.2)</td>
<td>771 (3.9)</td>
<td>15831 (79.2)</td>
<td>3398 (17.0)</td>
<td>1933 (9.7)</td>
<td>34.2</td>
<td>60.1</td>
</tr>
<tr>
<td>6</td>
<td>16602</td>
<td>207 (1.2)</td>
<td>920 (5.5)</td>
<td>733 (4.4)</td>
<td>14742 (88.8)</td>
<td>1127 (6.8)</td>
<td>940 (5.7)</td>
<td>18.4</td>
<td>22.0</td>
</tr>
<tr>
<td>12</td>
<td>14575</td>
<td>147 (0.9)</td>
<td>661 (4.3)</td>
<td>740 (4.8)</td>
<td>13927 (90.0)</td>
<td>808 (5.2)</td>
<td>887 (5.7)</td>
<td>18.2</td>
<td>16.6</td>
</tr>
<tr>
<td>18</td>
<td>14667</td>
<td>102 (0.7)</td>
<td>558 (3.8)</td>
<td>783 (5.3)</td>
<td>13524 (90.2)</td>
<td>660 (4.5)</td>
<td>885 (6.0)</td>
<td>15.5</td>
<td>11.5</td>
</tr>
<tr>
<td>24</td>
<td>14007</td>
<td>113 (0.8)</td>
<td>495 (3.5)</td>
<td>786 (5.6)</td>
<td>12613 (90.0)</td>
<td>608 (4.3)</td>
<td>899 (6.4)</td>
<td>18.6</td>
<td>11.3</td>
</tr>
<tr>
<td>30</td>
<td>13399</td>
<td>104 (0.8)</td>
<td>446 (3.3)</td>
<td>813 (6.1)</td>
<td>12036 (89.8)</td>
<td>550 (4.1)</td>
<td>917 (6.8)</td>
<td>18.9</td>
<td>11.3</td>
</tr>
<tr>
<td>36</td>
<td>12849</td>
<td>127 (1.0)</td>
<td>458 (3.6)</td>
<td>805 (6.3)</td>
<td>11459 (89.2)</td>
<td>585 (4.6)</td>
<td>932 (7.3)</td>
<td>21.7</td>
<td>13.6</td>
</tr>
<tr>
<td>42</td>
<td>12264</td>
<td>121 (1.0)</td>
<td>429 (3.5)</td>
<td>828 (6.8)</td>
<td>10886 (88.8)</td>
<td>550 (4.5)</td>
<td>949 (7.7)</td>
<td>22.0</td>
<td>12.8</td>
</tr>
<tr>
<td>48</td>
<td>11714</td>
<td>129 (1.1)</td>
<td>449 (3.8)</td>
<td>817 (7.0)</td>
<td>10319 (88.1)</td>
<td>578 (4.9)</td>
<td>946 (8.1)</td>
<td>22.3</td>
<td>13.6</td>
</tr>
<tr>
<td>54</td>
<td>11136</td>
<td>135 (1.2)</td>
<td>390 (3.5)</td>
<td>806 (7.2)</td>
<td>9805 (88.0)</td>
<td>525 (4.7)</td>
<td>941 (8.5)</td>
<td>25.7</td>
<td>14.3</td>
</tr>
<tr>
<td>60</td>
<td>10611</td>
<td>125 (1.2)</td>
<td>369 (3.5)</td>
<td>814 (7.7)</td>
<td>9303 (87.7)</td>
<td>494 (4.7)</td>
<td>939 (8.8)</td>
<td>25.3</td>
<td>13.3</td>
</tr>
<tr>
<td>All</td>
<td>152724</td>
<td>2472 (1.6)</td>
<td>7411 (4.9)</td>
<td>8696 (5.7)</td>
<td>134145 (87.8)</td>
<td>9883 (6.5)</td>
<td>11168 (7.3)</td>
<td>25.0</td>
<td>22.1</td>
</tr>
</tbody>
</table>

\(^a\)Tests performed where the patient was diseased
### Table 58: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with increased fibrosis progression rate by observation point

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>2473 (12.4)</td>
<td>3320 (16.6)</td>
<td>1123 (5.6)</td>
<td>13084 (65.4)</td>
<td>5793 (29.0)</td>
<td>3596 (18.0)</td>
<td>42.7</td>
<td>68.8</td>
</tr>
<tr>
<td>6</td>
<td>14207</td>
<td>447 (3.1)</td>
<td>1394 (9.8)</td>
<td>1097 (7.7)</td>
<td>11269 (79.3)</td>
<td>1841 (13.0)</td>
<td>1544 (10.9)</td>
<td>24.3</td>
<td>29.0</td>
</tr>
<tr>
<td>12</td>
<td>12366</td>
<td>344 (2.8)</td>
<td>1096 (8.9)</td>
<td>9829 (79.5)</td>
<td>1441 (11.7)</td>
<td>1440 (11.6)</td>
<td>23.9</td>
<td>23.9</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>10925</td>
<td>311 (2.8)</td>
<td>933 (8.5)</td>
<td>8593 (78.7)</td>
<td>1244 (11.4)</td>
<td>1399 (12.8)</td>
<td>25.0</td>
<td>22.2</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>9681</td>
<td>335 (3.5)</td>
<td>859 (8.9)</td>
<td>7452 (77.0)</td>
<td>1194 (12.3)</td>
<td>1370 (14.2)</td>
<td>28.1</td>
<td>24.5</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>8487</td>
<td>347 (4.1)</td>
<td>766 (9.0)</td>
<td>6384 (75.2)</td>
<td>1113 (13.1)</td>
<td>1337 (15.8)</td>
<td>31.2</td>
<td>26.0</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>7374</td>
<td>350 (4.7)</td>
<td>746 (10.1)</td>
<td>1019 (13.8)</td>
<td>1096 (14.9)</td>
<td>1369 (18.6)</td>
<td>31.9</td>
<td>25.6</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>6278</td>
<td>386 (6.1)</td>
<td>579 (9.2)</td>
<td>1036 (16.5)</td>
<td>4277 (68.1)</td>
<td>965 (15.4)</td>
<td>40.0</td>
<td>27.1</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>5313</td>
<td>395 (7.4)</td>
<td>509 (9.6)</td>
<td>1016 (19.1)</td>
<td>3393 (63.9)</td>
<td>904 (17.0)</td>
<td>43.7</td>
<td>28.0</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>4409</td>
<td>421 (9.5)</td>
<td>389 (8.8)</td>
<td>949 (21.5)</td>
<td>2650 (60.1)</td>
<td>810 (18.4)</td>
<td>52.0</td>
<td>30.7</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>3599</td>
<td>391 (10.9)</td>
<td>277 (7.7)</td>
<td>843 (23.4)</td>
<td>2088 (58.0)</td>
<td>668 (18.6)</td>
<td>58.5</td>
<td>31.7</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>102639</td>
<td>6200 (6.0)</td>
<td>10869 (10.6)</td>
<td>11292 (11.0)</td>
<td>74278 (72.4)</td>
<td>17069 (16.6)</td>
<td>17492 (17.0)</td>
<td>36.3</td>
<td>35.4</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased

### Table 59: Adjusted fibrosis progression sensitivity analyses—results using reference strategy with increased fibrosis progression rate by observation point and PPV at 25%

<table>
<thead>
<tr>
<th>Observation time (months)</th>
<th>Tests (n)</th>
<th>TP results n (%)</th>
<th>FP results n (%)</th>
<th>FN results n (%)</th>
<th>TN results n (%)</th>
<th>Positive results n (%)</th>
<th>Diseased* n (%)</th>
<th>PPV (%)</th>
<th>Sensitivity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20000</td>
<td>3037 (15.2)</td>
<td>6915 (34.6)</td>
<td>559 (2.8)</td>
<td>9489 (47.4)</td>
<td>9952 (49.8)</td>
<td>3596 (18.0)</td>
<td>30.5</td>
<td>84.5</td>
</tr>
<tr>
<td>6</td>
<td>10048</td>
<td>328 (3.3)</td>
<td>2124 (21.1)</td>
<td>454 (4.5)</td>
<td>7142 (71.1)</td>
<td>2452 (24.4)</td>
<td>782 (7.8)</td>
<td>13.4</td>
<td>41.9</td>
</tr>
<tr>
<td>12</td>
<td>7596</td>
<td>216 (2.8)</td>
<td>1330 (17.5)</td>
<td>403 (5.3)</td>
<td>5647 (74.3)</td>
<td>1546 (20.4)</td>
<td>619 (8.1)</td>
<td>14.0</td>
<td>34.9</td>
</tr>
<tr>
<td>18</td>
<td>6050</td>
<td>207 (3.4)</td>
<td>990 (16.4)</td>
<td>339 (5.6)</td>
<td>4514 (74.6)</td>
<td>1197 (19.8)</td>
<td>546 (9.0)</td>
<td>17.3</td>
<td>37.9</td>
</tr>
<tr>
<td>24</td>
<td>4853</td>
<td>153 (3.2)</td>
<td>789 (16.3)</td>
<td>295 (6.1)</td>
<td>3616 (74.5)</td>
<td>942 (19.4)</td>
<td>448 (9.2)</td>
<td>16.2</td>
<td>34.2</td>
</tr>
<tr>
<td>30</td>
<td>3911</td>
<td>139 (3.6)</td>
<td>635 (16.2)</td>
<td>277 (7.1)</td>
<td>2860 (73.1)</td>
<td>774 (19.8)</td>
<td>416 (10.6)</td>
<td>18.0</td>
<td>33.4</td>
</tr>
<tr>
<td>36</td>
<td>3137</td>
<td>135 (4.3)</td>
<td>493 (15.7)</td>
<td>284 (9.1)</td>
<td>2225 (70.9)</td>
<td>628 (20.0)</td>
<td>419 (13.4)</td>
<td>21.5</td>
<td>32.2</td>
</tr>
<tr>
<td>42</td>
<td>2509</td>
<td>158 (6.3)</td>
<td>358 (14.3)</td>
<td>262 (10.4)</td>
<td>1731 (69.0)</td>
<td>516 (20.6)</td>
<td>420 (16.7)</td>
<td>30.6</td>
<td>37.6</td>
</tr>
<tr>
<td>48</td>
<td>1993</td>
<td>151 (7.6)</td>
<td>290 (14.6)</td>
<td>247 (12.4)</td>
<td>1305 (65.5)</td>
<td>441 (22.1)</td>
<td>398 (20.0)</td>
<td>34.2</td>
<td>37.9</td>
</tr>
<tr>
<td>54</td>
<td>1552</td>
<td>126 (8.1)</td>
<td>219 (14.1)</td>
<td>234 (15.1)</td>
<td>973 (62.7)</td>
<td>345 (22.2)</td>
<td>360 (23.2)</td>
<td>36.5</td>
<td>35.0</td>
</tr>
<tr>
<td>60</td>
<td>1207</td>
<td>111 (9.2)</td>
<td>155 (12.8)</td>
<td>213 (17.6)</td>
<td>728 (60.3)</td>
<td>266 (22.0)</td>
<td>324 (26.8)</td>
<td>41.7</td>
<td>34.3</td>
</tr>
<tr>
<td>All</td>
<td>62856</td>
<td>4761 (7.6)</td>
<td>14298 (22.7)</td>
<td>3567 (5.7)</td>
<td>40230 (64.0)</td>
<td>19059 (30.3)</td>
<td>8328 (13.2)</td>
<td>25.0</td>
<td>57.2</td>
</tr>
</tbody>
</table>

*Tests performed where the patient was diseased
Figure 16: Results of strategies using data with adjusted fibrosis progression estimate
Comparison to ELUCIDATE data

The ELUCIDATE data contained 705 observations taken from 420 participants randomised to the ELF monitoring arm of the trial. After removing measurements following an ELF value of 9.5 or above for each individual (akin to the trial setting), the simulated data set contained 66,320 observations for 20,000 participants and the simulated data set with adjusted fibrosis progression included 59,000 observations for 20,000 participants. Analysis of the ELF value at the point of randomisation for each of the data sets showed similar results—mean (SD) for the ELUCIDATE data was 9.57 (1.21); for the simulated data 9.71 (1.15); and for the simulated data with adjusted fibrosis progression 9.83 (1.20)—with the mean value being slightly lower for the ELUCIDATE data than the two simulated data sets. The between-individual standard deviation was higher for the ELUCIDATE data than for the simulated data sets (0.93 for the ELUCIDATE data compared with 0.76 for the simulated data and 0.82 for the simulated data with adjusted fibrosis progression). The within-individual standard deviation was similar for the ELUCIDATE data and both simulated data sets (0.53 for the ELUCIDATE data and, 0.51 and 0.52 for the simulated and simulated with adjusted fibrosis progression data sets respectively). Results of analysis of randomisation ELF and analysis of variance on ELF at all recorded time points can be seen in Table 60.

The ELUCIDATE data modelled consisted of 429 observations from 153 participants, with each participant having a minimum of 2 and a maximum of 6 ELF observations and the average number of observations per person was 2.8. The number of observation points used from the simulation model was therefore capped to give a similar mean number of observations per person to the value seen in the ELUCIDATE data. Allowing more observations per person would introduce bias as patients with slower progressing disease will have more ELF measurements prior to having a test result of 9.5 or above. The bias seen here relates to comments made by Bellera et al when analysing monitoring data. The model fitted to simulated data used 26,429 observation points for 9,608 simulated participants and the simulated data with adjusted fibrosis progression rate used 23,972 observations for 8,779 simulated participants. For the simulated data sets the mean number of observations was 2.8 for the original data set with unadjusted fibrosis progression and 2.7 for the data set with adjusted fibrosis progression. Results of modelling the ELUCIDATE data, simulated data and adjusted fibrosis progression estimate simulated data can be seen in Tables 61-63.

Modelling of the ELUCIDATE data showed the increased in ELF per year to be 0.31 (95% CI (0.22, 0.39); p-value <0.001). Modelling of the simulated data showed the increase in ELF
per year to be comparable at 0.24 (95% CI (0.23, 0.26); p-value <0.001) and for the simulated data with adjusted fibrosis progression the increase in ELF per year was estimated to be 0.28 (95% CI (0.27, 0.30); p-value <0.001).

Table 60 Results of analysis of randomisation ELF and analysis of variance for ELF measurements at all time points

<table>
<thead>
<tr>
<th>Randomisation point ELF</th>
<th>ELUCIDATE data</th>
<th>Simulated data</th>
<th>Simulated data with adjusted fibrosis progression</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELF mean (SD)</td>
<td>9.57 (1.21)</td>
<td>9.71 (1.15)</td>
<td>9.83 (1.20)</td>
</tr>
<tr>
<td>Analysis of variance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between-individual SD</td>
<td>0.93</td>
<td>0.76</td>
<td>0.82</td>
</tr>
<tr>
<td>Within-individual SD</td>
<td>0.53</td>
<td>0.51</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Table 61 Results of multilevel model of repeated ELF measures from ELUCIDATE trial

<table>
<thead>
<tr>
<th>Estimate</th>
<th>95% Confidence Interval</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years</td>
<td>0.31</td>
<td>(0.22, 0.39)</td>
</tr>
<tr>
<td>Constant</td>
<td>8.73</td>
<td>(8.63, 8.82)</td>
</tr>
<tr>
<td>Between-individual SD</td>
<td>0.43</td>
<td>(0.36, 0.51)</td>
</tr>
<tr>
<td>Within-individual SD</td>
<td>0.48</td>
<td>(0.44, 0.52)</td>
</tr>
</tbody>
</table>

Table 62 Results of multilevel model of simulated repeated ELF measurements

<table>
<thead>
<tr>
<th>Estimate</th>
<th>95% Confidence Interval</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years</td>
<td>0.24</td>
<td>(0.23, 0.26)</td>
</tr>
<tr>
<td>Constant</td>
<td>8.84</td>
<td>(8.83, 8.85)</td>
</tr>
<tr>
<td>Between-individual SD</td>
<td>0.42</td>
<td>(0.41, 0.43)</td>
</tr>
<tr>
<td>Within-individual SD</td>
<td>0.47</td>
<td>(0.46, 0.47)</td>
</tr>
</tbody>
</table>

Table 63 Results of multilevel model of simulated repeated ELF measurements with adjusted fibrosis progression rate

<table>
<thead>
<tr>
<th>Estimate</th>
<th>95% Confidence Interval</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years</td>
<td>0.28</td>
<td>(0.27, 0.30)</td>
</tr>
<tr>
<td>Constant</td>
<td>8.86</td>
<td>(8.84, 8.87)</td>
</tr>
<tr>
<td>Between-individual SD</td>
<td>0.42</td>
<td>(0.41, 0.43)</td>
</tr>
<tr>
<td>Within-individual SD</td>
<td>0.46</td>
<td>(0.46, 0.47)</td>
</tr>
</tbody>
</table>
Discussion

Reference strategy

At the initial testing point a monitoring strategy will be identifying cases from a prevalent population where a large proportion of patients will have high ELF values. At subsequent time points those with a positive result will not be tested and the tested population will contain cirrhotic patients that have either been falsely negative at the previous testing point or have developed cirrhosis since the last testing point (incident cases), hence the difference in results at the initial monitoring time point compared with others. The percentage of false negative results generally increased with each time point as patients with low ELF trajectory have reached compensated cirrhosis but as they have a low ELF value for their disease stage they are required to progress further to have a positive test result using the simple threshold decision rule. The increasing percentage of false negative results as the testing points advance suggests the simple threshold should be reduced at later time points to account for the patients that have false negative results using the original threshold.

Comparing strategies with changes to individual components to the reference strategy

Inferior strategies

It was anticipated that the strategy with retesting would result in an increase in the number of tests per person required compared to the reference strategy but the percentage of patients with delay to diagnosis increased also. For the patients in the target retest range their test result was determined by combining both the initial and retest results and their corresponding measurement errors with the mean of the initial and retest result being used to classify a result as positive or negative. Due to the measurement error of both the initial and retest results some patients would have been positive on their initial test (as with the reference strategy) but using the mean of the initial and retest means they have a negative result. The slight increase in time to diagnosis when using a retest strategy will have a small effect on the percentage of participants with delay to diagnosis also.

The strategies using absolute and relative increase from last recorded value decision rules were notably worse than the strategies using absolute and relative changes from the initial recorded value. When using a decision rule based on detecting a magnitude of change between one value and another, the two values used to calculate the change will both have measurement error. The strategies with decision rules based on absolute and relative increases from the start value had
only the recorded value at the time of randomisation with measurement error incorporated, a constant value for each individual, to compare to when determining a positive or negative result. The strategies with decision rules based on absolute and relative increases from the last recorded value have a new value to compare with at each testing point, meaning a different measurement error to account for. Comparisons with the initial value will allow the detection of gradual increase in ELF across the entire monitoring period rather than sharp increases since the previous monitoring point.

The simple threshold strategy outperformed the strategies comparing current to previous values. This is in part due to the index of individuality (II), the ratio of within-individual and between-individual variation. If a test has a high II value, where an individual can have results spanning a wide range of the possible results for a group of people, comparison to constant thresholds will be more meaningful than for tests with low II values, where an individual will have tests results spanning only part of the possible range of results and comparisons to previous results will be more beneficial.236

‘Trade-off’ strategies
The reduced test frequency strategy showed a large decrease in the number of tests per person used for a small increase in the percentage of people with delay to diagnosis. It may be that for a substantial decrease in the number of tests required, and therefore the resource used, the slight potential for increased harm to patients (through later diagnosis) is acceptable.

Superior strategies
The linear regression strategy was the only strategy tested that showed a reduction in both the number of tests required and the percentage of patients with delay to diagnosis. By fitting a regression model using all previous observations for an individual and obtaining a prediction from this, the linear regression method utilised all available data and some allowance was made for the fluctuation in results due to measurement error. The linear regression strategy, however, only resulted in small benefits compared with the reference strategy. This modest improvement in monitoring strategy performance may not merit the extra complexity involved when using the linear regression method.

Estimates of test performance and disease progression (sensitivity analyses)

The results obtained when varying estimates in the simulation model and evaluating the reference strategy highlight the importance of accurate data. The increases and decreases in
estimates of test performance (measurement error and between-individual variability) and fibrosis progression rate affected the three measures of performance in different ways.

**Measurement error and between-individual variability**

The measurement error of a test affects the number of false positive test results, with larger measurement error resulting in more false positive results and smaller measurement error resulting in fewer false positive results. Between-individual variability will affect the underlying ELF values possible at each fibrosis stage. Providing ELF is related to fibrosis stage, if the between-individual variability is smaller it will be easier to correctly identify fibrosis stage from ELF resulting in less false positive results and more true positive results.

With less false positives and more true positives PPV will increase, the number of tests required will increase as the reduction in false positives means the number of patients correctly staying in the monitoring programme will increase. With reduced measurement error the observed values reflect more closely the underlying disease state of each patient, if the threshold does not adequately account for this patients will need to progress for longer to have a test value over the threshold indicating a positive result. When the between-individual variability is reduced, due to the increase in true positive results the percentage of patients with delay to diagnosis will decrease.

**Fibrosis progression rate**

Fibrosis progression rate will affect the number of diseased patients. With an increased fibrosis progression rate more patients will have compensated cirrhosis, which will lead to an increase in PPV. With increased fibrosis progression rate patients have positive results earlier in the strategy and the strategy will require fewer tests to be performed. If patients have increased fibrosis progression rate more patients will have been in cirrhosis for more than 12 months meaning more patients can be undetected for over 12 months.

**Limitations**

**Data sources**

The estimates from data sources used to inform the simulation model will have a large impact on the results of the simulation model. The suitability of data was assessed, by consultation with clinical colleagues, and where necessary estimates were adjusted for sensitivity analyses. However, as the model was dependent on the information it used, the quality and suitability of data used will always be a limitation. Just one cross-sectional study provided information on both the link between ELF scores and fibrosis stage and the distribution of fibrosis stages at
entry to the trial. When looking to identify an estimate of measurement error several sources were identified with the estimates from each found to be vastly different. The data linking ELF to fibrosis stage defined fibrosis stage by biopsy. Even though biopsy is the reference standard for staging fibrosis, biopsy is known to not be accurate in some cases.

The ELUCIDATE trial data used to assess the simulation model was not completely appropriate as the dataset contained repeated observations from only 153 participants with many participants having only two observations; more observations per person would allow the model to better estimate the error terms and the changes over time. ELF measurements only being taken until the point of a measurement being classed as positive also hinders the ability of the data to estimate the true progression of ELF over time as those with higher ELF values (and possibly more developed cirrhosis) cease to have ELF recorded and so progression beyond this ELF value cannot be assessed. Patients with lower ELF measures (below 9.5) continued monitoring meaning they had more measures, and therefore contributed more data to the model; however they were potentially very different to those with fewer measures, who were likely in a worse health state.

**Assumptions**
A major limitation of the simulation model is the vast number of assumptions required. Some of the estimates used to generate the monitoring data, such as fibrosis progression rate and measurement error, can be varied in magnitude and the results assessed to identify the impact of using data of insufficient quality or suitability in the model. However, there were many assumptions, made out of necessity, in the development of the model that cannot be varied and assessed so easily and the consequences explored. The model assumes fibrosis progression is constant and requires patients to have only positive fibrosis progression; patients with a decreasing or stable disease state had progression at 0.01 fibrosis units per year for the purposes of the simulation which would indicate a stable disease state. The model assumed linear increases in ELF between fibrosis stages, normally distributed ELF within fibrosis stage and constant fibrosis progression rate. The error associated with each observation was assumed to be normally distributed and a simple error term has been used with no distinction made between within-individual and analytical variation. The error used in the simulation may also be simplistic as the error term is randomly chosen from a distribution that is not only constant across individuals and time but also not linked to the magnitude of the ELF value. As no alternative data or substantiated opinion was available to enable modelling of these factors in any other way, these assumptions were necessary for the development of the model.
Longitudinal data sets with ELF values and biopsy recorded in addition to data from a biological variability study of ELF would be required to test these assumptions.

**Trial considerations**

Several criteria were required to allow the simulation model to generate data for a trial (described in Table 15). Whilst these criteria were included to avoid anomalies and were based on clinical advice, there is no data to support them.

**Further work**

A greater variety of strategies could be evaluated with multiple components assessed simultaneously. More complex decision rules and frequencies could be explored, for example a simple threshold decision rule where the threshold remains the same across patients but varies by time point within a monitoring strategy or changing the frequency of testing to be non-constant.

It may be possible for the simulation model to be adapted to account for usual care (and variation in usual care). If usual care could be allowed for, it may be possible to compare monitoring strategies with usual care and with further simulation work estimate differences in patient outcomes.

The model can be used to show lifetime progression for a time-matched cohort of patients with fibrosis (if the data is simulated with all patients starting at the onset of liver fibrosis). This data may be beneficial to the assessment of how a strategy would perform in practice rather that specifically in the trial setting as this would provide information on how newly diagnosed patients would benefit from monitoring.

**Conclusions**

Simulation can be used to obtain monitoring data for candidate monitoring strategies and to enable an appropriate strategy to be selected for full scale evaluation.

In the case of using ELF to monitor liver fibrosis, only the linear regression monitoring strategy showed better performance than the simple threshold strategy, and given the additional complexity and small benefit from using the linear regression method the simple threshold method may be most appropriate. Reducing the frequency of testing may be an alternative to the simple threshold strategy if the compromise between number of tests and delay to diagnosis is acceptable.
To generate monitoring data there has to be available evidence on the natural history of the disease and the performance of the monitoring test (measurement error and test accuracy)—this evidence can be from existing data sets, reviewing the literature or potentially expert opinion. If the data informing the simulation model is inaccurate the results obtained from evaluation of strategies will not reflect the truth. Inaccurate estimates will affect results in a complex way. The results of sensitivity analyses highlighted the importance, for this test and disease area, to have accurate estimates of test performance and progression.

When analysing data from the ELUCIDATE trial to compare with the simulated data it was clear that due to the design of the ELUCIDATE trial the analysis would have to limit the bias of recorded ELF results. Comparison of the ELUCIDATE data and the simulated data provided similar results. Bias in monitoring data particularly concerning the number of recorded results should be considered when analysing.
Chapter 8 - Methodological considerations in the optimization of monitoring biomarkers to meet value-based market access hurdles

Components of the work described in this chapter were published:

Background

Biomarkers are a central component of the proposed revolution in healthcare. Sometimes called personal, precision or stratified medicine – the defining characteristic of personalized medicine is the use of molecular (including genetic) and imaging information on the individual patient; i.e. biomarkers, to guide decisions on their clinical management.

The potential clinical applications of personalized medicine include:
- screening for the risk of developing disease,
- diagnosing the presence of disease,
- providing prognosis an individual patients disease progression
- identifying whether patients are likely to respond to particular treatments (pharmacogenomics); and
- identifying whether patients are at elevated risk of adverse events from particular treatments (pharmacogenomics/toxgenomics).

The biomedical knowledge that underlies personalized medicine also has significant potential in the discovery and translation of new therapies, however, consideration of these is outside of the scope of this project.

Personalised medicine technologies are being developed at a time when developed health care systems are under increasing pressure to increase the efficiency; i.e. to consider the value of new technologies in terms of what they produce in relation to how much they cost; and to review current clinical practices with a view to eliminate those activities that are of low or no value. The former process is implemented through health technology assessment processes, such as those undertaken by NICE in the United Kingdom, HAS in France, the PBAC in Australia and CADTH in Canada. In line with other new health technologies, personalized medicine innovations tend to come at a substantial financial cost. Commercially provided tests such a Oncotype Dx are magnitudes more expensive than conventional laboratory tests; and co-dependent therapies such as crizotinib even more expensive. These price
tags mean they are inevitably subject to formal health technology assessment prior to market access.

The latter process is being driven internationally by the Choosing Wisely Campaign. The elimination of unnecessary tests is at the forefront of Choosing Wisely campaigns. As a result there is an inevitable tension between the widespread adoption of additional tests that are required for personalized medicine and the societal and professional pressure to make less use of tests. In this context, biomarker tests will only be adopted if they can be demonstrated to be high value use of limited health care resources. This relatively new pressure to assess the value of new tests before they are adopted aligns testing with the processes of health technology assessment that have been used for drugs and some devices for many years. The focus of this theme of the programme grant has been the development of methods for assessing the value of personalized medicine technologies, with a particular focus on monitoring tests; i.e. tests that are applied repeatedly to the same patient over a period of time to inform the sequential clinical care decisions.

The remainder of this chapter is structured as follows. In section 2 we consider the difference between a conventional clinical utility for the individual patient approach to test optimization and an approach based upon the cost effectiveness of the test from a population health perspective. Section 3 describes a framework for characterizing personalized medicine technologies; Section 4 describes an existing method for optimizing diagnostic tests to meet cost effectiveness targets and its extension for monitoring tests. Section 5 describes the formal extension of this framework for a monitoring test with ‘n’ administrations and provides an illustrative example where there are 6 sequential administrations of the test. Section 6 then sets out some conceptual issues with the calculation of the value of information of additional research for a monitoring test.
Cost effectiveness in personalized medicine technologies

As described above, biomarker tests are at the centre of the personalized medicine revolution. Whilst some tests are dichotomous – i.e. they test the presence or absence of a particular biomedical characteristic, many tests measure a continuous biomedical parameter, e.g. blood glucose, Forced Expiratory Volume. The interpretation of the test result converts the continuous variable into a categorical variable by defining a certain test result as the transition point between normal and abnormal. This transition point is referred to as the test cut-point. Conventionally the choice of cut-point is selected on the basis of clinical utility. Clinical utility considers the risk benefit ratio of a test from the perspective of the individual patient. Conditional upon the treatment associated with a positive test result having no or only a low risk of an adverse event, maximizing clinical utility leads to preferring a highly sensitive diagnostic test cut-point over a highly specific cut-point.

Figure 17 shows how shifting the cut point for a diagnostic test changes the proportion of patients who receive false positive and false negative test results. The upper half of the figures show the test score distribution for individuals who actually have the condition of interest. The lower half of the diagram shows the test score distribution for individuals who do not have it. Between the points A and B, for any given test result it is possible that the individual may have or be free of the condition of interest. The initial cut-point is shown by the solid blue vertical line; and the dark red portions of each distribution show the proportion of the test results that are incorrect. Individuals who have the condition but receive a test score below the cut-point are defined as false negatives; and individuals free of the condition who receive a test score above the cut-point are defined as false positives.
Figure 17 Relationship between cut-point and test performance

Figure 18 shows how the proportion of individuals who receive false positive and false negative test results change as the cut-point is moved. As the test cut-point moves to the left (lower), the proportion of false negative results reduces; but the proportion of false positive results increases. By contrast, if the cut-point moves to the right, the proportion of individuals who receive a false negative result increases, whilst the number of individuals who receive a false positive result decreases.
As described above, because clinical utility is evaluated from the individual patient perspective, the value attached to a false positive tends to be considerably less than the value attached to a false negative. A false negative deprives the individual of the opportunity to receive appropriate treatment in a timely manner. By contrast, a false
positive exposes the individual to a treatment that they will not benefit from but is not expected to do them harm. However, when consideration of the value of test is expanded to include the resources consumed by the test and any treatments administered subsequent to the test result, as is the case for the health technology appraisal of tests, the perspective for establishing value moves from the individual patient to that of the population that the health care system is responsible for.

From a population health perspective, in a health system that is operates with budget that is fixed, or quasi-fixed in real terms, the cost of a technology is health foregone by others. We use the ‘bookshelf’ model developed by Culyer, McCabe and Edlin to illustrate how the cost effectiveness threshold represents the health foregone due to the premium cost of new therapies.

The cost-effectiveness ‘bookshelf’ is a graphical representation of the health system in which each available health technology is represented as a unique ‘book’ on the bookshelf. A broad definition of ‘technology’ is adopted for this purpose, which includes any health care intervention or service that consumes resources and provides value to the health system.

The width of each book represents each technology's budget impact if funded (i.e., the incremental cost of providing the technology to all patients in the relevant indication), while the height of each book represents each technology’s incremental cost-effectiveness ratio (ICER). For the purpose of this paper, the preferred unit of ‘effectiveness’ for a health technology is assumed to be the quality-adjusted life year (QALY), such that the height of each book represents the incremental cost per additional QALY provided by the technology. The books are stacked next to each other along the bookshelf (the x-axis of Figure 19) and sorted so that the most desirable technologies (represented by the shortest books) are at the far left of the bookshelf and the least desirable technologies (the tallest books) are at the far right (Figure 19). With a fixed health budget not all technologies can be funded. In choosing which technologies to fund, the decision maker maximizes the value produced by the health system by funding the technology at the far left of the bookshelf first, this is technology A in in Figure 19. The decision maker carries on funding each technology to its right in turn (B, C, D, etc.) until the budget is spent.
The least desirable technology to be funded (G) is referred to as the ‘marginal’ technology.

Since the health system has a fixed budget, funding any new technology will inevitably displace one or more existing, funded technologies. For the purposes of this explanation, we assume that the health system’s objective is to improve population health with the available budget. Hence, a new technology will only be funded if it provides more QALYs than are forgone through the displacement of currently funded technologies. For a new technology to do this it must produce more QALYs per $ spent than the least valuable technology that is currently paid for. In our bookcase example, a new technology must have a cost per QALY [incremental cost effectiveness ratio (ICER)], that is equal to or better than technology G, the marginal technology, if it is to produces more QALYs than is displaces. Hence, the health system budget provides an implicit value of health – the ICER of the marginal technology. This is the value that technologies, including personalized medicine, must target if they are to be attractive to decision makers.

Figure 19 The cost effectiveness bookshelf
Now consider the relationship between the test cut-point and cost effectiveness. In Figure 17 we showed the four categories of test result: True Positive and Negative; False Positive and Negative. Table 64 provides illustrative costs and QALYs associated with each of the categories. The rows CP1 and CP2 provide the distribution of tested individuals between the four possible outcomes for two different cut-points. In the fourth column we have calculated the expected Costs and QALYs for the test using CP1 and CP2. We can see the expected cost of the test using CP2 is $5,000 compared to $4,400 for CP1. However, the expected QALYs from the test using CP2 is also higher – 0.74 compared to 0.64. From a population health perspective, is it worth moving from CP1 to CP2? We know that for this to be the better value choice, the incremental cost per QALY produced by this substitution must be lower than the cost per QALY of the marginal technology. In our example the cost per QALY of the marginal technology is $50,000 (top right hand cell in Table 64); and the Incremental Cost per QALY of CP2 compared to CP1 is calculated as follows:

\[
ICER = \frac{5000 - 4400}{0.74 - 0.64} = \$6,000 \text{ per QALY}
\]

As $6,000 is considerably lower than $50,000, we would expect the implementation of the test using CP2 rather than CP1 to create more health than is displaced by the additional cost.
When using the cost effectiveness decision criterion, the increase in population is maximized by including new technologies up to the point where the ICER for new technology is exactly equal to the ICER of the marginal technology. Hence, in the cost effectiveness framework it is possible to go further than merely identifying whether a specific test cut-point is good value. The framework can be used to identify the most cost effective cut-point for a test. We describe how this is done in more detail in Section 4.

A framework for characterizing Personalised Medicine Technologies

The umbrella of Personalised Medicine covers a wide range of technologies and combinations of technologies. Identifying the appropriate methods for the economic evaluation of a specific personalized medicine technology will be significantly helped by a systematic approach to characterizing the components of the technology. In this section we describe a model for characterizing personalized medicine technologies in terms of their constituent technologies, for the purposed of economic evaluation.

As described above the foundation characteristic of personalized medicine is a test of the molecular, including genetic, characteristics of an individual or, or in the case of cancers, a disease. For screening and diagnostic tests, this may be the sole component of the technology. However, for prognostic tests and for test-treatment combinations, the technology will require combination of the molecular information and clinical expression (phenotypic) data. Prognostic technologies inherently link the molecular

---

**Table 64 Impact of changing cut-points on cost and outcomes**

<table>
<thead>
<tr>
<th>QALYs</th>
<th>True Positive</th>
<th>True Negative</th>
<th>False Positive</th>
<th>False Negative</th>
<th>Expected Value</th>
<th>Net Monetary Benefit</th>
<th>Value of Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Costs</td>
<td>$5,000</td>
<td>$500</td>
<td>$8,000</td>
<td>$8,000</td>
<td></td>
<td></td>
<td>$50,000</td>
</tr>
<tr>
<td>Distribution CP1</td>
<td>Expected QALYs</td>
<td>0.45</td>
<td>0.3</td>
<td>0.05</td>
<td>0.2</td>
<td>$27,600</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expected Costs</td>
<td>$2,250</td>
<td>$150</td>
<td>$400</td>
<td>$1,600</td>
<td>$4,400</td>
<td></td>
</tr>
<tr>
<td>Distribution CP2</td>
<td>Expected QALYs</td>
<td>0.5</td>
<td>0.2</td>
<td>0.1</td>
<td></td>
<td>$32,000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expected Costs</td>
<td>$2,500</td>
<td>$100</td>
<td>$1,600</td>
<td>$800</td>
<td>$5,000</td>
<td></td>
</tr>
</tbody>
</table>
information to phenotypic information, and thus are a combination of two testing technologies. The magnitude of an individual patient’s health benefit from treatment is dependent upon the phenotypic expression of the disease. The relationship between the molecular characterization of disease and phenotypic expression is uncertain in even the monogenetic disorders such as Gauchers, where homozygous twins have been shown to have radically different times to symptomatic presentation. In addition to the molecular and phenotypic test components, some PM technologies may have an additional pharmacogenomic test technology, to establish whether the individual will respond to the specific therapy; HER2 testing for Herceptin therapy in Breast Cancer being possibly the most well known example. The final potential component of a personalized medicine technology is the treatment itself. Some technologies – such as Kalydeco the gene-specific treatment for cystic fibrosis – are themselves personalized technologies, whilst others are more conventional treatments. However, their economic evaluation will require consideration of at least the phenotypic expression test as well as possibly the molecular data.

Figure 20 provides a graphical representation of this framework.

The framework allows us to highlight the potential for correlations between the different components of a specific PM technology. For example, the effectiveness of treatment is likely to be systematically related to phenotypic expression. This relationship may be positive or negative depending upon the nature of the treatment. For example a treatment that stops further progression but does not resolve accumulated disability will have a less valuable effect the greater the phenotypic expression; by contrast a treatment that resolves accumulated disability will have a more valuable effect in patients with greater disability.

Changes in the phenotypic expression of the population targeted for treatment will change the casemix of the patients to whom a pharmacogenomics test is administered, with potential implications for the test performance characteristics of the pharmacogenomics test, and by extension the expected effectiveness of the therapy in ‘responders’ identified by that test. For cost effectiveness analyses understanding relationships that impact upon the expected magnitude of benefit are clearly central.
In the next section we describe how to identify the optimum cut-point for a diagnostic test and consider how this framework can be extended for a monitoring test.

**Extending Phelps and Mushlin for monitoring tests**

Phelps and Mushlin (PM, 1988) use cost effectiveness analysis to assess diagnostic technologies. In their work, they describe how to identify the optimal test cut-off for a diagnostic test that meets a pre-specified cost effectiveness threshold. They consider a population where each person has some probability of illness (f), and model the simple case where a physician uses a dichotomous test to identify patients as being either sick or healthy depending on the test diagnoses. It is however recognized that the population on which these diagnostic technologies are applied is heterogeneous, and the probability for each individual being sick may vary.

The benefits and harms often associated with the use of diagnostic technologies depend on the true state of health of the patients on whom these technologies are applied. However, it is unlikely for an individual’s true health state to be observed, thus the actual patients benefits and costs will depend on how well diagnostic technologies identify these states of health. Ideally, tests will be optimized to identified individuals, but since resources in every health system are finite, the goal is
that the use of diagnostic devices will optimize population health, and this requires adding up the benefits and costs over the population eligible to use the test.

In developing the theoretical framework for assessing diagnostic devices, Phelps and Mushlin use the following notation;

\[ f = \text{probability that patient is sick} \]
\[ p = \text{sensitivity} \]
\[ 1 - q = \text{specificity} \]
\[ U_{st}(C_{st}) = \text{utility(cost) of sick person, treated} \]
\[ U_{sn}(C_{sn}) = \text{utility(cost) of sick person, not treated} \]
\[ U_{ht}(C_{ht}) = \text{utility(cost) of healthy person, treated} \]
\[ U_{hn}(C_{hn}) = \text{utility(cost) of healthy person, not treated} \]

Using the above information and a predetermined cost effectiveness threshold (\( g \)) they determine the net benefit for the diagnostic technology and maximize this net benefit in an optimization setting with respect to the test performance characteristics, specifically \( p \) and \( q \). PM define an expression for the net benefit as;

\[
\begin{align*}
NB &= f \left[ U_{st} + (1 - p) \times U_{sn} \right] + (1 - f) \left[ (1 - q) \times U_{ht} + q \times U_{hn} \right] \\
&\quad - g \left\{ f \left[ p \times C_{st} + (1 - p) \times C_{sn} \right] + (1 - f) \left[ (1 - q) \times C_{ht} + q \times C_{hn} \right] \right\} 
\end{align*}
\]

The combination of \( p \) and \( q \) which maximizes the NB can be obtained by varying the diagnostic test cut-off used to establish diagnosis. For a given cost-effectiveness threshold, the optimal choice of \( p \) and \( q \) which maximizes the expected net benefit is given by;

\[
\frac{dp}{dq} = \frac{(1 - f) \times (\Delta U_H - g \times \Delta C_H)}{f \times (\Delta U_S - g \times \Delta C_S)} = \beta_d
\]

The slope \( \beta_d \) corresponds to a cut-off \( k_d \) on the ROC curve of our diagnostic test technology of interest. Using the costs and outcomes of treatments of false positives

---

2 We can find combinations of \( p \) and \( q \) in which the net benefit remains the same, by taking the differential of Eq. (1), and allowing \( p \) and \( q \) to vary jointly, but holding the total change in NB to zero.
and negatives obtained from literature, in addition to the ROC curve for our diagnostic test, a cost effective test cut-off can be identified for a given threshold.

We have worked in this NIHR Applied Programme to extend the work of PM.\textsuperscript{290} The framework developed by PM for a diagnostic test has been used to evaluate the same test, CA125, as a monitoring test. Specifically, they consider a two period monitoring test where clinicians re-administer the same test to patients who are diagnosed as either positive or negative by the first test, [for the reason that these patients display a clear predisposition for the condition of interest.] Using the Ca125 test monitoring for relapse in ovarian cancer, we show how the repeated use of the initial cut-off can lead to a substantially increased false-negative rate compared with the monitoring cut-off - over 4% higher than in this example - with the associated harms for individual and population health.

In this monitoring scenario presented by Longo et al (2014), each of the different subgroups of patients that are retested have a systematically different prevalence (probability of being sick) in the period in which they are tested compared to the initial population.\textsuperscript{290} The prevalence in the subgroups in the second period depends on both the ability of the previous test to correctly classify patients as sick or healthy, and on the probability of developing the disease in the time between the two tests. We argue that if the prevalence within the subgroups of population that are being tested in the current period are unique and different, then likewise, the optimal combination of test performance characteristics that maximize the net benefit among these subgroups will also be unique and differ from $p$ and $q$ of the initial test. As a result, a unique cost effective test cut-off can be identified for each sub population on an ROC that corresponds to their test performance characteristics. We illustrate this by first updating the prevalence in Eq 1 to obtain the unique slopes in each sub-population, and they then identify the unique cut-offs for each sub-population as shown below:

\[
\frac{(1-f_{\text{+ve subgroup}})(\Delta U_1 - g \times \Delta C_1)}{f_{\text{+ve subgroup}}(\Delta U_s - g \times \Delta C_s)} = \beta_{\text{positive sub-population}}
\]
It is possible to build on the work, to develop a framework which identifies the prevalence and the corresponding optimal cost effective test cut-offs for subgroups in monitoring period for a set of n administrations of the test. For simplicity, and ease of computational burden, we use a discrete set of cut-off points between pre-specified test score limits of 0 and 1. We obtain the optimal pathway for our n-period monitoring, using a backward induction approach in a dynamic programming setting. We consider a monitoring regime where individuals with or at risk of disease of interest, presumably a chronic disease, are followed for i periods. Those who are diagnosed positive receive treatment while those diagnosed negative do not receive treatment, and we continue to monitor both the positive and negative sub-populations to the ith period. In each period, a physician decides on a test cut-off ($T_{c_{i,j}}$; where i refers to the monitoring period and j refers to a specific sub-population within a period\(^3\)) that will be used to test each sub-population. Let assume $T_{c_{i,j}}$ is measured on a continuous scale and lies anywhere between 0 and 1, then individuals with test scores above $T_{c_{i,j}}$ are diagnosed positive and vice versa.

\[ \frac{dp}{dq} = \frac{(1-f_{\text{ve subgroup}}) \times (\Delta U_H - g \times \Delta C_H)}{f_{\text{ve subgroup}} \times (\Delta U_S - g \times \Delta C_S)} = \beta_{\text{negative sub-population}} \]

\(^3\) Thus in period n=1, j=(1); in period i=2, j=(1,2); in period i=3, j=(1,2,3,4)
We consider a 3 period monitoring regime. In this monitoring regime, a physician decides on $T_{c_{1,1}}$ in period 1, and the initial population is tested with this cut-off. Individuals whose test score fall above $T_{c_{1,1}}$ are considered as our positive subpopulation, and those with test scores below are considered as the negative subpopulation. In period $i=2$, the physician decides on two test cut-offs ($T_{c_{2,1}}$ and $T_{c_{2,2}}$) that will be used to test the subgroups. The positive subgroup in period $i=2$ is tested with a cut-off $T_{c_{2,1}}$ while the negative subgroup is tested with a cut-off $T_{c_{2,2}}$. Based on the test diagnosis of each individual in each of the two sub-populations, they are further stratified into 4 subgroups in period $i=3$ as shown in Figure 21, and the physician has to decide on the test cut-off for each of these 4 subgroups. Thus for every $i$th period, a physician is faced with $2i-1$ sub-populations, and he decides the test cut-off that must be applied to each of these them. In Figure 21 above, every positive sub-population comprises of all the true positives (TP) and false positives (FP), while the negative sub-population comprises of the true negative (TN) and false
negatives (FN). The utilities and costs associated with individuals in each of these test outcomes is different.

The prevalence of a condition varies under sequential testing, and this compels a difference in test cut-offs. In period \( i=1 \), the prevalence in the initial population can in most cases be obtained from clinical experts, the registry or from epidemiological literature. Longo et al show that patients presenting for monitoring tests have a different prevalence compared with the general population presenting for the initial test. They estimate the prevalence for the positive and negative subgroups in period \( i=2 \) to be \( r1 + (1-r1) \rho \) and \((1-s1) + s1 \rho\) respectively, where \( r1 \) and \( s1 \) represent the positive and negative predictive values respectively, and \( \rho \) is the probability that a patient is sick when presenting for a test in period \( i=2 \), given that they were not sick when presenting for test in period \( i=1 \). Thus the prevalence of a condition in the current period is a function of the test performance characteristics in the previous period and the rate of disease progression.

Developing on the framework we described in Longo et al (2014), consider the case where individuals are monitored for more than two periods. We define the following notations; 

\[
i = \text{the monitoring period} \\
j = \text{set of sub-populations in period } i \\
\rho = \text{rate of disease progression} \\
f_{i,j} = \text{prevalence for } j\text{th sub-population in the } i\text{th period}
\]

Thus from Figure 21 above, we can see that at period \( i=2 \), \( j = (1, 2) \) and \( f_{i,j} = (f_{2,1}, f_{2,2}) \), and at period \( i=3 \), \( j = (1, 2,3,4) \) and \( f_{i,j} = (f_{3,1}, f_{3,2}, f_{3,3}, f_{3,4}) \).

We also define \( k \) such that:

---

\(^4\) The utilities (costs) associated with a TP, FP, TN and FN are \( U_{TP}(C_{TP}) \), \( U_{FP}(C_{FP}) \), \( U_{TN}(C_{TN}) \) and \( U_{FN}(C_{FN}) \)
Given the above notation, the prevalence for a sub-population $j$ in period $i$ can be generalized as:

$$f_{i,j} = \frac{p_{i-1,k} \times f_{i-1,k} + \left[ (1 - q_{i-1,k}) \times (1 - f_{i-1,k}) \right] \times \rho}{p_{i-1,k} \times f_{i-1,k} + (1 - q_{i-1,k}) \times (1 - f_{i-1,k})} \quad \forall \ j \text{ that is odd}$$

and

$$f_{i,j} = \frac{(1 - p_{i-1,k}) \times f_{i-1,k} + \left[ (1 - q_{i-1,k}) \times (1 - f_{i-1,k}) \right] \times \rho}{(1 - p_{i-1,k}) \times f_{i-1,k} + (1 - q_{i-1,k}) \times (1 - f_{i-1,k})} \quad \forall \ j \text{ that is even}$$

If $i = 1$ then we have the case of a diagnostic. For $i \geq 2$, we have a monitoring scenario. It must also be noted that $f_{1,1}$ is exogenous, and can be obtained from epidemiological literature.

For a two period monitoring test, Longo et. al (2014) show how to calculate the disease prevalence for each decision node of the second period using the positive predictive value, the negative predictive value, and the disease progression rate from the first period. In our model we generalize on Longo et. al (2014) and derive the formulas for calculating the respective prevalence at each decision node when the monitoring test is administered repeatedly in $N$ periods. Then, we maximize the expected net-benefits resulting from the repeated administration of the test by finding the optimal cut-off test scores in each period.

Our illustrative model based uses the dynamic programming method to calculate the optimal cut-off test scores for a monitoring test with 6 periods. In this model the sensitivity and specificity are calculated at each decision node, which can be achieved by estimating the distributions of the test scores for the sick and healthy patients at the respective node. Next, we show an optimization example for a monitoring test administered in 6 periods ($N=6$). Let the model parameters be denoted as follows:
\[ f : \text{Probability that patient is sick} \]
\[ U_{st}(C_{st}) : \text{Utility(cost) of sick person, treated} \]
\[ U_{sn}(C_{sn}) : \text{Utility(cost) of sick person, not treated} \]
\[ U_{ht}(C_{ht}) : \text{Utility(cost) of healthy person, treated} \]
\[ U_{hn}(C_{hn}) : \text{Utility(cost) of healthy person, not treated} \]
\[ g : \text{Inverse of cost effectiveness ratio} \]
\[ \rho : \text{Disease progression rate} \]

The parameters take the values reported in Table 65.

### Table 65 Parameter values for illustrative monitoring model

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>Positive</th>
<th>False</th>
<th>False Positive</th>
<th>True</th>
</tr>
</thead>
<tbody>
<tr>
<td>QALYs</td>
<td>0.9</td>
<td>-2</td>
<td>0.5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Costs ($s)</td>
<td>500</td>
<td>800</td>
<td>400</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(g)</td>
<td>(1/50,000=0.00002)</td>
<td>(f)</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\rho)</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Furthermore, let \(T_s\), \(0 \leq T_s \leq 1\), denote the test scores of sick patients and \(T_h\), \(0 \leq T_h \leq 1\), denote the test scores of healthy patients. We use beta distributions for \(T_s\) and \(T_h\) because the beta distribution is restricted to values in \([0, 1]\) and can be used to model distributions that are symmetric or skewed. Let \(T_s \sim \beta(2,5)\) and \(T_h \sim \beta(5,2)\) for all decision nodes in all periods. However, the model is developed such that the distributions for \(T_s\) and \(T_h\) can be defined distinctly for each decision node.

To solve the optimization problem and find the optimal cut-offs for each decision node in each period, we have discretized \(T_s\) and \(T_h\) with the granularity 0.1. Thus, \(T_s\) and \(T_h\) \(\in\{0,0.1,...,1\}\). Using the backward induction method of dynamic programming results in the optimal cut-offs as depicted in the following table: (Table 66)
<table>
<thead>
<tr>
<th></th>
<th>Decision for 1st Period</th>
<th>1st Period Outcomes</th>
<th>2nd Period Outcomes</th>
<th>3rd Period Outcomes</th>
<th>4th Period Outcomes</th>
<th>5th Period Outcomes</th>
<th>6th Period Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3</td>
<td>+</td>
<td>0.5</td>
<td>+</td>
<td>0.6</td>
<td>+</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>-</td>
<td>0.5</td>
<td>+</td>
<td>0.6</td>
<td>-</td>
<td>0.6</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>-</td>
<td>0.5</td>
<td>-</td>
<td>0.6</td>
<td>+</td>
<td>0.6</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>-</td>
<td>0.5</td>
<td>-</td>
<td>0.6</td>
<td>+</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>-</td>
<td>0.5</td>
<td>-</td>
<td>0.6</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>-</td>
<td>0.5</td>
<td>-</td>
<td>0.6</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>-</td>
<td>0.4</td>
<td>-</td>
<td>0.5</td>
<td>+</td>
<td>0.5</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>-</td>
<td>0.4</td>
<td>-</td>
<td>0.5</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>-</td>
<td>0.4</td>
<td>-</td>
<td>0.5</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>-</td>
<td>0.4</td>
<td>-</td>
<td>0.5</td>
<td>-</td>
<td>0.4</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>-</td>
<td>0.3</td>
<td>+</td>
<td>0.5</td>
<td>+</td>
<td>0.6</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>-</td>
<td>0.3</td>
<td>+</td>
<td>0.5</td>
<td>+</td>
<td>0.6</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>-</td>
<td>0.3</td>
<td>+</td>
<td>0.5</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>-</td>
<td>0.3</td>
<td>+</td>
<td>0.5</td>
<td>-</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 66 Optimising cut-offs
Some observations on the estimation of the value of information for monitoring tests

The methods we have described above develop Phelps and Mishlin’s work to take account of how the change in the casemix of patients changes after each test administration. However, it has assumed that the ROC curve is the same for each administration. However, it is known that changing the casemix of the population to which a test is applied will change the sensitivity and specificity and thus, the ROC curve for the test at administration of the test will be significantly different from the previous test and for the individuals who tested positive compared to those who tested negative in the previous administration. Figures 22 and 23 illustrate how the data driving the sensitivity and the specificity for the second administration a test would differ from the data driving the sensitivity and specificity for the first administration. [Strictly these figures are correct under the assumption that there is no disease progression.]

Perfect information for a diagnostic test can be characterized as being able to characterize the distribution of test scores for each true health state that the test is designed to measure. With this information for each test score it would be possible to define the probability that an individual who received that score was a true/false
positive or true/false negative. Perfect information for a monitoring test requires the same information plus knowing with certainty how many individuals will have progressed from a true negative state to a true positive state in the time interval between the tests. This would allow the calculation of sensitivity and specificity, assuming that the test score distribution for the new true positive individuals is the same as for the true positive individuals at the time of the first administration of the test; and hence allowing the construction of separate ROC curves for the two groups defined by the initial test.

It follows that there are three distinct sources of uncertainty in a monitoring test:

- test score distribution for true positive individuals;
- test score distribution for true negative individuals; and
- progression rate from true negative to true positive.

The choice of the value of health used in the analysis plays a more complex role in the estimation of the value of information for a monitoring technology than for an
effectiveness parameter. This is because it the choice of the value of health determines the cut-point for the initial and subsequent administrations of the test. In determining the initial cut-point it determines the ROC curves for the subsequent test administrations. This points to a much more complex relationship between the value of health and the value of information than is observed for therapies.

Uncertainty on the true health state is unlikely to be evenly distribution across the range of test scores. Extreme test scores (either high or low) are systematically less likely to be false positives/negatives than test scores in the mid-range. The implication of this is that the location of the initial cut-point will have a direct effect on the contribution of the test score distributions for true positive and true negatives to the total value of information about the test performance. Further, the choice of initial cut-point will impact upon the degree to which disease progression can lead to a change of status after the monitoring test, and hence the value additional research on this parameter.

The pivotal role of the choice the value of health in determining the initial cut-point and hence the value of further research, value of information analyses will be health system specific to an even greater degree than is the case for research on the effectiveness and safety of therapeutic interventions. For monitoring test technologies that are developed for multiple markets it may be more efficient to undertake substantial global pre-market studies that characterize all 3 parameters (test score distribution for true positive and true negative cases, and the rate of progression) in some detail as this may be more efficient than multiple studies that target portions of the evidence base for each parameter.

Conclusions
The work described in this chapter and published in Longo et al, 2014, provides a new insight into the use of biomarkers for monitoring tests when the principal evidence of benefit is the cost effectiveness of the healthcare services which will be managed in the context of the biomarker test results. Rigorous study designs and careful attention to the choice of cut off are essential.
1) By applying Phelps and Mushlin’s framework to monitoring tests, we have shown that the test cut-off that is cost effective in monitoring may be quite different from the one that is cost effective in the first administration of the test.

2) We show that the extent of this difference, and hence its importance to clinicians, patients, and decision makers depends upon the underlying prevalence of the condition, test performance characteristics of the test at its first administration, rate of disease progression, and the interval between test administrations, as well as the target cost-effectiveness threshold.

3) Decision makers are likely to be interested in the impact of changing the cost-effectiveness threshold on the optimal case definition cut-off value; hence, we present the cost-effectiveness threshold curve, which is able to identify the optimal case definition cut-off value corresponding to each cost-effectiveness threshold.

4) The framework developed by Phelps and Mushlin, and extensions, such as the ones we have described here, offer a formal decision analytic approach identifying the cut-off that optimizes the contribution of these technologies to population health in a given healthcare system.
Chapter 9 - An overview and patient perspective on biomarker guided research
Introduction

Does a marker-guided strategy lead to better outcomes for patients? Despite a wealth of literature on the development and validation of biomarkers, relatively few studies have directly addressed this question. The best evidence to answer this kind of question comes from a pragmatic RCT looking at comparative clinical and cost-effectiveness. Few such trials have been conducted and there have been some major disappointments. In oncology, one of the most active areas for biomarker research, McShane and colleagues concluded in 2005 that, “the number of markers that have emerged as clinically useful is pitifully small”, an uncomfortable conclusion which has been echoed several times over the subsequent decade.\textsuperscript{32, 293-296} It seems therefore that, despite a very active biomarker discovery process, researchers cannot yet claim that the “biomarker pipeline” is working successfully in patient benefit terms.

In this chapter we pull together the salient conclusions of the work in chapters 2-8, both to provide an overview from the perspective of researchers and healthcare professionals. However, in the light of this, we were keen to explore the patient perspective on our overview of the “state-of-the-art” in this field. The main findings of the Methodological workstream were distilled and presented to a group of patients and family members, and their opinions sought. This chapter also describes that exercise.

The technical language used in methodological writing tends to make the key arguments inaccessible to a non-technical readership, so we decided to proceed in two main stages. In Stage 1, the key points were identified in discussion between this chapter’s lead author (JH) and two health researchers (MT and JanB), chosen to be experienced in PPI but not familiar with the methodological literature on biomarkers or test evaluation. This initial step entailed extracting key messages from nearly a hundred pages of methodological writing. JH prepared a description of the main findings, which itself ran to several pages; MT and JanB read the original chapters as well as JH’s summary, then over three meetings key concepts were agreed and provisional plans drawn up for how the material might be presented to patient and public representatives. In Stage 2, the points were first discussed with one of the Programme’s highly experienced PPI representatives (JoanB) and amendments made,
before the agreed material was presented (by MT and JanB) to a group of patients and family members from the Liver North patient support group.

Further details of the Stage 2 work are provided towards the end of this chapter. The section to follow presents the results of the Stage 1 work. For completeness, it incorporates a few points that chapter authors added after the PPI consultation exercise was completed, but none of the additions was substantive enough to change the key messages.

The starting point of WS1 was that a marker is a type of test, so when we took a close look at the way marker-guided strategies have been evaluated, we could draw on what is known about evaluating tests. This is a well-trodden path: most of the basic principles of test evaluation have been known for a long time. However, as the chapters in this section have shown, these basic principles are not consistently applied to the evaluation of a biomarker-guided patient management strategy. The impact of such oversight can be substantial, a reality which seems not to be widely appreciated. It may be that researchers are unclear about the mechanisms at work, and few patients or members of the public are likely to have “lifted the lid” and questioned how the performance of a test may be judged. To ensure the planned PPI consultation was meaningful therefore, it was necessary to ensure first that the researchers had a good understanding of the WS1 outputs, in order for them to be able to explain in non-technical terms the background to the WS1 work, the results of the studies themselves, and also some of the implications. The Stage 1 document was the means to that end and is quite lengthy as a result. Methodologist readers may wish to skip this section (and rejoin the chapter on page x), as it will present them with no surprises and indeed they may find much of the material self-evident.

**Stage 1: Understanding outputs from the methodological workstream and agreeing key areas for discussion with patient and public representatives.**

This section begins by rehearsing a few terminological basics to ensure there is enough common ground for the discussion to follow.
The terminology being used is ACCE, which divides the research pipeline into labelled components: Analytical validity is followed by Clinical validity, then Clinical utility and finally, Ethical, legal and social considerations. Analytical validity, ‘A’, is about the quality of the measure used for the biomarker and how best to handle the inevitability of measurement error. Analytical validity is in itself a complex construct, with three main components: pre-analytical variability (the effect on samples of different storage conditions, transportation times etc.); analytical variability (differences between reagents, analysers, software, laboratory QA procedures etc.); and biological variability (within-patient differences, reflecting samples collected at different times of day, times since last meal, etc.). Assorted standardisation and callibration procedures can reduce variability from these sources, but never eliminate it, so estimating residual measurement error is an important part of the ‘A’ phase of test development.

Clinical validity, here ‘C1’, is about the relationship between the measure and the clinical condition, summarized using a variety of parameters. Measurement variability in the marker will necessarily set a limit to the strength of any such relationship and hence to the achievable performance of a marker-based test.

- The case-definition threshold (the cut-point on a continuous measure) is used to divide the tested population into test positives and test negatives. Choosing a cut-point always necessitates trading off detection rates and false alarms. The balancing act may be data-driven, e.g., using ROC analysis to identify the cut point which minimises misclassifications, or there may be a policy-driven focus on one parameter, e.g., the test must detect at least 60% of cases.

- The performance of the test at the chosen cut-point can be reported using the familiar sensitivity (how well the test performs in relation to cases) and specificity (how well the test performs in relation to non-cases). ‘Test accuracy’ as conventionally calculated uses the same principles, and essentially reports the proportion of the total sample (cases and non-cases) who were correctly classified.

- If the sensitivity and specificity figures observed in one study are replicated in one or more very similar studies, Clinical validity (C1) is considered to have been established, justifying the use of terms such as “a validated test”.
Clinical utility, the second ‘C’ in ACCE, refers to information on comparative clinical and cost effectiveness, information which is usually obtained via a pragmatic RCT. The way the ACC components of the pipeline work has major implications for the last component, ‘Ethical, legal and social considerations’, so consideration of ‘E’ is deferred until later in this chapter.

Other authorities have labelled up the stages of biomarker development in different ways, but the basic notion of “a pipeline” is widely accepted. Why then have so few clinically useful biomarkers emerged when researchers have followed it?

Some of the answers to this question are familiar from the wider literature on evaluating tests. Evaluation of marker-guided care often requires a test-and-treat strategy, with associated increase in sample size etc. At its most basic level, “benefit” from early identification and treatment depends not only on the effectiveness of treatment – the usual consideration when designing an RCT – but also on the number of people who could benefit if found, together with the ability of the test to find them. Standard treatment trials can take these as read and incorporated into eligibility criteria and recruitment projections. “Test and treat” trials on the other hand need to include them, or assumptions about them, in their planning, including in their power calculations. If Treatment As Usual includes existing tests or Clinical Decision Rules, then it follows that studies of comparative effectiveness will need to be larger still.

The working of the biomarker pipeline is returned to below, following a brief rehearsal of the complications entailed when a test is used for monitoring purposes. Most of the above insights have come from methodological work on the evaluation of tests used for diagnostic purposes, and guidelines for researchers in this area have been available for some time. Our present interest however lies elsewhere, namely in tests used for monitoring purposes, the evaluation of which has received, by comparison, little methodological attention.

**Monitoring tests**

Consistent with the wider test literature, the use of biomarkers for monitoring purposes has been very little studied, despite their widespread use in patient care. This
section therefore begins by revisiting some of the basic considerations entailed in a test-and-treat clinical utility evaluation, as described above for a diagnostic test.

It is important in designing any clinical utility study to know the number of people who could benefit if found. In the diagnostic context, the relevant figure is the prevalence of the condition in the sample at the time of testing (‘time 1’). In the monitoring context, it follows that information is also needed on changes in prevalence between test occasions.

And what about the ability of the test to find the people who could benefit from treatment? In a diagnostic context, if the prevalence of the condition and the sensitivity of the test are both known, then the number of cases detected can be calculated. Real cases, correctly detected are called ‘true positives’ (‘TPs’) in test parlance. But not all the people with positive test results will be true positives. Some will be ‘false positives’ (‘FPs’), and their numbers can be calculated from the proportion of non-cases in the sample, together with the specificity of the test. If no confirmatory testing is carried out, FPs will not be identified as such and will not be returned to the monitoring sample. In these circumstances, all those testing positive, i.e., FPs as well as TPs, will be managed in the same way, but it would be reasonable to assume that only the latter might be able to benefit. If the new management regime is associated with the potential for any kind of harm (further tests of an invasive nature, treatment side effects) then it must be remembered that FPs as well as TPs will be subject to these.

Any cases missed by the diagnostic test (the ‘false negatives’, or ‘FNs’) will remain in the monitoring sample. Over time, these will be supplemented by new “cases”, arising as a result of disease progression. To estimate the number of cases in the sample on a subsequent monitoring occasion (‘time 2’), it is necessary to know three things: the false negative rate at time 1, the progression rate of the disease, and elapsed time. Consequently, it cannot be assumed that prevalence at time 2 will be the same as that at time 1 – it could be higher, or it could be lower. The performance of the monitoring test will reflect these new circumstances, and the pattern will repeat itself over subsequent monitoring occasions.
The well-established route along the biomarker research pipeline does not readily generate all the information indicated above, or make use of it in the design of clinical utility studies, such as a trial to evaluate a monitoring strategy. Does this matter? If some figures aren’t available and plausible values have to be assumed, how much hangs on the assumed values being “about right”? And at what point in the pipeline should health economic considerations be addressed? Enhanced monitoring or the initiation of treatment for people unlikely to benefit will waste resources and can overburden health care systems, so simply widening the definition of who is eligible for enhanced care is seldom the best approach, even if the enhanced regime is acceptable and relatively benign.

WS1 aimed to address the above questions and help to bridge the gap between clinical validity and clinical utility studies. Our topic was the design of pragmatic trials aimed at evaluating the clinical utility of a marker used for monitoring purposes, and our focus was the role of key test and patient parameters in the design of such studies. We approached the task in a number of different ways, and from a number of different academic perspectives, but some very similar messages emerged.

A case study

Chapter 3 asked if key test and patient parameters had informed published guidelines on the use of prostate specific antigen for monitoring prostate cancer recurrence. Little evidence of such a systematic approach was found. When to test, for example, “appeared to be almost exclusively determined by standard follow-up schedules rather than being based on any scientific evidence.” Although the potential for variation in measurements was usually accepted by guideline developers, they showed little interest in the potential effect of such variation on the interpretation of test results. Relevant evidence was not always available, but even when it was available, it was not always used. A systematic review of biological variation in levels of prostate-specific antigen found mean variability of 20%. Based on this figure, it was calculated that, “to be 95% sure that a change in total prostate-specific antigen is not due to random variation, the change needs to be about 50% of the previous measurement.” This review was not however cited by any of the seven guidelines subsequently
published. Recommendations on when to take action were based on consensus statements or retrospective case series rather than on calculations of optimal values for cut offs and monitoring frequency in the target population.

**Trial evidence**

Chapter 4 asked if the randomised controlled trial design had been successfully used to evaluate strategies for monitoring disease progression or recurrence. Fifty eight relevant trials were found, with the intervention usually taking the form of adding an assessment to an existing schedule of care. The reviewers observed that, “although one might expect both the rate of disease progression and the degree of measurement variability associated with a given test to be taken into account when setting test frequency, test intervals were apparently determined by convenience or by fear of missing a key clinical event.” The test strategy which delivers a high detection rate (i.e., few missed key events) will inevitably also deliver false alarms, but in the published trials, “There was little acknowledgement of the potential for false-positive results…”. The reviewers also noted that many of the trials seemed to be underpowered. About a fifth of the 58 trials were stopped early, often because of a lower than expected event rate in the control groups. For these and no doubt other reasons, “Only a small proportion of trials reported statistically significant results for the primary outcome”:

**The methodology literature**

Chapter 5 found that the design and evaluation of strategies to monitor disease progression have received relatively little methodological attention. There is however relevant work on the statistical analysis of monitoring data and on the development and evaluation of screening strategies, with the starting point for most approaches being quite a detailed understanding of how test results fluctuate in the absence of disease progression or other “signal” of interest.

How big does a change need to be in marker scores to justify clinical interest? In Clinical Validity studies, if patient groups known to differ in clinically important
terms show a marker score difference of a particular magnitude, then consensus can emerge that that constitutes a "clinically important difference". In Analytical Validity studies, marker scores alone may reveal patterns of interest. Reference Change Values (RCVs) are a way of quantifying score fluctuations likely to occur in the absence of a real change in the patient’s underlying condition. Assuming the statistics of the normal curve apply, then a difference between two results greater than the RCV is taken to indicate a real change has occurred.

Some work has been done on using score variability to develop decision rules, but in the absence of a demonstrable link to patients’ clinical condition, the contribution of “signal to noise” approaches to choosing a monitoring strategy will remain limited. Overall, there is little evidence that methodological work has informed the functioning of the biomarker research pipeline.

**Patient outcomes and monitoring**

Chapter 6 explored two ways in which monitoring strategies may improve patient outcomes. If test results do not influence patient management, then a testing strategy is unlikely to improve patient outcomes. The purpose of introducing a new test needs first to be made explicit however: is it to replace an existing test on the grounds of improved accuracy, or reduced invasiveness, for example? Or is earlier diagnosis the aim? Or identification of an “at risk” group who are to receive further investigation? The causal pathways through which benefit is achieved are not necessarily simple in any of these circumstances however, so understanding how the relevant pathways operate and, if necessary, modifying the testing strategy in order to optimise them, may be necessary if maximum patient benefit is to be obtained. Three key approaches to structuring the problem in relation to tests for screening and for diagnosis were identified in the literature and applied to tests used for monitoring, drawing on 58 previously identified trials (see above) for illustrative purposes.

The focus of the first approach was the trade-off between benefits and harms under new testing regimes as compared to standard care, when tests are introduced for different purposes. This approach highlighted the extent to which the impact of a
monitoring strategy depends not only on the effectiveness of treatment, but also on the properties and timing of testing, and hence the extent to which the size and the management of different subgroups of patients differ in practice between the new and the standard regimes. Prior to conducting an RCT, these figures should be identified from pilot or feasibility studies.

In the second approach, a new testing strategy was considered to be a complex intervention, and its components picked apart in those terms. Clinicians’ and patients’ trust in the monitoring protocol, and their willingness to comply with it, were identified as important considerations which should not simply be assumed, but should be assessed prior to a trial, and in some circumstances, be the subject of research in their own right.

Lastly, the circumstances in which an RCT for the evaluation of patient benefit was both necessary and timely were scrutinized. Conclusions overlapped with those drawn in the two previous paragraphs: an RCT may indeed be required to evaluate the net impact of a new monitoring strategy, because the latter is likely to be exerting its effect via a network of individual mechanisms. However that RCT will be most informative if it is based on good estimates of important population, test and treatment parameters.

**The first modelling study**

The aim (Chapter 7) was to see if modelling could help identify a “best bet” monitoring strategy, for potential evaluation in a subsequent trial. The approach was to look at different rules for defining a test positive result and, on the assumption that true diagnostic status (“caseness”) would later be known, compare the different rules in terms of important indicators such as delay to diagnosis. The statistical model that enabled informative comparisons to be made was found in sensitivity analyses to be heavily reliant on detailed information about disease progression and test performance. In the chapter, the latter was characterised in terms of two components: measurement error and between-individual variability. The nature and role of measurement error and - in the biomarker field - its relationship to Analytical validity
('A'), have been described earlier. Terminology is unfortunately not standardised in this area, so here it also needs to be noted that “between-individual variability” is one way of characterising Clinical validity (‘C1’), as it refers to the variation in marker scores observed amongst individuals belonging – according to a reference standard - to the same diagnostic category (e.g., cirrhosis, or a particular fibrosis stage). Such variation within a diagnostic category is likely to be associated with overlap in the score ranges observed in different categories, which in turn brings the potential for misclassification of individuals when that is based on marker score – the conventional measure of ‘test accuracy’ characterises the performance of a test essentially in terms of the overall amount of such misclassification.

Putting the disease progression and test performance parameters together, it can be seen that we are on familiar territory, because together they determine the number of “cases” in the study sample at a point in time, and also the biomarker test’s ability to identify them.

Regarding the test performance parameters, available information on ELF, generated through the traditional research route, proved to be insufficient for comparison purposes. On disease progression, longitudinal descriptive data was needed, and even cross-sectional prevalence information would not have sufficed.

Using the simulated data to compare monitoring strategies for liver fibrosis, it was found that a simple threshold case rule performed very well. A more complex rule using linear regression to summarize changes in score over time did perform marginally better, but would have been potentially more difficult to apply in practice. It was emphasised that these comparisons did depend on the assumptions made. For example, tests which show only limited within-individual variation in their scores lend themselves to rules which incorporate changes over time, whereas comparison to a simple threshold is more meaningful for tests generating less stable scores.

The second modelling study
Because of its dependence on the ELUCIDATE trial, the details of this study are provided in the WS3 section of this report (Chapter 21). However, most of the results were available in time for the WS1 patient consultation exercise, so key points are reported here. In this simulation study, the measure of success of the monitoring strategy was not time to diagnosis, but time to severe complications. Reference standard diagnostic information is not available in many monitoring contexts, so cannot be used to check the clinical validity of the test, or be the determinant of patient management. This was the case in the ELUCIDATE trial, and further, any “diagnostic” information that became available would be defined differently in the two arms of the trial: clinically in the control arm, and using the ELF test in the intervention arm. Consequently, test accuracy considerations were not part of this simulation. In addition, the clinical utility of any test-and-treat strategy is dependent not only on successful diagnosis but also on successful treatment, so simulating the success of the monitoring strategy as evaluated in the trial required the modelling of patient outcomes. The recent HTA report on non-invasive liver biomarkers noted the value of this approach, on the grounds that it would provide a hard end point without the need for liver biopsy.¹²⁹

The key modelled relationship – as in the previously published ELF data - was between ELF scores and the rate of development of severe complications. Measurement error and the test performance of ELF would have contributed to that initial relationship, so were understood to take a similar role here and hence were not specifically addressed in the simulation – although the success of that approach does depend on these parameters remaining similar in the two settings. In planning the trial, a cut point of 12.5 on ELF was initially adopted as the marker-defined “diagnostic” threshold, but as a clearer picture emerged of the ELF score distribution in the trial population, a decision was taken to adopt the lower value of 9.5 to define an “at risk” subgroup, eligible for enhanced care.

In this study, trial participants’ time to severe complications was modelled, using early information from the trial itself. In the ELF arm, the initial distribution of ELF scores, together with the observed cumulative incidence of above-threshold scores), was used to model the relationship between a starting ELF value and the likelihood of passing the 9.5 threshold after a given period of time. In the control arm, the relationship between starting ELF values and the observed cumulative incidence of
clinically diagnosed cirrhosis was modelled, as well as the relationship between clinical diagnosis and concurrent ELF scores. Previously published data on the relationship between ELF ranges and time to severe complications were then used to model the occurrence of the latter in the two arms of the trial. An assumed treatment effect was then applied to all “diagnosed” patients and the two simulated arms compared over different durations of follow up. This simulated trial enabled the achieved power of the actual trial to be calculated, and provided an illustration of how the approach could be used for designing future trials.

Measurement error and test accuracy were both indirectly incorporated into this simulation model via the association reported in the literature between ELF scores and the subsequent rate of development of cirrhosis. Subject to certain assumptions therefore, it could be said that the traditional research route had provided enough of the evidence needed to plan a trial. It was however clear that it would have been preferable to have information on key change parameters, including those relating to sample composition, available prior to designing a trial, ideally in this case from a cohort study incorporating sequential ELF values. The conclusion about the need for better longitudinal data in order to model the likely numbers of “test positives” more accurately is similar to one drawn in the first simulation study.

The third modelling study

In Chapter 8, a third modelling study had a different focus again. It was conducted from a health economic perspective, and specifically addressed how to optimise the benefit to patients when using an imperfect test for monitoring. The starting point for such an exercise is that the anticipated effects on all four outcome groups of testing (true and false positives, true and false negatives) need to be included in the calculation of benefit. It follows that the number in each of the groups also needs to be known, and that these will depend not only on the numbers of actual cases, but also on the definition of a test-positive in terms of a threshold score.
In the context of diagnostic testing, the academic literature already contained a method for choosing a cut-point in such a way that it would maximise overall patient benefit, taking into account the consequences for patients and budgets of incorrect as well as correct diagnoses. The third modelling study applied these methods to the different needs of monitoring tests, and in so doing highlighted once again the importance of understanding how sample composition is likely to change over time.

Health economists can make reasonable estimates of the benefits and the harms, as well as the costs, consequent on each of the four possible outcomes of testing, namely the patient has the tested-for condition which is correctly identified, or it is missed, and the patient who is well and correctly identified as such, or mistakenly diagnosed as having the condition. The overall benefit to the patient population of different testing strategies is calculated by multiplying up these “per patient” figures by the numbers of patients in the four outcome categories. These numbers will change according to the cut-point used for the test, but also, and crucially, they depend on the proportion of patients in the tested sample who actually have the condition. It follows that the optimum cut-point is sample dependent, since the lower the prevalence of the condition in the sample, the higher will be the proportion of “false alarms” among the patients identified by the test as having the condition. How then might the composition of a monitored patient cohort change over time? And what would be the implications for the optimum cut-point for the monitoring test?

First, the optimum “diagnostic” threshold was calculated, using the best available information on test accuracy and the consequences for the four resulting categories: true and false positives, and true and false negatives. Considering first the patients not identified as having relapsed (the “diagnostic test negatives”) at that time point, it could be anticipated that at any subsequent surveillance point, this sample would contain a mixture of patients who had stayed well, new cases of cancer, and cancer cases missed (“false negatives”) at the earlier administration of the test. Estimating the numbers of new cases required information on the rate of disease progression as well as the time elapsed since the previous testing, while estimating the numbers of previously missed cases drew on knowledge of the accuracy of the diagnostic test, in particular the likelihood that a test negative result would prove to be correct (a test parameter known as the “negative predictive value”). Under plausible estimates of
these values, the sample at the second testing occasion might have a higher or lower disease prevalence than the sample at the time of “diagnostic” testing. A new optimal “monitoring” threshold could then be calculated, reflecting the new prevalence value. If the next step for the “diagnostic test positive” subgroup (i.e., patients identified as having relapsed according to the initial test) was to be continued monitoring, equivalent calculations could be performed, but would produce a different optimum cut point. It was noted that in principle, the approach could be applied to monitoring across multiple occasions.

The potential of a Value of Information approach was explored towards the end of the chapter. For a diagnostic test, “perfect information” consists of the distribution of test scores in “cases” and “non cases”. This information can be used to calculate the probability that a patient with a given test score has been correctly classified, noting that uncertainty about the patient’s true health state will be systematically greater for mid range test scores than for extreme ones. Putting this argument in slightly different terms, misclassification rates are likely to be greater for mid range test scores than for very low or very high test scores. Since mid range scores are also going to be more affected by the choice of cut point (choosing a lower cut point, for example, will increase the numbers of mid range scorers testing positive), it follows that the distribution of uncertainty between the test positives and the test negatives will vary according to the cut point adopted.

In the monitoring context, true disease progression data is required, but additional testing complexities also need to be considered. First, the distribution of test scores in the new cases may differ from that observed in the cases present on the first testing occasion. Second, the likelihood that a previously test negative individual will have become a test positive individual will depend not only on their true change in health status, but also on how likely it is that their test score will have crossed the relevant cut point boundary – which in turn depends on where that boundary is set. Adoption of a low cut point on the first, “diagnostic” occasion, will lead to a reduced likelihood that such a stringently defined subgroup will have changed their real status the next time around, i.e., it reduces the value of information in the monitoring of originally defined test negatives. Finally, as the modelling reported earlier in Chapter 8 showed, the optimum cut point will always reflect the chosen value of health, suggesting a
very complex relationship between the value of information and the value of health in the monitoring context.

**Implications for the design of clinical utility studies**

The literature reviews and the case study identified similar issues. The ways in which factors such as disease progression rate, measurement error, choice of cut-point and monitoring interval can affect clinical utility - and hence, trials to evaluate clinical utility - were then clearly illustrated by the modelling studies.

All three of the modelling exercises concluded that information on changes in the tested sample over time would greatly improve the usefulness of their models. Disease progression is clearly a major element of such changes, and the need to characterise it could perhaps be regarded as self-evident: if the aim of a monitoring intervention is to change the progress of a disease over time, then comparison data on the progress of the disease in the absence of the intervention is clearly a prerequisite for evaluation purposes. Disease progression data is not however provided by the conventional research pipeline, which may partly explain why its importance has been persistently overlooked.

A trial of a marker-guided monitoring strategy will entail specifying that strategy in terms of thresholds, monitoring intervals, etc. How though should those elements be chosen? And is it always necessary to know who were the true “cases” and who were not? Modelling Study 2, and the trial in WS3 of this Programme, were based on identifying an “at risk” group rather than diagnosing “cases” as such. Patients were defined as at risk in terms of ELF score ranges previously observed to be associated with subsequent differences in patient outcomes. By applying an assumed treatment effect, a trial could be modelled and its size calculated, based in the usual way on the amount of benefit it was judged important to detect. If the results of such an adequately powered trial suggested it was effective and cost-effective to monitor with the threshold and the interval used, then it would be tempting to conclude that the research pipeline was working. But how were the threshold and intervals chosen and
were they the best ones? Maybe the threshold adopted to define an at risk group was too low, tipping the balance too much towards detection of every possible case at the expense of increasing the numbers of false alarms? Maybe such frequent testing was unnecessary and longer intervals would have been perfectly adequate? And what if no evidence of effectiveness or cost-effectiveness was found? Where might the explanation lie? And what would then be the way forward – more trials with different combinations of thresholds and intervals?

Modelling Studies 1 and 3 suggest a different approach might produce better value from the trials (Clinical utility) budget. However, in addition to the disease progression information previously mentioned, this approach depends on eventual knowledge of true “caseness”, and so would require more and better test performance information (Analytical validity and Clinical validity) than the pipeline currently supplies. If available, the information could be used to estimate the effects of different monitoring strategies on patients much more precisely than is possible at present, enabling “best bets” to be identified for subsequent evaluation in trials. Better estimates of measurement error, for example, would contribute to increased test accuracy, to reducing misclassifications and hence to increasing benefit (via fewer false negatives) without disproportionate cost or harm (from treating the false positives who do not benefit).

It may be that part of the problem in the functioning of the biomarkers research pipeline is the prominence given at the clinical validation stage to just two indicators of test performance, sensitivity (how well the test performs in relation to cases) and specificity (how well the test performs in relation to non-cases). Both metrics can be derived from the case-control studies which play a prominent role in test development and initial validation, and it is perhaps insufficiently appreciated that people using tests in practice – and also people evaluating test use in practice - are in a fundamentally different position. As in the ELUCIDATE trial, if it is the test result that is intended to be used by the people making patient management decisions, then test performance metrics will be needed which take that starting point into account. The test evaluation literature identifies two such metrics, capturing first, how many test positives turn out to be cases (the Positive Predictive Value, or PPV) and second, how many test negatives turn out to be non-cases (the Negative Predictive Value, or
NPV). To illustrate how misleading reliance on sensitivity and specificity can be in this context - and remembering that even a very specific test will incorrectly identify a small proportion of non-cases as cases (i.e., it has a false positive rate) - then in a low prevalence sample (i.e., one mainly consisting of non-cases), the basic arithmetic of applying a small percentage to a large number will result in many test positive results being generated by non-cases, and a consequent lowering of the proportion of the test positives coming from cases. In a higher prevalence sample, arithmetic dictates that this proportion – the PPV – will be higher.

Applying the same logic to test negatives, in a low prevalence sample, i.e., one containing a small number of cases, a reasonably sensitive test will miss very few of these, and hence add only a very small number of test negative results to the large number generated by the non-cases. It follows that the proportion of test negatives who are indeed non-cases – the NPV – will decrease as prevalence increases, but also that this parameter will change very little across a range of low prevalence values. It also follows that in a low prevalence sample, a lot of the people receiving treatment will not be able to benefit because - even using a very accurate test with high specificity – although they tested positive, they were in fact false positives. Lowering the cut point used to define “caseness” will increase the numbers of people testing positive, but without good prevalence data, it is not possible to estimate what proportion of the test positives have the potential to respond to treatment.

As explained, both PPV and NPV change with the composition of the tested population, i.e., they reflect prevalence - and that is why they cannot be calculated from case-control studies alone. With the correct input parameters, modelling of all these effects would enable a much more realistic picture to be built up of the magnitude of benefit that could potentially accrue from a specific monitoring regime being applied to a specific patient population, although it might also raise challenging questions about the rationale for the monitoring intervals and the thresholds currently in use. From a research perspective, there would be clear advantages in applying such a model to the design of a clinical utility trial.

So, should funders insist on better longitudinal data – on disease progression in the population that will be the subject of the trial AND on how marker scores change over
time? And what about the analytical validity of the marker as measured outside of a research laboratory? And test accuracy results, including PPV and NPV, from a relevant population? Should these be required before funding a monitoring trial? At the conclusion of the Value of Information section of Chapter 8, it was argued that the most efficient way forward for monitoring technologies might be the improved characterisation of test score distributions in cases and non-cases, and of progression rates. The same could perhaps be said to those funding research on the evaluation of monitoring technologies: extra time and effort spent on providing better quality information of this kind would almost certainly lead to a better designed trial - and the overall research duration and budget might not be very different.

**Stage 2: Obtaining patient and public perspectives**

**Establishing a shared understanding**

Following a number of meetings between the three researchers to agree understandings and discuss possible content, the key areas to be taken forward for discussion with PPI representatives were identified as:

- Assumptions about test development from the laboratory bench to the clinic
- Variations in test scores within and between individuals (signal to noise),
- Test accuracy (relationship to clinical condition),
- Detection of cases and non-cases (sensitivity and specificity),
- Prevalence of disease and using prevalence in interpreting tests,
- Rate of disease progression and its impact on treatment strategies,
- Development of treatment strategies to improve patient care
- Development of guidelines and communication about tests between professionals and patients.

**Initial presentation to PPI representatives**

An initial meeting with a PPI representative (one representative was unable to attend) was held in Leeds on 31st July 2015. The key concepts of test measurement, accuracy,
performance and interpretation were highlighted in the context of the NIHR liver biomarkers programme. It was agreed that these issues were important and interesting for patients and the public to discuss. Preparation for a future workshop with a larger group of people was discussed.

Liver North Workshop

Ten people (8F: 2M) took part in the PPI workshop. A range of conditions were represented including:
- 3 x Liver Transplant Patients
- 3 x Primary Biliary Cirrhosis
- 3 x Carer
- 1 x Non Alcoholic Fatty Liver Disease
- 1 x Overdose with potentially hepatotoxic drug
- 1 x Primary Sclerosing Cholangitis

It became clear from questions asked and from the discussion following the presentation that patients were interested in methodological issues about biomarker evaluation. The group members understood the concepts and asked sophisticated questions about the biomarker test development process. The key discussion areas were as noted above.

Patients assumed that all parts of the biomarker pipeline from laboratory bench to clinic were equally researched and considered. Any new biomarker test that was evaluated in a trial or implemented into a service should be “fit for purpose”. Participants stated, “Don’t set up a test, then move on and think we’ve cracked it”. All aspects of test development and evaluation were thought to be equally important and should carry a similar weight in the research pipeline.

Patients stated that it should be made clear what the level of accuracy of the test is in practice and also the number of false positives and false negatives that the test produces. The signal to noise issue was thought to be important and patients indicated that in their opinion only tests that reached a certain level of performance should be taken forward for further evaluation.
The variation in test scores within and between patients was acknowledged with several patients offering examples of how their test results had impacted on their care. At least one carer had noted fluctuations in changes in test scores over time and he used these to monitor the patterns in the health of his partner, concluding they meant very little on their own. He readily understood the notion of a false positive test result, and when he thought about how particular test results could make a big difference to his partner’s care, he was alarmed to think the likelihood of such an outcome might not be well understood by clinicians. Patients agreed that fluctuations in scores or changes in scores were poorly explained by doctors if at all. The fact that snap shot data “how you are today” test results were used to guide patient care and to instigate treatment changes was thought to be poor practice in the context of chronic conditions. All agreed that tests should be part of a wider clinical assessment. With the advent of new technology and the ability to monitor and log routine data it was thought that understanding patterns of change within and between a cohort of patients with Liver Disease over a series of different points in time was really very important. Patients could clearly extrapolate this example to other conditions as well. Patients found it hard to believe that disease progression rates – of central importance to their experience as patients - were not routinely available and routinely used for research purposes. They were almost as taken aback by the lack of interest shown by researchers as well as clinicians in the practical implications of test inaccuracy, and were “astonished” that the routine monitoring of patients was not captured and made use of to support the assessment of new biomarker tests. Participants could pick up on the notions of false positives and false negatives without any difficulty, and because the implications for an individual (clinical but also psychological) of either sort of misclassification were very apparent, they expected high standards to prevail at every stage of the biomarker pipeline.

Patients stated that it was important that all the information required to interpret a test result such as variation in scores (signal to noise), accuracy, prevalence, performance in practice and evidence based treatment strategies should be communicated to patients. Any information that was put into guidelines should be developed with the involvement of patient and public representatives and someone with a plain English remit. Any guidelines developed should be undertaken at a national level and be
reviewed within a relevant timeframe to account for regular updates. A lay summary should be available for patients if they would like to know more information.

Participants stated that the way in which test results are communicated to patients needs careful consideration and involvement of patients and the public to make messages clear. The role of the doctor in promoting good communication about any monitoring test and about the “possibility” of there being other interpretations and outcomes of test results should be part of their education programme. Patients wanted to know about how accurate the test was and the chance of false positives and negatives. How such results might impact on patients in an acute or emergency situation and a routine situation was considered. Supported decision making was advocated for both situations. Further awareness of these issues with patients and the public should be implemented with careful thought being given as to how information is presented and the types of examples that are used. Patients clearly stated that decision making should be facilitated in partnership with health professionals.

**Patient perspectives and the “E” in ACCE**

The “E” in ACCE – the Ethical, legal and social considerations attendant on biomarker testing - have not been addressed in this account so far. Emphasis in the literature to date has been on the potential implications of test results, e.g., genetic tests for late onset conditions, rather than on the quality and efficiency of the research pipeline. But are there downsides for patients of the present, “let’s try it and see” approach to biomarker development? The PPI work reported above suggests first, that quality and efficiency matter to patients in relation to the research pipeline, as well as in direct care, and second, that the widely recognised right of patients to make their own trade-off between length and quality of life when making treatment decisions may also need to be more consistently and transparently applied to decisions about tests. The Health Economic approach provides very useful tools of thought here, in its explicit calculation of harms (physical and psychological) and benefits foregone, as well as the costs, of over- and under-investigation and treatment.
Patients quite rightly expect trials to be well-designed in all relevant respects, but it could be argued that in some of the respects discussed here, they are not well served by the existing evidence pipeline. Research funding mechanisms are currently “tuned” towards traditional A and C1 work in laboratories, and towards clinical trials, but descriptive longitudinal work is not currently attractive to funders, possibly because its contribution to bridging the evidence gap is insufficiently appreciated. Secondary use of data collected for other purposes may offer one practicable way forward.

Patient perspectives are returned to in Chapter 24, towards the end of this report.

The Workstream 1 methodology studies were reported using the PRISMA 2009 Checklist.
Chapter 10 - Biomarker Pipelines – Ensuring Clinical Translation Using Renal Cancer and Kidney Transplantation as Exemplars
This chapter highlights the main aims of the Clinical Translation Workstream, provides the background to this aspect of the Programme and the clinical context and outlines the various elements of the work undertaken which is further described in detail in the following chapters.

Main Aims of the Clinical Translation Workstream (WS2)

Within the context of the overall aim of the Programme in developing strategies to enable the rapid evaluation and translation of promising protein biomarkers into the NHS for patient benefit with a main focus on renal and liver diseases, WS2 focusses on the clinical biochemistry aspects with the main aim of:

- Establishing and maintaining a multicentre sample and clinical data bank in chronic liver disease, renal cancer and renal transplantation, together with the development of a robust system for evaluation of promising emerging biomarkers and related assays, to facilitate rapid biomarker assessment prior to large-scale trialling in practice within the NHS.

Within this, specific objectives are to:

- Undertake systematic reviews of biomarker status in renal cancer particularly relating to prognosis, and in renal transplantation relating to acute complications and long-term outcome with a focus on delayed graft function

- Prioritise selected biomarkers for further evaluation at the multicentre level and develop and employ validation strategies for technical validation of relevant biomarker assays, including determination of significant pre-analytical issues.

- Design and undertake studies examining the clinical utility of the selected biomarkers within the chosen disease areas and identify those biomarkers which justify evaluation in a clinical trial leading to applications for funding.

- Create a clinical sample and data resource to underpin studies within this Programme but also to facilitate future rapid and cost-effective biomarker
evaluation, with the infrastructure and strategies developed providing a blueprint for future similar studies in other disease areas.

Critical Elements in the Biomarker Discovery and Translation Process

General Concepts

With the massive investments in large scale “omics” studies accompanying the increased technological and bioinformatics capabilities, our understanding of the molecular changes and mechanisms underlying many diseases has increased dramatically. It is now understood that within a single disease phenotype or diagnosis there exists underlying molecular heterogeneity, giving rise to previously unrecognised disease sub-groups. With this, there is a drive to use such information to tailor treatment of each patient on an individual basis, recognising the heterogeneity within a disease, variously described as “personalised”, “precision” or “stratified” medicine.\textsuperscript{300-304} Consequently the development of “targeted therapies” has revolutionised treatment in many diseases, particularly many cancers.\textsuperscript{305} This vision is of course predicated on the use of biomarkers to optimise all aspects of the patient pathway from early and accurate diagnosis, through determining the extent and activity of the disease and prognosis, predicting response to therapy allowing optimal treatment selection, and monitoring for treatment response/toxicity or disease progression. Using biomarkers in this way has major potential benefits for patients and the NHS in terms of improving outcomes and providing more cost-effective care, for example supplementing or replacing invasive or expensive procedures or imaging tests, improving patient safety and quality of life and avoiding unnecessary or suboptimal treatment or toxicities. Interestingly however and using therapy-related biomarkers as an example although it extends to other biomarkers types equally, there are currently only 26 FDA-approved “official” companion diagnostics, 10 of which relate to HER2 testing for trastuzumab (Herceptin), although there are considerably more Laboratory Developed Tests (LDTs).\textsuperscript{305, 306} The need to refine existing methods used in economic evaluations of companion diagnostics to include additional characteristics of the test performance itself given the increasing complexities has also been highlighted.\textsuperscript{307}
Genomic analysis of tissue or cells is now routinely used not just in inherited diseases but increasingly in many cancers to allow selection of specific targeted therapies. However many routinely measured biomarkers are proteins present in body fluids and as such remain the province of clinical biochemistry labs. Protein biomarkers have the advantages of being comparatively cheap, easy to measure, relatively non-invasive and providing dynamic information. At the time of application for this Programme in 2008, the depth of knowledge being generated as described above and the extent of its impact in terms of effectively revolutionising several aspects of patient care in some diseases areas weren’t apparent, although expectations were high. Likewise, with the plethora of potential novel protein biomarkers arising from the surge of discovery efforts in clinical proteomic studies, expectations surrounding the introduction of new biomarkers to clinical practice were high. However, even in 2008 this was accompanied by an increasing awareness that the translation of novel biomarkers into clinical practice was not being realised. A pivotal reason for this was the absence of a clearly defined pathway linking biomarker research to health services research, and this was compounded by the technical inadequacies of some of the studies negating any likely translation of the findings. The stages in the “biomarker pipeline” have been variously defined, as illustrated (Figure 24).

*Figure 24 Linear depiction of the phases in the biomarker pathway.*
The dynamic nature of test evaluation in this pathway from biomarker to medical test is more apparent when summarised as a cyclical process (Figure 25) as recently described by the Test Evaluation Working Group of the European Federation of Clinical Chemistry Laboratory Medicine.\textsuperscript{308}

**Figure 25 Cyclical framework for the evaluation of in vitro medical tests.** This framework illustrates that the key components of the test evaluation process are driven by the purpose and role of using a test in the clinical pathway. Reciprocally the key test evaluation elements may influence or modify existing clinical pathways. The outer circle linking the various elements of the test evaluation cycle highlights the interplay between the various components and how e.g. analytical performance may impact clinical performance and vice versa; how clinical performance or effectiveness of a test may call for improved analytical performance and sets new analytical goals for improving the clinical and cost-effectiveness of the test-treatment pathway. Reproduced with permission from.\textsuperscript{308}
Certainly in 2008, the necessary evaluative framework involving multiple stakeholders including academia, industry, healthcare providers and regulatory authorities and with wide-ranging considerations including technology transfer, cost-effectiveness, methodological workflows and regulatory hurdles, wasn’t clear or effective. Without a comprehensive evaluation framework, the route from stringent evaluation to clinical implementation (including evaluation of impact on outcomes) and the realisation of the substantial potential of biomarkers to contribute to improving patient care and health service provision was recognised to be at risk with various issues highlighted from multiple perspectives including academia and industry.\textsuperscript{1-9, 296, 309, 310} Encouragingly though, the need for national strategies to overcome this was also increasingly being appreciated at this time by the NIH in the USA and by the Royal College of Pathologists in the UK, for example.\textsuperscript{11} This Programme was essentially established to explore and address some of these issues, with health economic and trial methodology covered in WS1 and WS3 and with WS2 directed at exploring aspects of infrastructure related to sample banking and assay validation and focussing round renal cancer and renal transplantation as exemplars for biomarker testing.

So has the situation changed in the intervening period and decreased the value of this Programme? The answer to this is a resounding no if the situation is reflected in the
The perceived obstacles and pitfalls in taking a biomarker (or biomarker panel) along the pathway from lab to clinic essentially remain the same in more recent studies as those highlighted earlier.\cite{308, 311-315} Focussing on the lab and clinical biochemistry perspectives, some key aspects are discussed further below and form the subject of specific WS2 activities.

**Specific Lab-based/Clinical Biochemistry Aspects**

Although this Programme covers the later parts of the biomarker pipeline, it is still appropriate to briefly consider some of the issues which are particularly challenging in biomarker discovery and those which contribute to the high level of biomarker attrition at the early stages. Particular technical challenges for novel biomarker discovery in clinical fluids such as serum or plasma include the vast dynamic range of protein concentrations, spanning from albumin at approximately 40-60 g/L down to cytokines at <1 ng/L; a range exceeding the analytical capability of proteomic technologies and necessitating extensive fractionation/enrichment strategies,
particularly with just 22 proteins constituting 99% of the entire plasma protein content.\textsuperscript{10} For urine, low protein concentration and high salt content present challenges, although as approximately 70% of urinary proteins are thought to be kidney-derived it is an attractive alternative source of enriched biomarkers, particularly for renal diseases.\textsuperscript{316} However, normalisation of results is an issue; creatinine is most commonly used although this is not ideal as it is affected by many factors such as muscle mass and renal function. Results from many studies have failed to be confirmed, with a major contributory factor being poor initial study design. This includes insufficient statistical power, inherent study bias due to lack of attention to awareness of potential confounding factors, and the impact of pre-analytical factors on sample quality, although such issue are being addressed with promising results emerging.\textsuperscript{312-315, 317-320} To improve study design and reporting, several guidelines have now been published, such as REporting recommendations for tumor MARKer prognostic studies (REMARK) and STAndards for Reporting Diagnostic accuracy (STARD) in relation to diagnostic markers.\textsuperscript{321, 322}

Other factors preventing further progression include the lack of the necessary tools or resources such as the availability of suitable numbers of stringently collected clinical samples with the appropriate associated clinical data including long-term follow-up data which can take years to accumulate. Biobanking is not coordinated internationally, nor is there a complete central database holding such information about available resources although there are very good examples of integrated activities within certain disease areas or countries, for example the Organisation of European Cancer Institutes (OECI) Pathobiology Working Group through its maintenance of the OECI-TuBaFrost exchange platform, the String of Pearls Initiative in the Netherlands for chronic kidney disease and specific large-scale more general and less disease-focussed national initiatives within the USA, Japan, Iceland, Korea and China for example.\textsuperscript{323, 324, 325} Similarly, examples of publications describing specific biobanks and including indicators of the quality of the samples are relatively few and limited, with some notable exceptions such as that from the Mayo Clinic in Arizona describing the RCC samples held within the Multidisciplinary Genitourinary Diseases Biospecimen Bank, and hence awareness of potential samples for validation purposes is limited.\textsuperscript{326} Encouragingly given the importance of such resources, the
impact of sample processing protocols, pre-analytical factors and quality on biomarker research is increasingly appreciated and the need for standardised approaches to processing, quality assessment and recording of critical variables through initiatives such as SPIDIA, BRISQ and SPREC promoted.\textsuperscript{317, 327-329} Other aspects such as governance, ethical issues, patient involvement, ensuring long-term social, operational and financial sustainability and the importance of considering and collecting indicators of impact, all critical issues for funders as well as researchers, are also assuming a higher level of prominence.\textsuperscript{325, 330-334} Importantly, concerns over the issue of underuse of samples has been raised in a survey of 456 biobanks in the USA and this adds weight to the importance of ensuring visibility and accessibility of such resources to maximise their value and avoid duplication of efforts.\textsuperscript{335, 336}

Gradually these issues these are being addressed. Another bottleneck is the lack of appropriately validated assays. Awareness of potential pitfalls in the level of validation of assays is increasing, including commercial assays, and the need for consolidation/harmonisation and appropriate use of guidelines for validation highlighted.\textsuperscript{337-347} Additionally a disjoin in many cases between the “discovery” researchers and laboratories and the test implementer laboratories, i.e. the clinical biochemistry community with their wealth of experience in using assays and knowledge of the level of performance needed in a routinely available clinical test, has also contributed to the mismatch between expectations and delivery. Undoubtedly, the decline in academic activities of clinical biochemists in the UK, driven by the pressures to deliver an ever increasing hospital workload and with little time to devote to biomarker-related research, hasn’t helped the situation.\textsuperscript{259, 348} Thoughtful reviews by clinical biochemists and colleagues about biomarker progression also add the perspective of the final implementation in hospital labs, not just in terms of the need for evidence-based use but also the practical considerations needed for routine adoption and considerations of reimbursement.\textsuperscript{259, 296, 309}

For the regulatory bodies there is a recognition of the need to be flexible and also to accelerate the approval processes, particularly with the increasing emergence of multiplex marker panels but also in terms of retaining safety whilst not slowing drug
development in the case of companion diagnostics. Efforts by governments and funding bodies to increase the progress made in precision or stratified medicine generally and biomarkers specifically is evident from initiatives in areas such as the development of the infrastructure required to help generate evidence on the clinical and cost-effectiveness of a given *in vitro* test (e.g. the NIHR Diagnostic Evidence Cooperatives (DECs); [www.nihr.ac.uk/about/diagnostic-evidence-co-operatives](http://www.nihr.ac.uk/about/diagnostic-evidence-co-operatives)) and the Innovate UK Precision Medicine Catapult ([www.catapult.org.uk/web/precision-medicine/home](http://www.catapult.org.uk/web/precision-medicine/home)). In the USA and spanning the whole of the biomarker pipeline, the recently established National Biomarker Development Alliance is a trans-sector initiative addressing the issues of a “dysfunctional and disjointed status of biomarker R & D” and aiming to develop widely-accepted standards, best practices and guidelines through an “end to end systems approach.”

### Establishing the Pipeline in Renal Cancer and Kidney Transplantation

#### Renal Cancer

**Clinical Context**

Over one third of a million cases of renal cancer are diagnosed annually worldwide with >143,000 deaths ([www.globoCan.iarc.fr](http://www.globoCan.iarc.fr)). In the UK, it is the 8th most common cancer with ~10,000 new cases annually and the incidence is increasing. For example, in the period between between 1975-1977 and 2009-2011, rates have more than doubled, increasing by 132% overall and 168% in females ([www.cancerresearchuk.org/cancer-info/cancerstats/](http://www.cancerresearchuk.org/cancer-info/cancerstats/)). Occurring with a male to female ratio of 3:2, risk factors include smoking, obesity and hypertension. Most renal cancers (~90%) are renal cell carcinomas (RCC), cancers arising from the renal parenchyma with the most common histological subtype (70-80%) being the conventional (clear cell) RCC (ccRCC). Other main histological subtypes include papillary (10-15%), chromophobe (5-10%), and collecting duct (<1%) tumours and benign oncocytomas (2%-5%), arising from various kidney cell types. Novel subtypes of RCC are continually being defined, largely on the basis of morphology although genetic characterisation is providing further insights, with >24 subtypes of renal cancer now included in the most recent Vancouver classification.  

---

270
Subtypes have broadly differing clinical behaviour and underlying genetic changes although also heterogeneous and complex within a subtype. Major increases in our understanding of the underlying molecular genetic and epigenetic changes and the heterogeneity within the subtypes is leading to changes in clinical management of patients with the most marked impact currently being in the development of targeted therapies.356, 357

The majority of patients have few or no symptoms and a recent international prospective study involving 4288 patients with renal masses found diagnosis was incidental in 67% of patients.358 Approximately two-thirds of patients with RCC present with disease localised to the kidney and surgery or ablative therapies are the standard treatment. Although largely curative, 30-40% of patients will subsequently relapse. Renal cancer is inherently resistant to chemotherapy and radiotherapy and for patients with metastatic disease or those at high risk and warranting adjuvant therapy, the increased biological knowledge has led to the rational design of therapies targeting specific pathways. In the majority (>80%) of sporadic cases of the ccRCC subtype, the Von Hippel Lindau (VHL) tumour suppressor gene has been implicated.359-362 A major role of the VHL protein relates to its role as a ubiquitin ligase, targeting proteins such as members of the transcription factor hypoxia-inducible factor (HIF) family for ubiquitination and subsequent proteasomal degradation. Loss of VHL function leads to accumulation of HIF with consequent transcriptional activation and upregulation of genes including vascular endothelial growth factor (VEGF). In the past nine years, seven agents targeting VHL-related pathways either through kinase inhibition or antibody-based targeting, namely sorafenib, sunitinib, pazopanib, axitinib, everolimus, temsirolimus and bevacizumab, have been approved for treatment of metastatic ccRCC.356, 363 Several other genes have also now been implicated along with epigenetic changes which may lead to further insights into clinical behaviour and therapeutic exploitation.354, 364 Recent developments in immunotherapy with immune check-point inhibitors based on antibodies to CTLA4, PD1 and PDL-1 are also showing promise in RCC.365 As studies start to define the underlying molecular changes in the non-clear cell subtypes, it is likely that novel therapeutic targets may be found.366-370 However, it is clear that no dominant driver gene equivalent to VHL exists, although MET represents a promising possibility in papillary tumours. MET gene mutations are only evident in
12% of sporadic cases but copy number gain has been found in 81% and 46% of types I and II papillary RCC respectively and MET protein is overexpressed in up to 90% of cases. A range of tyrosine kinase inhibitors with activity against MET including foretinib, capmatinib, crizotinib and volitinib are currently in clinical trial or planned.

There are currently no circulating biomarkers which are routinely used in RCC although clear clinical needs have been identified and such developments are a priority research area. Clearly circulating biomarkers have the advantage of being accessible and measured relatively non-invasively and are able to provide information longitudinally, even prior to surgical or other treatment. Biomarkers could potentially impact the patient pathway by enabling earlier diagnosis, determining prognosis for stratification of follow-up, detecting relapse and selecting patients for specific therapies by predicting response. A particular challenge, both in terms of diagnosis and prognosis, relates to the management of small renal masses (≤ 4 cm), up to 25% of which are benign and the risks of surgery or ablative procedures, particularly in elderly patients with comorbidities, have to be balanced against the risk and timeframe of the tumour progressing.

At the time of the application for this Programme grant there were few emerging diagnostic biomarkers or predictive biomarkers although several prognostic biomarkers seemed to have apparent potential utility and hence the focus was on prognostic biomarkers. We have recently reviewed existing potential tissue and fluid biomarkers and there are now also some promising initial studies in the areas of diagnosis and prediction of response to therapy. For example, urinary concentrations of the two proteins aquaporin-1 (AQP-1) and perilipin-2/adipophillin (PLIN-2/ADFP) have been shown in several studies by the same group to be significantly elevated in patients with RCC compared with healthy, benign renal and surgical controls, declining post-nephrectomy. These findings have been extended to 720 patients undergoing CT scans for a variety of indications using a newly developed ELISA for AQP-1 where a ROC AUC of 0.99 was achieved although elevated concentrations were also seen in other malignancies. Clearly these and some other markers show promise but lack of the necessary level of evidence in terms of numbers and sizes of studies, independent validation and
availability of robust assays precludes a focus on exploring these further within this Programme at present. However it is anticipated that such biomarkers may be the subject of future studies utilising the sample banks assembled within this Programme.

Determining prognosis in RCC

The ability to stratify patients according to risk is highly desirable. Detection of relapse in patients with localised disease is based on repeated imaging, in some cases with biopsy. This is expensive and has the associated risks of cumulative radiation exposure and the potential morbidity associated with the biopsy procedure. Although postoperative surveillance protocols have been developed, detection of relapse is often not optimal and the ability to rationally guide the surveillance regimens on an individual basis and also to identify high risk patients for adjuvant therapy has obvious benefits. A critical determinant of a patient’s prognosis is the stage of the cancer and currently this is determined using the TNM system based on the extent of the primary tumor (T), whether local nodes are affected (N) and whether metastatic disease is present (M). Other recognised independent clinicopathological prognostic factors include tumour grade and the presence of necrosis and various prognostic models or algorithms have been developed incorporating these and other factors. Problems with these include the subjective nature of some elements such as tumour grade and the grouping of patients into a limited number of risk groups, meaning that estimates of risk can be wide for individual patients. This is exemplified in the widely used system for patients with localised disease developed at the Mayo Clinic which integrates pathological T stage, N stage, tumour size, nuclear grade and tumour necrosis. This divides patients into low, intermediate and high-risk groups with estimated 5-year metastases-free survival rates of 97.1%, 73.8% and 31.2% respectively. However the challenge is being able to further stratify these patients, particularly those placed into the intermediate risk category.

For patients with metastatic disease, the most widely used clinical prognostic model is that proposed by the International Metastatic RCC Database Consortium (IMDC) which similarly groups patients into risk categories depending on the number of poor prognostic features present (Karnofsky performance status of less than 80%, less than 1 year from diagnosis to treatment, anemia, hypercalcaemia, neutrophilia, and
For the favourable, intermediate and poor risk groups, median overall survival of 43.2, 22.5 and 7.8 months respectively have been reported. For the favourable, intermediate and poor risk groups, median overall survival of 43.2, 22.5 and 7.8 months respectively have been reported.

Pre-operative nomograms which don’t include histopathological features but are based around parameters such as age, gender, symptoms, and CT-determined tumour size, T stage and metastasis have also been developed with the advantage of potentially being used to determine optimal treatment strategies both surgically and in terms of neo-adjuvant therapies. Tumour expression of selected proteins such as B7-H1, survivin and Ki-67 have been shown to have independent prognostic significance for CSS either alone or within an algorithm “BioScore”, and to add prognostic value to established clinicopathological models, including the SSIGN score. Other similar examples with prognostic utility include Ki-67, p53, endothelial VEGFR-1, epithelial VEGFR-1 and epithelial VEGF-D combined with performance status and T stage, and in metastatic RCC, carbonic anhydrase IX (CAIX), PTEN, vimentin and p53 combined with T stage and performance status. Recent genomic and transcriptomic studies are adding to this, for example a 34-gene classifier, ClearCode34, assigns patients to good risk (ccA) or poor risk (ccB) groups and a 16-gene signature categorises patients into low, intermediate or high risk groups. In terms of circulating biomarkers, several routinely measured haematological and clinical biochemistry factors including the neutrophil/lymphocyte ratio (NLR), thrombocytosis, haemoglobin, serum sodium, calcium and CRP have been found to be prognostic, as have several more directly tumour-related proteins including VEGF and carbonic anhydrase IX (CAIX) - attractive possibilities in terms of their ease of measurement even pre-operatively (reviewed in Chapter 12). Several of these have sufficient evidence to justify their exploration in a large-scale Programme of this type, and in particular to determine whether as a multiplex panel or in combination with additional clinicopathological parameters or models, they could provide superior performance to existing models.

Kidney Transplantation
**Clinical Context**

Chronic kidney disease (CKD) affects approximately 8-16% of the adult population and is increasing dramatically, at least in part linked to global epidemics in diabetes, hypertension and obesity.\(^{397-399}\) It is estimated that 2-4% of people with CKD will eventually develop end-stage kidney disease (ESKD) and require renal replacement therapy (RRT; dialysis or transplantation) with 2.6 million people receiving RRT in 2010 and a similar number being unable to access it and dying prematurely.\(^{397}\) Kidney transplantation represents the gold standard treatment for patients with ESKD, providing improved quality of life and survival compared to commencing dialysis.\(^{400-402}\) A functioning kidney transplant is able to re-establish many of the important functions that the kidneys perform beyond what dialysis can deliver. For example, the kidneys have important endocrine functions in terms of the regulation of blood pressure, producing erythropoietin (stimulates red blood cell production) and activated vitamin D3 in addition to controlling electrolyte, acid/base and fluid balance. It has also been recognized that kidney transplantation is of economic benefit to society compared to treatment with dialysis.\(^{403, 404}\) In the first year the cost of kidney transplantation is equal to that of haemodialysis, however in subsequent years the cost is halved. The average cost of maintaining a patient with ESKD on dialysis is £30,800 per year. The cost of a kidney transplant is £17,000 per patient per transplant leading to a cost benefit in the second and subsequent years of £25,800 pa. At the end of March 2009, over 23,000 people in the United Kingdom had a functioning kidney transplant, saving the NHS over £512M for that year in the dialysis costs that they would need if they did not have a functioning kidney transplant (www.kidney.org.uk). At a time of financial pressure to all health care systems this is an important aspect to consider.

There are a number of different sources of kidneys for transplantation. The current classification can be broadly divided into deceased-donor or living-donor kidney transplantation depending on the source of the donor organ.\(^{405, 406}\) Deceased-donation is further classified as donation after brain death (DBD) and donation after circulatory death (DCD). Kidneys donated after a circulatory death (DCD) experience a longer period of warm ischaemia (reduced blood flow) prior to surgical retrieval in comparison to those kidneys donated after brain death (DBD). The length of time between the surgical removal of the donated kidney and its implantation into the
recipient is a major factor in determining whether there is immediate kidney function or delayed kidney function, referred to as delayed graft function (DGF). Therefore deceased-donation is associated with a higher rate of DGF compared to living-donation and DCD kidney transplantation has a higher rate of DGF compared to DBD kidney transplantation. Living-donor kidney transplants are characterized as genetically related between parents or siblings (living-related) or non-related between husband and wife, partners and friends (living-unrelated). More recently there has been an increase in the number of altruistic donors who wish to donate a kidney to anyone in society who is on the kidney transplant waiting list.

The need for kidney transplantation is far greater than the availability of organs. In 2012 there were 2,998 kidney transplants performed in the UK. However there are greater than 6000 people on the UK kidney transplant waiting list.\textsuperscript{407} The reality is that people will die unnecessarily whilst on the waiting list for a kidney transplant. Due to the ongoing shortage of kidney transplants the average waiting time for a deceased donor kidney transplant is two to three years. There have been a number of initiatives to increase the number of organ donors. These initiatives have included the recently introduced opt-out system in Wales where the assumption is that people are willing to donate unless they opt out, and accepting kidneys from expanded-criteria donors (ECD).\textsuperscript{406} Expanded criteria donors are defined as those aged > 60 years or aged $\geq$ 50 years with two of the following: a terminal serum creatinine level $\geq$ 1.5 mg/dl (132 µmol/L) , a cerebral vascular accident as the cause of death or a donor history of hypertension. Patients receiving ECD kidney transplants are consequently at a higher risk of developing DGF. Pooled living kidney-donation has been introduced to allow pairs of kidney donors and potential recipients who could not donate to one another because of immunological barriers to be matched with other pairs in the UK to whom they could donate.\textsuperscript{405} Further advances in medicine have allowed kidney transplantation to occur across blood group and HLA incompatibility.\textsuperscript{408, 409} The government has recognized the importance of kidney transplantation and has increased the number of nurses specialized in organ donation (SNODs) to help identify more potential donors.

Unfortunately kidney transplants have a finite lifespan. A number of factors determine how long a transplanted kidney lasts.\textsuperscript{405, 406} These include donor age and
the nature of the donation e.g. living donation versus deceased donation, immunological matching with respect to blood group and HLA compatibility, baseline function of the donated kidney, immediate versus delayed graft function, the primary kidney disease and the health of the recipient. Overall the average kidney transplant survival times are about 95% in the first year, 85-90% at 5 years and 75% by 10 years.\textsuperscript{410} Living-donor kidney transplants have the better long-term outcomes than those from deceased donors. This is in part due to the controlled nature of the donor assessment and in part due to the reduced ischaemic time and subsequent DGF.

Following kidney transplantation there are a number of complications that can occur. The early complications are those common to any surgical procedure including post-operative chest or wound infection.\textsuperscript{411} Early post-operative complications that are specific to kidney transplantation include renal vein thrombosis, DGF, acute rejection (AR) and calcineurin inhibitor toxicity. All of these complications result in a failure of the transplanted kidney to function effectively. Renal vein thrombosis is detected on renal tract ultrasound scanning and is rarely reversible with devastating consequences. A kidney transplant biopsy is required to distinguish between DGF, AR and calcineurin inhibitor toxicity. This procedure is not without risk and is resource intensive. Non-invasive diagnostic techniques are needed to allow rapid diagnosis and appropriate treatment.

\textbf{Delayed Graft Function (DGF)}

Overall DGF has a higher risk of mortality, transplant loss, need for dialysis, kidney biopsy and increased hospital stay.\textsuperscript{412, 413} Management strategies promoting renal function post-transplant have major implications for patient benefit and resource savings.\textsuperscript{414} DGF is secondary to the ischaemia-reperfusion injury (IRI) that follows retrieval of the kidney from the donor with attendant loss of perfusion followed by implantation into the recipient, with restoration of perfusion. During this, the kidney is stored on ice and transported to the designated kidney transplant unit. This period of time is referred to as the cold ischaemic time (CIT). Generally speaking the longer the CIT the more likely the kidney will experience IRI which is manifest clinically as DGF.
The multiple definitions of DGF have hindered the ability to characterize its incidence and outcomes. As many as eighteen definitions have been used in the literature.\textsuperscript{415} The most common definition proposed has been the receipt of dialysis within 7 days following transplantation.\textsuperscript{412} However this is still a crude definition and holds back the opportunity to stratify different degrees of ischaemia-reperfusion and outcomes. Within this Programme, any analysis has used the following definitions and using the creatinine reduction ratio (CRR = (day 0 SCr minus day 7 SCr)/day 0 SCr): immediate graft function where there is a significant and sustained fall in creatinine within the first 48 hours (IGF; CRR \geq 0.7), slow graft function where serum creatinine fails to fall significantly in the first 48 hours but the patient does not receive dialysis in the first week (SGF; CRR < 0.7), and DGF where the patient receives dialysis in the first week except for isolated hyperkalaemia.

DGF ranges from 5-50\% following deceased donor transplantation and depends on many factors.\textsuperscript{416, 417} Not surprisingly DGF is significantly lower in live donations at 4-10\%.\textsuperscript{418} As previously described the variations in the reported incidence will depend upon the definition of DGF that has been used.\textsuperscript{419} The United Network of Organ Sharing database (www.unos.org) recorded DGF in US patients at 21.3\% in early 2011. A number of donor and recipient factors have been identified as contributing to the development of DGF. There has been an attempt to develop risk calculators for DGF.\textsuperscript{420, 421} These factors can be categorized as recipient related and include gender, ethnicity, previous transplantation, diabetes mellitus, HLA-mismatch, peak panel reactive antibodies, previous blood transfusions, body mass index, duration of dialysis before transplantation. Donor related factors include donor age, CIT, warm ischaemic time (WIT), DBD versus DCD, hypertension, baseline serum creatinine, cause of death, and weight.\textsuperscript{412, 420-424}

DGF is extremely important as an independent risk factor for both early and late kidney transplant loss and increased mortality.\textsuperscript{22, 425, 426} DGF lasting for more than six days strongly decreases long-term transplant survival.\textsuperscript{427} It has been proposed that the kidney donor type may be more important than current DGF definitions in understanding the impact of DGF on longer-term outcomes. This is corroborated by the fact that the prediction of poor outcomes associated with DGF is largely independent of many of the definitions in use and provides further support for the
stratified approach to defining DGF that is proposed. It has been shown that the severity and duration of native acute kidney injury (AKI) secondary principally to ischaemia-reperfusion injury (IRI) predicts the risk of chronic kidney disease. There are obvious parallels between DGF and AKI but with the caveats that there are other factors at play such as immunological responses and immunosuppressant medications.

Delayed graft function is associated with other adverse outcomes such as acute rejection (AR). Acute rejection is more likely to occur during an episode of DGF due to increased exposure of donor epitopes and has a significant impact on kidney transplant survival. It is therefore important to identify the underlying cause of the DGF following kidney transplantation. It is usual to perform a kidney transplant biopsy to confirm the histological diagnosis of tubular injury from IRI and exclude the possibility of AR. The clinical management of DGF involves close attention to detail with appropriate management of the patients medications, acid/base, electrolyte and volume status. There can be a risk of volume overloading patients particularly if they are oliguric. The dose of specific immunosuppressants e.g. calcineurin inhibitors, is usually lowered in the setting of IRI due to their vasoconstrictive effects whereas in the case of AR the dose of immunosuppression is increased.

In summary the early identification of DGF and the specific underlying pathology has significant potential to improve immediate patient management, allowing fluid volume status optimisation, timely appropriate dialysis, and avoidance of unnecessary investigation and treatment. With the increased use of donor kidneys from deceased and extended criteria donors to meet the demand for transplants there is a concurrent need for new biomarkers to improve assessment of the quality of donated kidneys prior to transplantation and enable more objective decisions about viability. The opportunity to stratify patients and identify those with significant IRI may allow the individualisation of immunosuppressive regimens (e.g. avoidance or lowered dose of calcineurin inhibitors) at an earlier time point may in turn result in improved long-term outcomes. Similarly improved monitoring of kidney function, prediction pre- or post-transplant or earlier diagnosis of early post-operative complications such as DGF or AR, or longer-term prognostic information would allow earlier intervention, avoidance of biopsies and tailoring of longer term immunosuppression. Efforts to
identify novel biomarkers of DGF to improve upon the use of serum creatinine are increasing and potential new biomarkers are reviewed in Chapter 12.

**Overview of the Work Undertaken in the Clinical Translation Workstream (WS2) to Develop and Use These Pipelines**

An overview of the activities and key deliverables of WS2 is shown below (Figure 26).

*Figure 26 An overview of WS2. Key activities and deliverables are shown together with the inter-relationships with WS1 and WS3.*

Sample Banks and Clinical Data

Essentially to overcome the inherent inertia in the later stages of the biomarker pipeline where often readily available sample sizes are inadequate to provide sufficient statistical power, sample collection has been inconsistent, clinical data may
be incomplete and follow-up time is not mature enough to produce enough events, a key element of WS2 is the assembly of fluid samples together with associated clinical data, including long-term follow-up, from cohorts of patients with chronic liver disease, renal cancer or ESKD undergoing renal transplantation, the latter two being major areas of interest in Leeds. This will provide a valuable underpinning resource for this Programme and future collaborative biomarker studies aimed at evaluating whether specific putative novel biomarkers are likely to benefit patients and health services. This has been achieved through the RCT in WS3 for liver diseases with cross-sectional sample collection and through involving centres across the Local Research Networks for longitudinal and cross-sectional sampling in renal transplantation and renal cancer (10 and 11 centres respectively), and is described in detail in Chapter 11. In total this has involved 1,967 patients and 149 healthy volunteers and 5,976 sample timepoints, with many of these sample timepoints also including multiple sample types being banked such as serum, plasma and urine and each stored in multiple aliquots to maximise future use. Standard operating procedures (SOPs) have been used for the collection and processing of samples and data collected according to specific case report forms (CRFs) via the Leeds Clinical Trials Research unit (CTRU) with samples shipped to Leeds and ultimately stored within a licensed research tissue bank (RTB) to maximise future use and benefits. The theoretical advantages to this and the level of stringency are thus similar to that of a clinical trial (Figure 27).
These prospective samples provide the ability to evaluate biomarkers in the following contexts:

- **Chronic liver disease** – identification of patients at risk of the subsequent development of cirrhosis and major liver events such as hepatocellular carcinoma.
- **Renal transplantation** – prediction/earlier diagnosis of patients with acute post-transplant complications such as DGF or acute rejection and prognostic stratification for long-term outcome.
- **Renal cancer** – prognostic stratification and longitudinal monitoring with other possibilities including diagnosis and prediction of response to therapy.
- **Healthy controls** – enabling determination of reference ranges and effects of factors including gender, age, ethnicity and diet.
Biomarker Prioritisation, Assay Validation and Evaluation of Clinical Utility

Focussing on the renal diseases, the identification of potential biomarkers to evaluate currently is based on our ongoing discovery activities relating to novel biomarkers, a systematic review of the relevant literature which is the focus of Chapter 12, and by approaches from other groups. A certain level of evidence must exist already (e.g. significant independent association with prognosis) with selected biomarkers already having been the subject of initial validation studies and agreed by a sub-panel of applicants as being of suitable scientific quality. Where promising biomarkers exist but with a lower level of evidence or requiring further technical assay validation, further evidence has been sought in some cases through using our own local Leeds Multidisciplinary RTB. This phased approach ensures independent validation (or failure to validate) whilst conserving the NIHR Programme-related samples for final multicentre validation studies and ultimately biomarkers emerging from this would be the subject of future clinical trials similar to that described in WS3. This strategy is illustrated below in Figure 28.

Figure 28 The phased approach to biomarker validation with evidence-based progression and utilisation of sample banks depending on the stage of the biomarker.

Within the timeframe of this Programme, assays for several potential biomarkers prioritised following systematic review have been systematically evaluated and validated, including in some cases analysis of specific pre-analytical aspects, as
described in Chapter 13. Following this the prognostic use of serum and plasma VEGF, plasma osteopontin and CAIX and serum CRP, alone or in combination, together with extensive clinicopathological variables, have been explored in RCC and this is reported in Chapter 14. In renal transplantation, suitable promising biomarkers have been similarly prioritised and once the outcome data is mature enough in the next 2-3 years, appropriate studies will be taken forward.

**Additional Deliverables**

The deliverables relating to the activities described above are summarised in Chapter 15, additionally highlighting others including industry partnerships, academic collaborations, generation of intellectual property, the additional utilisation of the RTB and plans for long-term sustainability of the resources generated.
Chapter 11 - Establishment of multicentre prospective observational cohorts with sample banks for biomarker validation
This chapter describes the development of three multicentre prospective observational cohorts, with high quality biospecimens, in renal cell carcinoma, renal transplant and liver disease. These cohorts have been established to enable the rapid clinical validation of new biomarkers, as exemplified in Chapter 14. In addition to providing a summary of the final cohorts, this chapter details and discusses the issues pertaining to study design, management and governance, providing some generalizable learning for future researchers establishing similar resources.

**Disease areas**

Patients were recruited to provide samples and data for future biomarker research from the major diseases under study, including renal cancer and renal transplantation as described in the previous chapter, liver disease through the linkage to WS3 and the clinical trial, and healthy controls. This latter cohort was recruited to determine reference ranges and potential biological or technical confounding factors such as age, gender and length of storage. Over the duration of the programme the three studies recruited 2116 participants in total and 5976 samples. These include:

- 847 liver disease patients, with 847 serum samples
- 514 patients on the transplant waiting list including 312 subsequently transplanted, with 3806 samples, each sample including multiple aliquots of serum, plasma and urine
- 706 renal cancer patients with 1132 samples, each sample including multiple aliquots of serum, plasma, buffy coat and urine.
- 149 healthy volunteers with 191 samples, each sample including multiple aliquots of serum, plasma and urine

**Renal Cancer**

**Objectives and Endpoints**

The primary objective for establishing this cohort was to provide prospectively collected high quality clinical samples and data, from multiple centres, to validate prognostic and longitudinal monitoring biomarkers of RCC. The end points for such studies would be to determine the association between these markers and outcome (disease-free survival, cancer-specific survival and overall survival) and their ability to detect relapse, when measured longitudinally.
Eligibility criteria

RCC Inclusion criteria:

- Newly diagnosed suspected RCC (all stages)
- All histological types of RCC
- No prior treatment for renal cancer
- Ability and willingness to provide written informed consent;
- Ability and willingness to co-operate with study procedures, including blood and urine sampling;
- Age ≥18 years.

RCC Exclusion criteria:

- Diagnosed familial RCC e.g. VHL syndrome
- Renal cancer acquired following/during renal dialysis
- High risk or known HIV/AIDS, HBV and HCV etc

Healthy Volunteer Inclusion criteria:

- Able to provide consent
- Willingness to cooperate with study procedures
- Age ≥18 years.

Healthy Volunteer Exclusion criteria:

- History of any cancer
- High risk or known HIV, AIDS, HBV and HCV.
- History of diagnosed renal disease
- Current/recent (within the last 3 months) UTIs

Study Design
A multicentre prospective observational cohort design, for retrospective blinded biomarker validation was adopted. Blood and urine samples were requested from eligible patients, attending eleven participating centres shown in Table 67, who were diagnosed with suspected RCC, according to the Study Site Operating Procedure (SSOP) as summarised below:

Cross-sectional study (target of n=500 prior to surgery or other treatment):
- a single baseline blood (12-18 mls – for serum, plasma and buffy coat)
- mid-stream urine sample

Longitudinal monitoring study (target of n=200 prior to nephrectomy):
- two separate baseline blood (for serum, plasma and buffy coat) samples
- Two separate baseline urine samples
- Additional blood and urine samples at 3-6, 12, 18 and 24 months post-nephrectomy, ceasing at relapse if earlier with a sample taken at that time.

Table 67 Centres recruiting patients with renal cell carcinoma

<table>
<thead>
<tr>
<th>Site Code</th>
<th>Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>N0000015</td>
<td>Charing Cross Hospital - Imperial College Healthcare NHS Trust (London)</td>
</tr>
<tr>
<td>N0000039</td>
<td>Nottingham City Hospital</td>
</tr>
<tr>
<td>N0000050</td>
<td>St James's University Hospital (Leeds)</td>
</tr>
<tr>
<td>N0000069</td>
<td>Freeman Hospital (Newcastle)</td>
</tr>
<tr>
<td>N0000131</td>
<td>Lister Hospital - East and North Hertfordshire NHS Trust</td>
</tr>
<tr>
<td>N0000132</td>
<td>Northwick Park Hospital</td>
</tr>
<tr>
<td>N0000153</td>
<td>Churchill Hospital (Oxford)</td>
</tr>
<tr>
<td>N0000221</td>
<td>Stepping Hill Hospital (Stockport)</td>
</tr>
<tr>
<td>N0000352</td>
<td>University Hospital of Wales (Cardiff)</td>
</tr>
<tr>
<td>N0000361</td>
<td>Western General Hospital (Edinburgh)</td>
</tr>
<tr>
<td>N0000537</td>
<td>Gartnavel General Hospital (Glasgow)</td>
</tr>
</tbody>
</table>
For each patient undergoing nephrectomy, a representative formalin-fixed paraffin-embedded (FFPE) block of tumour tissue was also collected. This was not included in the Programme application and was funded from elsewhere but provided added value to the sample collection through enabling multiple centres to participate in tissue-based studies in the future given that the clinical data was already being collected. In Leeds Teaching Hospitals Trust (LTHT) frozen tissue was also stored as this was being undertaken routinely already as part of local research.

Cross-sectional blood and urine samples were also collected from healthy volunteers across the different centres (target of n=200; relatives, hospital staff etc) to allow the determination of biomarker reference ranges and effects of factors including gender, age, ethnicity and diet.

The study schema is outlined in Table 68.
### Table 68 Renal Cell Carcinoma Study Schema

<table>
<thead>
<tr>
<th>Activity</th>
<th>Screening</th>
<th>Pre-treatment (baseline)</th>
<th>Nephrectomy</th>
<th>3-6 mth</th>
<th>12 mth</th>
<th>18 mth</th>
<th>24 mth</th>
<th>36 mth</th>
<th>48 mth</th>
<th>60 mth</th>
<th>Relapse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consent</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eligibility</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-sectional cohort (n=500)</td>
<td>Blood Sample&lt;sup&gt;b&lt;/sup&gt;</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Urine Sample</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Longitudinal cohort (n=200)</td>
<td>Blood Sample&lt;sup&gt;b,**&lt;/sup&gt;</td>
<td>x (2)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>X&lt;sup&gt;f&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Urine Sample&lt;sup&gt;**&lt;/sup&gt;</td>
<td>x (2)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Tissue Sample</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographics</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk factors</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symptoms</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Radiology data</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Pathology data</td>
<td>x&lt;sup&gt;f&lt;/sup&gt;</td>
<td>x&lt;sup&gt;**&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sites of disease at baseline and relapse</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Treatment details</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>ECOG PS</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comorbidities</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concomitant medications</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Haematology, biochemistry</td>
<td>x</td>
<td></td>
<td></td>
<td>x&lt;sup&gt;***&lt;/sup&gt;</td>
<td>x&lt;sup&gt;**&lt;/sup&gt;</td>
<td>x&lt;sup&gt;**&lt;/sup&gt;</td>
<td>x&lt;sup&gt;**&lt;/sup&gt;</td>
<td>x&lt;sup&gt;**&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survival assessment</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>
**Longitudinal cohort only:** 2 pre-treatment baseline samples taken if possible, 4-60 days apart; aHaematology, biochemistry data collected retrospectively at the next annual follow up, if available; bA final sample to be taken at relapse; ^ cases where patient undergo biopsy in the absence of nephrectomy; ++ cases where patients undergo nephrectomy only; bBlood denotes serum and plasma samples with buffy coat.
Statistical considerations
The design and analysis/reporting of many prognostic marker studies has been criticised and we will conform to REMARK guidelines.\textsuperscript{321, 432-434} There are many imponderables involved in calculating the sample size for biomarker studies. Patient recruitment in terms of numbers has been powered based on a consideration of our experience of ongoing marker analysis, for example cathepsin D where a 15% difference is seen between groups at 2 years and CRP where the difference is 40-50%.\textsuperscript{435, 436} Based on relapse rates ranging from 12.5% to 27.5% (i.e. 86 to 190) of the population diagnosed with clear cell RCC undergoing nephrectomy or ablation, we have modelled numbers ranging from 80% power to 90% power of correctly separating prognostic groups. A marker with 15% separation would require 216 to 380 patients in total to ensure numbers are sufficient for a power of 90% assuming event frequencies range from 12.5% to 27.5% at 2 years whereas a marker separating groups by 40% would require a total number of 72 patients. Given the projected figures of 400 clear cell patients in total then these targets are easily met. Cox multivariate proportional hazards model analysis will be used to identify multiple marker combinations, with significance levels adjusted for the number of variables included. About four times as many patients are required to detect interaction affects for a pair of dichotomous variables as for an individual variable, so only combinations of markers with relatively large effects are likely to be identified definitively. Simulations and sensitivity analyses will be performed to validate and confirm significance levels for such analyses.

For the longitudinal monitoring, methodology for evaluating biomarker test performance is more poorly developed. Experience suggests that, with up to 5 marker values throughout the follow-up period on 200 patients we would have adequate numbers to show the predictive capacity of relevant markers in this setting, and to develop mathematical models, as appropriate, to enhance our understanding of the disease process in these patients.\textsuperscript{437}
Renal Transplant

Objectives and Endpoints
The primary objective for establishing this cohort was to validate biomarkers for use in monitoring and diagnosis of early/acute kidney transplant complications. The secondary objective was to validate biomarkers for use in predicting patient outcomes and loss of transplant function. The endpoints of such studies would be to determine the association between the concentration of these biomarkers and diagnosis or prediction of disease/outcome.

Eligibility Criteria
Inclusion criteria:
- All sites: Active patients on the renal transplant waiting list
- Ability and willingness to provide written informed consent;
- Ability and willingness to co-operate with study procedures, including blood and urine sampling;
- Age ≥18 years.

Exclusion criteria:
- High risk or known HIV/AIDS, HBV and HCV or similar infectious diseases.
- Patients in the custody of HM Prison Service

Study Design
A multicentre prospective observational cohort design, for retrospective blinded biomarker validation was adopted.210 Blood and urine samples were requested from eligible patients attending participating centres and who were on the waiting list for renal transplantation. The ten centres participating in this study, shown in Table 69, were to recruit up to 850 renal transplant patients from the waiting list, aiming for a target 300 deceased donor and 40 live transplantations within the recruitment period of the study. The study was focused on patients receiving deceased donor kidney transplants due to the increased risk of DGF and chronic transplant dysfunction longer term and therefore providing sufficient samples and events for assessment of biomarkers.
Blood and urine samples were to be obtained following consent, whilst on the waiting list. Where possible a second baseline sample was to be collected immediately pre-transplant. Samples were then collected daily during the first week of hospital stay (~5 samples), then at the following intervals post-discharge: weekly for one month, then at 2, 3 and 6 months. The study schema is outlined in Table 70.

**Table 69 Renal Transplant Study Centres**

<table>
<thead>
<tr>
<th>Centre No</th>
<th>Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>N0000046</td>
<td>Royal Liverpool Hospital</td>
</tr>
<tr>
<td>N0000050</td>
<td>St James's University Hospital (Leeds)</td>
</tr>
<tr>
<td>N0000065</td>
<td>York Hospital</td>
</tr>
<tr>
<td>N0000069</td>
<td>Freeman Hospital (Newcastle)</td>
</tr>
<tr>
<td>N0000078</td>
<td>Hull Royal Infirmary</td>
</tr>
<tr>
<td>N0000110</td>
<td>Queen Alexandra Hospital (Portsmouth)</td>
</tr>
<tr>
<td>N0000118</td>
<td>Derriford Hospital (Plymouth)</td>
</tr>
<tr>
<td>N0000230</td>
<td>Southmead Hospital (Bristol)</td>
</tr>
<tr>
<td>N0000232</td>
<td>Northern General Hospital (Sheffield)</td>
</tr>
<tr>
<td>N0000299</td>
<td>St Luke's Hospital (Bradford)</td>
</tr>
</tbody>
</table>
### Table 70 Renal Transplant study Schema

<table>
<thead>
<tr>
<th>Activity</th>
<th>Screening</th>
<th>Pre-Transplant</th>
<th>Daily during week 1 of hospital stay</th>
<th>At discharge from hospital</th>
<th>Weekly for next month</th>
<th>2, 3 and 6 months post-discharge</th>
<th>Annually for up to 5 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consent</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eligibility</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research Samples</td>
<td>Blood</td>
<td>x</td>
<td>x</td>
<td>X</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Samples</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Urine</td>
<td>x</td>
<td>x</td>
<td>X</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical Data</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Follow up Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

- Blood denotes serum and plasma samples
- Samples for research to be taken at the same time as clinical samples.
Statistical Considerations

The frequency of sampling used in this study was designed to enable the clinical validation of biomarkers of potential use for distinguishing between DGF and AR, detecting DGF or rejection earlier and potentially predicting chronic kidney transplant dysfunction and patient outcomes. At the time of the application it was only possible to provide approximate estimates of sample sizes for this proposal. Firstly, several different patterns could be examined in the data that may predict DGF, AR and later clinical or sub-clinical dysfunction. For instance, a sudden rise in marker levels, marker levels elevated over time, a slow steady rise, or a sharp rise followed by a sudden drop. Secondly, normal variability in putative markers is uncertain, and will have to be determined for each individual biomarker. With 20 samples per patient, a slow steady rise, which is likely to be the most difficult pattern to distinguish, can be estimated fairly accurately. Preliminary simulations suggest that, in this case, assuming a linear increase in marker levels, with a ±10% 'measurement' error (due to patient, pre-analytical and analytical factors) we could clearly detect a 10% increase over six months. Rates of increase can then be used as predictors of DGF, AR and later rejection/dysfunction. We have experience with alternative models, which can be accommodated with similar magnitudes of variability. Assuming we have accurate parameters from such models, they can be used to predict DGF, AR or later clinical or sub-clinical rejection. Such predictions need to be highly discriminatory to make them clinically useful. For instance, to differentiate between patients with a 10% chance of DGF ± AR versus a 40% chance. This requires sample sizes of approximately 100 patients (with 90% power, 5% false-positive), assuming the model parameters are clearly estimated with a small standard error. We also have experience with other projects involving serial measurements to guide us, where sample sizes of the order of 200 patients, and multiple measurements over time, were sufficient to enable us to distinguish clinically meaningful effects (see, for example Migdal et al, 1994).^{438}

Recognising the variation arising from multiple underlying aetiologies and that associated with different biomarkers (neither sources being readily quantifiable), we set out to recruit 340 patients, with between 10-20 samples over six months to ensure that any such variation was accounted for.
The frequency of sampling and the numbers proposed should provide adequate power, recognising the variation arising from multiple underlying aetiologies and that associated with different markers (neither sources being readily quantifiable).

Liver Disease

Samples from patients with liver disease were obtained as part of the main ELUCIDATE trial (Chapters 16-23). From the initial 878 randomised patients 847 consented to take part in the translational research aspect of the trial and provided a single blood sample in a plain clot serum activator tube at baseline. The serum was then processed according to the ELUCIDATE Site Study Operating Procedure (SSOP), stored locally at -80°C, and then shipped on dry ice to the Research Tissue Bank (RTB) in Leeds for storage. Patients were then followed up and data collected as specified within the ELUCIDATE protocol. A detailed summary of this cohort is presented in Table 80.

Study Management

Investigator authorisation

Patients were recruited once all appropriate authorisations being granted for any centre and collection of the appropriate regulatory paper work. This included but was not limited to:

- Investigator contact details, including contact details for research nurse;
- Up-to-date, signed and dated, curriculum vitae (CV) for each individual on the signature and responsibilities log, including dates of Good Clinical Practice (GCP) training;
- Written confirmation of local Trust Research and Development (R&D) approval for the study.
- Copy of the SSI, signed by the Principal Investigator (PI).
- Signed PI Declaration
- Authorised signature logs
Recruitment

Patients were recruited from UK centres with slight overlap between our selected disease areas. For example two of the eleven centres recruiting patients for the RCC study were also part of the ten centres recruiting patients for the RT study. Research centres were required to have completed a feasibility assessment, obtained local NHS management approvals and undertaken a site initiation meeting with the Study/Trial Management team prior to the start of recruitment. Leeds CTRU were responsible for the monitoring of patient recruitment and consent, and also for coordination and storage of clinical data and logging of sample collection information whilst details of actual samples received were maintained within our local Laboratory Information Management System (LIMS) SENTRY.

Patient Consent

Patients at each site were approached by clinicians or research nurses trained in taking informed consent with study-specific Patient Information Sheets regarding the study and consent forms. In all cases the information in the Patient Information Sheet was explained fully to patients and they were given the opportunity to ask any questions. In the Renal Transplant Study, some centres sent patients a letter with the Patient Information Sheet, informing them of the research study and giving them time to consider their participation, prior to coming in to clinic. The right of the patient to refuse consent without giving reasons was respected. Wherever possible informed consent was obtained at a hospital visit prior to when the sample were obtained. However due to the patterns of referral of many of these patients this was not possible in many instances. In such cases, full explanations were provided and the patient given as much time as they needed to make a decision. Healthy participants were also recruited at each centre by approaching visitors and hospital staff.

Broad consent was sought to store and use patient data and samples for this project and also for subsequent unspecified studies given the intended storage within an RTB, if residual tissues and fluids remain, with examples of the types of study provided in the information sheet. Patients were specifically asked to opt in or out, using a multilevel consenting procedure for more sensitive aspects of the research including; whole genome sequencing.
studies; permission to contact GP regarding clinically relevant findings; permission to contact relatives regarding clinically relevant findings; and studies conducted by other groups and commercial partners,

No specific additional risks were involved for the patients participating in this study as wherever possible blood samples were obtained at the time of routine venepuncture and no surgical or treatment procedures other than those planned as part of their standard treatment were undertaken for this project.

Should a patient have required a translation of the study documentation, it was the responsibility of the individual Investigator to translate the Patient Information leaflet and the Consent form, using locally approved translators. The consent form must then be appropriately signed and dated by the patient, the investigator and the translator.

Original copies of the consent form were retained in the Investigator Site File; a copy of the consent form was given to the patient, a second copy filed in the patient’s healthcare records (as per local practice), and a third copy returned to the Leeds CTRU.

All patients were free to withdraw from the study at any time at his/her own request, without prejudice; or may be withdrawn from the study at any time at the discretion of the Investigator. Unused samples and data would, after the notice of withdrawal, be disposed of securely and respectfully. Part discontinuation or withdrawal involved using the subject’s samples already obtained up to the point of withdrawal. Full discontinuation or withdrawal involved having the subject’s samples destroyed. This required breaking the study code so that anonymised samples could be identified. In the event that analysis had already been performed on the sample(s), requests to destroy molecular data could not be honoured because the associated data may be required for audit purposes by a regulatory authority. Additionally, the sample may have been pooled with other samples such that it would not be possible to isolate a particular sample from the pool, and the pool may continue to be analysed until it is used up. A standard letter for withdrawing consent was included alongside the patient information sheet at recruitment.
Patient registration

Once a patient was confirmed as being eligible for the study and had given written informed consent, the investigator or designee completed the supplied registration case report form (CRF) and contacted the Leeds CTRU’s 24 hour registration telephone line. The Leeds CTRU recorded basic patient details (date of birth, initials, and confirmation of consent), and then allocated a unique trial number to the patient; this trial number was then recorded in the patient’s medical notes as well as on all study CRFs. Confirmation of the patient’s registration was then faxed/emailed back to the participating site, and the original registration and eligibility form was sent to the CTRU via post.

Data collection and storage

Data collection was managed by the Leeds CTRU. Relevant clinical and demographic data was collected, by research nurses and clinicians, on standardised case report forms (CRFs) appropriate to the participant group. These were then copied and the originals submitted to the Clinical Trials Research Unit (CTRU) and held securely in paper and electronic form. Participating sites were expected to maintain a file of essential study documentation, and to keep copies of all completed forms for at least the duration of the study. All information collected during the course of the trial was kept strictly confidential and the CTRU ensured compliance with all aspects of the 1998 Data Protection Act. On completion of the study, data will eventually be transferred to the MRC Informatics Centre in Leeds and will be held securely in compliance with all aspects of the 1998 Data Protection Act for a minimum of 10 years. Site files will be archived by the participating NHS Trusts for 10 years and arrangements for confidential destruction will then be made.

Sample collection and storage

Blood and urine samples were processed promptly at each centre by either the research nurse or staff of the Clinical Chemistry labs according to SOPs provided (Appendix 1) and frozen in multiple aliquots at -80°C until shipped. Details about the time of sampling and processing was recorded. Tissue samples (FFPE blocks and frozen tissue at LTHT only) were processed
by the staff in the Pathology department at each centre as part of the routine tissue processing procedures.

All samples (tissue and biological fluids) were shipped by courier at regular intervals and stored at an HTA licensed establishment (University of Leeds/Leeds Teaching Hospitals NHS Trust – 12279) in compliance with the licensing provisions of the Human Tissue Act (2004). Each set of fluid samples were mirror banked between two secure monitored/logged -80°C freezers linked to a central alarm system to alert to temperature fluctuations, located in secure rooms. Frozen tissue (from Leeds only) was stored securely in the bank’s liquid nitrogen dewars. FFPE tissue blocks were stored in a secure filing system within the banking facility. In all cases samples were pseudonymised and assigned unique storage numbers. Their locations have been logged on a secure Laboratory Information Management system (SENTRY) in Leeds and will also eventually be mirrored also on Medical Achiever being established by the Trust/University.

**Monitoring and quality assurance**

All returned Consent forms and CRFs were checked for compliance with the study protocol, inconsistent data, missing data and timing. Study staff were in regular contact with study centre personnel to check on progress and deal with any queries that they may have had. In the event of unclear data, the Leeds CTRU issued a manual query form, giving details of which information was missing or unclear. Responses to queries were made on the query forms, which were copied and returned. In addition data cleaning and checking was undertaken by researchers prior to embarking on research studies.

A Study Steering Committee (SSC) met every 6 months during the recruitment phase and annually (meeting or telecom) during the follow-up phase and was responsible for the overall supervision of the study.

**Study Management Results and Discussion of Issues**
Consent

Within this programme a multi-level approach to consent was adopted, where patients were given a choice concerning whether or not they wished to participate in certain more sensitive aspects of research. This was considered to be the most appropriate method of balancing the wishes of patients, whilst ensuring use of their samples and data in a wide a range of research projects.

Refusal to consent to the optional elements was low (<10%), with the exception of 18% of patients in the RCC cohort not wishing to consent to their relatives being contacted regarding any health findings, if they themselves could not be contacted. This concern amongst patients reflects similar ongoing discussions in the ethical field concerning the implications of genetic testing on families, and balancing the obligation for health care professionals to disclose, with an individuals right to know their relative’s disease status, or not.\textsuperscript{439}

The number of consenting issues was quite high, with issues in 79 of 706 (11.19%) RCC forms and 41 of 514 (7.98%) Renal Transplant forms. The vast majority of these issues were around the optional consent elements not being appropriately deleted. The cost and time associated with chasing, correcting and administering the optional consent elements were considered by the trials unit to be more than for a typical trial.

Taken together these findings provide some evidence to support the case for a broad, as opposed to multi-level, consent process, as the number of patients who would not have participated would have been low. These conclusions are supported by a recent workshop funded by the US National Institute for Health.\textsuperscript{440} This reported that broad consent is acceptable and pragmatic as long as participants are provided with sufficient information to make a reasonably informed decision and that sufficient processes are in place to provide independent oversight and approval of future research. They argue that broad consent can protect the rights of donors, whilst minimising the cost and administrative burden on researchers.
Where multi-level consent is required careful attention should be paid to the design of the consent form. With hindsight, the delete options used in these forms were not clear enough but were what was mandated locally at that time. This was also very much an emerging area with the use of RTBs and also larger scale genomic analysis becoming more widely used. The Leeds Multidisciplinary RTB form which provided the template for the NIHR Programme RTB has since been adapted based on patient and nurse feedback and is now presented much more clearly and with lower error rates.

Sample Processing

Technical pre-analytical factors, including specimen collection, processing, transport and storage are a major source of measurement uncertainty in biomarker discovery, clinical trials and clinical laboratory medicine. In order to minimise the effects of pre-analytical technical factors on samples, a multi-centre localised approach was established, as opposed to a centralised processing strategy (e.g. UK Biobank). Localized strategies have several benefits in comparison to centralised strategies, including short processing and cryostorage times, leading to better preservation of biomarkers, and batch shipments, reducing the overall cost of transportation. However, localised strategies require appropriate facilities, equipment and staff at each centre, increasing the burden of training and quality control.

To ensure consistency across all centres: standardised sample packs (see Figures 29 and 30) were prepared centrally and shipped out to centres, all sites underwent an initiation and research staff were trained in sample processing and handling procedures. Study Site Operating Procedures (SSOPs) were carefully prepared for all sample collections taking into account our previous experience in Leeds in processing samples and developing protocols for biomarker discovery studies and published guidance, where appropriate. However, for some parameters, best practice guidelines were not deemed to be pragmatic for “real world” sample collection in the NHS and also some analytes are not optimally processed under such conditions. However, the most important aspect is consistency and recording conditions so that this can be accounted for in terms of suitability for specific analytes. Table 71 outlines some of the key pre-analytical parameters and highlights differences between examples of published standards and guidance and the SOPs. Whilst the IARC guidance has been
developed specifically for biobanking, the CLSI procedures have been developed for use in the context of routine clinical chemistry analysis. 443-445

Compliance data was collected on key parameters, in-particular sample processing times, summaries of this data are presented in Table 72, along with the percentage of samples meeting the SSOP specification. Compliance within the RCC cohort was high, with over 94% of samples centrifuged within two hours and over 89% frozen within two hours. Compliance with the renal transplant cohort was lower, especially for bloods, with only 48% of patient samples being processed under two hours. The primary reason for this became apparent during study set up at several centres, where most in-patient routine blood samples are collected in the very early morning (~6am), but most of the local sample processing staff did not begin work until 8am. Therefore a pragmatic decision was made to process these blood samples in the shortest time frame possible. Compliance with urine samples was better as these were generally collected fresh by the research nurses.
Figure 29 ELF sample and Liver Biobank sample kit. From top left clockwise: Safebox® (Royal Mail) for shipping ELF sample; ELF Test Sample shipping form; 10x 0.5mL liver biobank tubes (Sarstedt); biobank blood tube label (CILS International); pastettes (Scientific Laboratory Supplies Ltd); 7 mL bijou (Scientific Laboratory Supplies Ltd); Sample form; ELF Test blood tube label (CILS International); ELF Test sample tube (Nunc).
Figure 30 Renal Transplant and Renal Cell Carcinoma sample kit. From top left clockwise: Sample tube kit (FluidX Ltd); pastettes (Scientific Laboratory Supplies Ltd); Sample from; Sample tube caps (FluidX Ltd); 150 mL urine collection pot (Scientific Laboratory Supplies Ltd); 50mL centrifuge tube; 20mL barcoded universal (Scientific Laboratory Supplies Ltd); 7 mL Bijou (x2) (Scientific Laboratory Supplies Ltd).
Table 71 Sample processing SOPs compared to examples of published standards and guidelines.

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Pre-analytical factor</th>
<th>SOPs</th>
<th>IARC (H18-A4)</th>
<th>CLSI (GP16-A3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urine</td>
<td>transport temperature</td>
<td>room temp</td>
<td>ambient</td>
<td>room temp (if &lt;2 hours)</td>
</tr>
<tr>
<td>Urine</td>
<td>time to processing</td>
<td>&lt;2 hours</td>
<td>minimum</td>
<td>&lt;2 hours</td>
</tr>
<tr>
<td>Urine</td>
<td>centrifugation speed</td>
<td>2000g</td>
<td>not specified</td>
<td>not specified</td>
</tr>
<tr>
<td>Urine</td>
<td>centrifugation time</td>
<td>10 mins</td>
<td>not specified</td>
<td>not specified</td>
</tr>
<tr>
<td>Urine</td>
<td>storage temperature</td>
<td>&lt;70C</td>
<td>-80C</td>
<td>not specified</td>
</tr>
<tr>
<td>Serum</td>
<td>tube type</td>
<td>Plain Clot Activator</td>
<td>without anticoagulant</td>
<td>clot activator</td>
</tr>
<tr>
<td>Serum</td>
<td>tube volume</td>
<td>8-10mL</td>
<td>not specified</td>
<td>not specified</td>
</tr>
<tr>
<td>Serum</td>
<td>transport temperature</td>
<td>room temp</td>
<td>room temp</td>
<td>room temp</td>
</tr>
<tr>
<td>Serum</td>
<td>collection procedure</td>
<td>invert 5x</td>
<td>not specified</td>
<td>invert 5-10x</td>
</tr>
<tr>
<td>Serum</td>
<td>clotting time</td>
<td>&gt;45 mins</td>
<td>&gt;30 mins</td>
<td>5-30 minutes</td>
</tr>
<tr>
<td>Serum</td>
<td>venepuncture to storage</td>
<td>&lt;2 hours</td>
<td>&lt;1 hour</td>
<td>&lt;2 hours</td>
</tr>
<tr>
<td>Serum</td>
<td>centrifugation speed</td>
<td>2000xg</td>
<td>1500xg</td>
<td>not specified</td>
</tr>
<tr>
<td>Serum</td>
<td>centrifugation time</td>
<td>10 mins</td>
<td>10mins</td>
<td>not specified</td>
</tr>
<tr>
<td>Serum</td>
<td>centrifugation temperature</td>
<td>room temp</td>
<td>room temp</td>
<td>20-22C</td>
</tr>
<tr>
<td></td>
<td>Serum storage temperature</td>
<td>Plasma tube type</td>
<td>Plasma transport temperature</td>
<td>Plasma collection procedure</td>
</tr>
<tr>
<td>----------------</td>
<td>--------------------------</td>
<td>------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td></td>
<td>&lt;-70C</td>
<td>EDTA</td>
<td>room temp</td>
<td>invert 5x</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EDTA</td>
<td>room temp</td>
<td>not specified</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>20-22C</td>
<td>invert 5-10x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>20-22C</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 72 Sample processing times and percentage compliance with SSOP

<table>
<thead>
<tr>
<th></th>
<th>Blood</th>
<th>Urine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median time to centrifugation in hours</td>
<td>% within 45 - 120 mins</td>
</tr>
<tr>
<td>RCC</td>
<td>01:11</td>
<td>98%</td>
</tr>
<tr>
<td>Renal transplant</td>
<td>01:24</td>
<td>78%</td>
</tr>
</tbody>
</table>
An inspection of some biomarker results highlighted that a single centre was using inappropriate blood tubes for the collection of serum samples and this is described in detail in Chapter 14. In this instance the confusion stemmed from the tube manufacturer (Greiner) making both a red top serum tube and a red top EDTA plasma tube. In both the SOPs and the on site training, careful attention was paid to ensure that sites selected the appropriate tube type for the collection of serum and plasma. At the time of CRF development it was thought to be overly burdensome to request sites to record the catalogue and lot numbers of each blood collection tube used, instead researchers selected the manufacturer from a tick list. The lack of standardised colour coding of blood collection tubes has been previously highlighted as a patient safety issue and the European Federation for Clinical Laboratory Medicine has called for harmonisation. Our experiences support this view that the current heterogeneity in tube colours presents a significant pre-analytical risk for multicentre biomarker studies and trials and until blood collection tubes are harmonised researchers must remain vigilant. Unfortunately, it is difficult to standardise research blood collection tubes by manufacturer within multicentre studies, as the tubes must be compatible with the local venepuncture system. Patient preference and the terms of ethical approval, usually dictate that research bloods are collected within the same cycle of venepuncture, using the same venepuncture apparatus as routine samples.
Clinical Data

Compliance to the study protocol and schedule were generally high across both RCC and Renal Transplant Cohorts as demonstrated in Tables 73 and 74 by the >90% and >85% overall CRF compliance for the respective studies. However, in both cohorts, sampling fatigue and/or loss to follow up was observed. This is more apparent in the RCC cohort than the transplant cohort, with only 48% of RCC patients in the longitudinal cohort providing their final sample (at 2 years), versus 74% of transplant patients (at 6 months). This is potentially due to both differences in the length of follow-up and whether follow-up occurs at the same hospital as recruited, and also the higher frequency of hospital visits required by transplant patients. However, as follow-up is still ongoing, compliance may change in time.

Table 73 RCC CRF compliance

<table>
<thead>
<tr>
<th>CRF</th>
<th>CRFs Received</th>
<th>CRFs Due</th>
<th>CRF return rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F01 Eligibility Form</td>
<td>705</td>
<td>706</td>
<td>99.86%</td>
</tr>
<tr>
<td>F02 Baseline Assessment</td>
<td>705</td>
<td>706</td>
<td>99.86%</td>
</tr>
<tr>
<td>F03 Surgery Details</td>
<td>697</td>
<td>706</td>
<td>98.73%</td>
</tr>
<tr>
<td>F04 Sample Form (Initial)</td>
<td>699</td>
<td>706</td>
<td>99.01%</td>
</tr>
<tr>
<td>F04 Sample Form (3-6mth)</td>
<td>139</td>
<td>169</td>
<td>82.25%</td>
</tr>
<tr>
<td>F04 Sample Form (12mth)</td>
<td>107</td>
<td>148</td>
<td>72.30%</td>
</tr>
<tr>
<td>F04 Sample Form (18mth)</td>
<td>73</td>
<td>148</td>
<td>49.32%</td>
</tr>
<tr>
<td>F04 Sample Form (24mth)</td>
<td>70</td>
<td>147</td>
<td>47.62%</td>
</tr>
<tr>
<td>F06 Follow up 1 year</td>
<td>624</td>
<td>629</td>
<td>99.21%</td>
</tr>
<tr>
<td>F06 Follow up 2 years</td>
<td>359</td>
<td>468</td>
<td>76.71%</td>
</tr>
<tr>
<td>F06 Follow up 3 years</td>
<td>142</td>
<td>200</td>
<td>71.00%</td>
</tr>
<tr>
<td>F06 Follow up 4 years</td>
<td>15</td>
<td>47</td>
<td>31.91%</td>
</tr>
<tr>
<td>F06 Follow up 5 years</td>
<td>0</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>Overall compliance</td>
<td>4335</td>
<td>4780</td>
<td>90.69%</td>
</tr>
</tbody>
</table>

Table 74 Renal Transplant CRF Compliance
<table>
<thead>
<tr>
<th>CRF</th>
<th>CRFs Received</th>
<th>CRFs Due</th>
<th>CRF return rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F01 Baseline Assessment</td>
<td>419</td>
<td>433</td>
<td>96.8%</td>
</tr>
<tr>
<td>F01 RT Eligibility and Registration</td>
<td>512</td>
<td>512</td>
<td>100.0%</td>
</tr>
<tr>
<td>F06 RT Sample Form (Pre Transplant)</td>
<td>442</td>
<td>512</td>
<td>86.3%</td>
</tr>
<tr>
<td>F03 RT Post-Op Investigations</td>
<td>291</td>
<td>297</td>
<td>98.0%</td>
</tr>
<tr>
<td>F06 RT Sample Form (Hospital Stay)</td>
<td>227</td>
<td>298</td>
<td>76.2%</td>
</tr>
<tr>
<td>F04 RT Intra Post Op</td>
<td>290</td>
<td>296</td>
<td>98.0%</td>
</tr>
<tr>
<td>F05 RT Follow Up week 1</td>
<td>272</td>
<td>292</td>
<td>93.2%</td>
</tr>
<tr>
<td>F06 RT Sample Form week 1</td>
<td>252</td>
<td>292</td>
<td>86.3%</td>
</tr>
<tr>
<td>F05 RT Follow Up week 2</td>
<td>272</td>
<td>289</td>
<td>94.1%</td>
</tr>
<tr>
<td>F06 RT Sample Form week 2</td>
<td>244</td>
<td>289</td>
<td>84.4%</td>
</tr>
<tr>
<td>F05 RT Follow Up week 3</td>
<td>266</td>
<td>289</td>
<td>92.0%</td>
</tr>
<tr>
<td>F06 RT Sample Form week 3</td>
<td>244</td>
<td>289</td>
<td>84.4%</td>
</tr>
<tr>
<td>F05 RT Follow Up week 4</td>
<td>268</td>
<td>288</td>
<td>93.1%</td>
</tr>
<tr>
<td>F06 RT Sample Form week 4</td>
<td>240</td>
<td>288</td>
<td>83.3%</td>
</tr>
<tr>
<td>F05 RT Follow Up month 2</td>
<td>251</td>
<td>284</td>
<td>88.4%</td>
</tr>
<tr>
<td>F06 RT Sample Form month 2</td>
<td>223</td>
<td>284</td>
<td>78.5%</td>
</tr>
<tr>
<td>F05 RT Follow Up month 3</td>
<td>232</td>
<td>280</td>
<td>82.9%</td>
</tr>
<tr>
<td>F06 RT Sample Form month 3</td>
<td>199</td>
<td>280</td>
<td>71.1%</td>
</tr>
<tr>
<td>F05 RT Follow Up month 6</td>
<td>219</td>
<td>266</td>
<td>82.3%</td>
</tr>
<tr>
<td>F06 RT Sample Form month 6</td>
<td>197</td>
<td>266</td>
<td>74.1%</td>
</tr>
<tr>
<td>F05 RT Follow Up year 1</td>
<td>159</td>
<td>203</td>
<td>78.3%</td>
</tr>
<tr>
<td>F05 RT Follow Up year 2</td>
<td>71</td>
<td>97</td>
<td>73.2%</td>
</tr>
<tr>
<td>F05 RT Follow Up year 3</td>
<td>13</td>
<td>15</td>
<td>86.7%</td>
</tr>
<tr>
<td><strong>Overall Compliance</strong></td>
<td><strong>5803</strong></td>
<td><strong>6639</strong></td>
<td><strong>87.4%</strong></td>
</tr>
</tbody>
</table>
Recruitment to time and target

The recruitment rates to the RCC and Renal Transplant cohorts are shown in Figures 31 and 32 and 33 and 34 respectively. The accrual graphs clearly show that recruitment was slower than expected across both studies, with RCC aiming to have completed by December 2012 and Renal Transplant by September 2013. This was in part due to the length of time taken to set up individual centres (see Tables 78 and 79) and slower than predicted monthly recruitment. However, 5 of 11 RCC centres and 8 of 10 Renal Transplant Centres managed to achieve or exceed their target number of patients by the end of the study period, as shown in Figures 32 and 34. Over recruitment was encouraged and two RCC and six transplant centres over recruited by more than 50% of their original target figure. However, we observed that there were limited incentives within the system to encourage over recruitment and upwards revision of target recruitment figures. One centre in particular excelled at recruiting patients and hit their recruitment target within 6 months of opening, but declined to increase their target.
Figure 31 Monthly and cumulative accrual of patients with suspected RCC. Figures are shown for both the longitudinal and...
cross-sectional cohorts.

Figure 32 RCC targets and accrual by centre
Figure 33 Renal Transplant: monthly and cumulative recruitment from the waiting list and transplantation
Figure 34 Renal Transplant targets and accrual by centre
Study Governance Aspects

Ethical considerations

The studies were all performed in accordance with the recommendations for biomedical research involving human subjects adopted by the 18th World Medical Assembly, Helsinki, Finland, 1964, and subsequent amendments. Permission was sought from each NHS organization and national/local principles of Research Governance were adhered to. The possibility of samples being used in collaboration/partnerships with commercial companies has been made explicitly clear in the patient information sheet and consent form. All procedures and processes will be compliant with the Human Tissue Act 2004 and the management of tissues and fluids stored centrally in Leeds will be the responsibility of the RTB Management Committee.

Establishment of a Research Tissue Bank

To establish the renal transplant and renal cancer cohorts, two separate approaches to gaining ethical approval were available at the time of set up:

i) Project-specific ethical approval

ii) Research Tissue Bank “generic ethical approval”

Traditionally, researchers would apply to a Research Ethics Committee (REC) for project specific ethical approval, which would enable them to use the material collected in the study for the duration and purpose of the study. However, if these materials had value beyond the duration of the initial ethical approval, researchers would need to do one of the following before the end of the study period:

i) Apply for a renewal of project-specific approval

ii) Obtain a HTA licence and set up a Research Tissue Bank

iii) Transfer samples to a research tissue bank

Therefore, rather than applying for multiple project specific approvals and setting up an RTB at the end, we opted to set up an RTB from the outset and to seek generic ethical approval for the collection and use of patient data and samples in the current and future projects.

Ethical approval was sought to establish the Leeds NIHR Biomarker Research Tissue Bank (ethical approval 10/H1306/6) and a Management Committee appointed to oversee the collection, storage and release of material for research (chaired by Profs Banks and Selby,
and including the local HTA Designated Individual, Quality Assurance Manager, RTB Manager, nurse representative, patient representative, clinical representatives and others). This committee monitors and reviews the bank and related activities and considers applications for access and proposed use of the samples by other research groups once the samples have been used to meet the remit of this programme. Priority is given to collaborative applications and groups need to provide a clear and well defined research plan according to the criteria laid out in the application process. Additionally, research groups have to complete a Material Transfer Agreement (MTA) and demonstrate provision of correct storage, logging and records of storage as required. Groups who are granted permission to access samples from the bank are expected to meet the costs of having the samples shipped under appropriate conditions to their establishment. Samples will only be provided if appropriate consent has been given by the patient i.e. a sample would not be released for a project involving genetic analyses if the patient has not given their consent for their sample to be used in such studies. The bank is promoted to researchers and patients through the Joint Multidisciplinary Research Tissue Bank web site (www.multirtb.leeds.ac.uk).

The Human Tissue Act 2004 (HTA) requires that the storage of relevant material which essentially is cellular material for all ‘research in connection with disorders, or the functioning, of the human body’ is licenced by the Human Tissue Authority. NHS Research Ethics Committees can give generic ethical approval for an RTB to collect store and release tissues for use in research, providing the bank is on a HTA-licensed premises. This is advantageous in terms of maximising use of samples for future research, ensuring consistent and high level governance procedures and minimising administrative burden.

Local and National NHS approvals

The Renal Cell Carcinoma and Renal Transplant studies both utilised the generic ethical approval of the RTB (REC 10/H1306/6) from the start, for recruiting patients and collecting samples. The liver disease serum samples were collected under the project specific approval of the ELUCIDATE trial (10/H1313/2) but sought broad patient consent for use of the samples in future biomarker research. The Liver samples and data were then subsumed under the governance of the RTB following the completion of recruitment and sample collection.
Governance Results and Discussion of Issues

Research RTBs are purported to have several advantages over project specific approvals, notably the ability to collect, store and use samples for a wide range of applications without seeking specific ethical approval for individual projects, and no requirement for local R&D approvals for Tissue Collection Centres (TCCs) (Table 75).

Table 75 Pros and Cons of Research Tissue Banks vs Project-Specific Approval

<table>
<thead>
<tr>
<th></th>
<th>Project-Specific Approval</th>
<th>Research Tissue Banks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portfolio Adoption</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Human Tissue Act applies</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Usage of samples</td>
<td>Within scope of original project-specific approval</td>
<td>Within scope of original RTB approval – but requires approval of RTB committee</td>
</tr>
<tr>
<td>Local NHS R&amp;D approval</td>
<td>Required</td>
<td>Not required</td>
</tr>
</tbody>
</table>

However, the reality was more complicated, presenting numerous challenges, including:

NIHR CRN Portfolio Adoption

Adoption onto the NIHR CRN Portfolio is a fundamental enabler to run successful multicentre studies in the NHS. Portfolio adoption enables access to NHS service support, without which most recruiting centres would not have participated.

However, the guidelines for portfolio eligibility ([https://www.crn.nihr.ac.uk/can-help/funders-academics/nihrcrn-portfolio/which-studies-are-eligible-for-clinical-research-network-support/](https://www.crn.nihr.ac.uk/can-help/funders-academics/nihrcrn-portfolio/which-studies-are-eligible-for-clinical-research-network-support/)) specifically exclude “banking of biological samples or data except where this activity is integral to a self-contained research project designed to test a clear hypothesis”. This meant that the RTB itself was deemed to be ineligible, therefore individual studies (RCC and RT) were submitted to the CRN portfolio, using the ethical approval of the RTB.
NHS Permissions

Although the individual projects were eventually adopted onto the portfolio, the creation of projects meant that local NHS approval was required for every centre participating in the studies. One of the advantages of RTBs is that Tissue Collection Centres do not need local approval. Furthermore, the unfamiliarity of local teams with RTBs and the complexity of the project approval process was a source of much confusion and delay at almost every NHS centre (Tables 78 and 79).

Human Tissue Act (HTA) Relevant Material

Tissue samples or buffy coats or any cellular material collected under a project-specific REC approval are not bound by the governance of the HTA for the duration of the project. However, such samples (“Relevant Material”) collected within an RTB infrastructure are governed by the HTA. On the whole, HTA governance did not impact greatly on our procedures, as they are well aligned with best practice and GCP. However, one aspect mandated by HTA i.e. “Storage incidental to transportation” was a significant cause for concern and incurred significant expense in meeting. The HTA defines storage as incidental to transportation if “tissue is held for a matter of hours or days (but never weeks) pending transfer to a licensed establishment.”. As buffy coat samples for example (serum, plasma and urine are exempt providing acellular) were to be stored on site and shipped in batches to save money (£150 for 1 sample vs £400 for 100) the time spent on local sites will have exceeded the time HTA consider to be acceptable for storage incidental to transportation. For centres that had licenced premises this meant that all HTA relevant materials needed to be transferred from the local clinic to the licenced premises within 7 days, incurring additional risk, time and cost. For centres that did not have licenced premises, this meant shipping all samples back to Leeds within 7 days of collection, significantly increasing the shipping costs and the amount of administration.

NHS Service Support Costs

Prior to the introduction of the “Attributing the costs of health & social care Research & Development (AcoRD)” guidelines in 2012 there was ambiguity as to what classified as a research or NHS service support cost. Also the reimbursement costs of biomarker studies as opposed to cTIMPs were at a much lower level and didn’t vary whether single sample and
little information collected or multiple samples and multiple CRFs. In 2009 when this programme was funded, many of the clinical research costs incurred by sites, including CRF completion for example, were assumed to be covered by the NHS service support costs and were therefore not funded separately by NIHR. For the sample and CRF-intensive renal transplant cohort this meant that there was insufficient reimbursement to incentivise centres to participate. Following discussions with the CRN Renal Speciality Group a substantial amendment was produced that halved the number of samples and pages of CRFs to those included here and used, making the study more attractive to potential centres. The impact of this amendment can be observed in the boost to recruitment following March 2013 in Figure 33. However, sampling fatigue was an ongoing issue (refer to Tables 73 and 74), suggesting that further incentives may help to improve compliance.

Local and National NHS Approvals

The process of gaining ethical approval for the RCC and Renal Transplant studies, through the establishment of a research tissue bank (RTB) took 5 months from research ethics committee (REC) submission (15/01/2010) to REC approval being granted (15/06/2010). This was in part due to the committee requesting further information and some revision of documentation on 08/02/2010 (response 03/03/2010) and again on 01/04/2010 (response 15/04/2010). All three studies gained adoption onto the NIHR Clinical Research Network (CRN) portfolio through the Coordinated System for gaining NHS Permission (CSP), ensuring access to local service support. The timescales for gaining national approval for each study are shown in Table 76. In April 2010 the NIHR introduced the CRN High Level Objectives (http://www.nihr.ac.uk/documents/policy-and-standards/Faster-easier-clinical-research/NIHR-Metrics-Comparison-CCF-December-2013.pdf), which aimed to increase the proportion of studies in the CRN portfolio that deliver to their planned recruitment time and targets, reduce the time taken to achieve NHS permission; and reduce the time taken to recruit first participants. These include a 40 day target for gaining NHS permission and a 30 day target for recruitment of first patient following permission. These targets were not achieved for both the RCC and Renal Transplant cohorts. In total, the RCC and Renal Transplant cohorts took 6 and 12 months respectively, from CRN portfolio submission to recruiting the 1st patient.

Adoption onto the CRN portfolio was a challenge, with both the RTB status and also queries
concerning peer review causing delays. Delays in study-wide national sign off were mainly due to these being some of the first RTB studies to come through the CLRN, who were, at the time, unfamiliar with the process. The concept of having ethical approval for the Leeds NIHR Biomarker RTB but not for the specific studies caused confusion, as did having protocols for the specific studies that sat alongside the RTB protocol. At a local permission level, NHS Trust R&D offices were also unfamiliar with the RTB ethical approval status, asking questions around the need for study-specific ethical approval. It is worth noting that towards the end of the programme, some centres sent letters confirming that local R&D approval was not required due to the RTB status. RTB status also introduced another source of delay, as Material Transfer Agreements (MTAs) had to be signed off by the trust’s Designated Individual for research, as they are responsible under the Human Tissue Act for any relevant materials. These individuals were often unknown to R&D Departments, or were not in post, were difficult to contact or simply refused to sign the agreements. Transfer of patient data at several centres was also a cause of delay, despite using the standard procedures of an accredited CTRU.

Participating NHS hospitals were identified initially by the chief investigators, but later through word of mouth and the invaluable support of CRN cancer, liver and renal clinical specialty groups. Tables 77 and 78 show the time taken by individual centres to submit their site-specific information (SSI) form (research application), gain local R&D approval, open the centre following R&D approval and recruit the 1st patient. The main source of local delay in the RCC study set up was the R&D approval process, taking an average of 3.8 months (range 0.0-7.0), whilst the main source of local delay for the Renal Transplant study was the submission of the SSI form, taking an average of 4.2 months (range 0.6-7.8). This difference in R&D approval timing may be a reflection of the introduction of the “NIHR Performance in Initiating and Delivering Clinical Research exercise” in the autumn of 2011. This initiative set a 70 day benchmark for local NHS providers to recruit their first patient, following receipt of a valid research application. The two studies span this transition period, with the majority of RCC centres (73%) starting in 2011 and the majority of Renal Transplant centres (90%) starting in 2012/13. None of the RCC centres achieved the 70 day benchmark (Mean 218; Range 76-369 days), whilst 40% of Renal Transplant centres did (Mean 115; range 26-287 days). By combining the two cohorts together a downwards trend in R&D approval times over time is observed, Figure 35. This supports the CRNs own analysis, which reports that in 2014-15 83 percent of studies are now receiving NHS permission within 40 days at all sites.
It was queried during the study period whether this reduction in local NHS permissions might be in part due to “gaming”, where centres would withhold submitting their SSI forms until the end of the process, in order to improve their metrics. This hypothesis is supported to some extent in the doubling of mean SSI submission times between the cohorts (2.3 vs 4.2) and the trend towards increased SSI submission times as a percentage of local review and approval times, Figure 36. However, Figure 37 does suggest that there might be a trend towards shorter overall set up times over time, suggesting that improvements in set up time may be being achieved and that this initiative may be having the desired impact on pushing the pace of UK clinical research. It should be noted however, that although of interest, these observations are inconclusive and could be due to a multitude of other factors, such as different centre makeups in the comparisons of the cohorts, differences in the complexity of the study set up for the cohorts etc. Further evaluation at a national level should be conducted, taking into account the time from transfer of SSI forms to submission, data for which are logged within IRAS.

Table 76 Timescales for gaining national approval

<table>
<thead>
<tr>
<th>CRN portfolio approval (months)</th>
<th>NHS R&amp;D approval (months)</th>
<th>Time from R&amp;D approval to first patient (months)</th>
<th>Time from first to last patient recruited (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCC Study 06/01/2011-24/02/2011 (1.6)</td>
<td>02/02/2011-01/04/2011 (1.9)</td>
<td>01/04/2011 - 14/07/2011 (3.4)</td>
<td>14/07/2011-30/06/2014 (36)</td>
</tr>
<tr>
<td>Renal Transplant Study 19/04/2011-12/03/2012 (10.8)</td>
<td>16/12/2011-05/03/2012 (2.6)</td>
<td>05/03/2012 - 26/04/2012 (1.7)</td>
<td>26/04/2012-30/04/15 (37)</td>
</tr>
</tbody>
</table>

*date of last transplant was 24/9/15
Table 77 RCC study timescales for gaining local centre approval. Sites have been anonymised for this purpose and therefore site numbers below are arbitrary.

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Time from transfer of SSI form to Submission (months)</th>
<th>Time from submission of SSI form to R&amp;D Approval (Months)</th>
<th>Time from R&amp;D approval to open for recruitment (months)</th>
<th>Time from open to 1st Patient (months)</th>
<th>Total (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.3</td>
<td>0.0</td>
<td>1.3</td>
<td>2.3</td>
<td>6.8</td>
</tr>
<tr>
<td>2</td>
<td>3.6</td>
<td>1.2</td>
<td>2.3</td>
<td>5.0</td>
<td>12.1</td>
</tr>
<tr>
<td>3</td>
<td>5.8</td>
<td>4.9</td>
<td>1.8</td>
<td>0.7</td>
<td>13.2</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
<td>3.3</td>
<td>2.0</td>
<td>1.7</td>
<td>7.6</td>
</tr>
<tr>
<td>5</td>
<td>0.0</td>
<td>5.8</td>
<td>0.9</td>
<td>5.4</td>
<td>12.2</td>
</tr>
<tr>
<td>6</td>
<td>7.0</td>
<td>3.9</td>
<td>1.6</td>
<td>1.8</td>
<td>14.3</td>
</tr>
<tr>
<td>7</td>
<td>0.7</td>
<td>2.5</td>
<td>1.3</td>
<td>0.2</td>
<td>4.8</td>
</tr>
<tr>
<td>8</td>
<td>1.7</td>
<td>7.0</td>
<td>3.4</td>
<td>0.5</td>
<td>12.5</td>
</tr>
<tr>
<td>9</td>
<td>1.7</td>
<td>7.0</td>
<td>3.4</td>
<td>0.5</td>
<td>12.5</td>
</tr>
<tr>
<td>9</td>
<td>0.0</td>
<td>2.3</td>
<td>2.3</td>
<td>0.5</td>
<td>5.1</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>3.5</td>
<td>0.3</td>
<td>0.1</td>
<td>4.1</td>
</tr>
<tr>
<td>Median</td>
<td>1.2</td>
<td>3.7</td>
<td>1.9</td>
<td>0.6</td>
<td>12.2</td>
</tr>
</tbody>
</table>
Table 78 Renal Transplant study timescales for gaining local centre approval. Sites have been anonymised for this purpose and therefore site numbers below are arbitrary.

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Time from transfer of SSI form to Submission (months)</th>
<th>Time from submission of SSI form to R&amp;D Approval (Months)</th>
<th>Time from R&amp;D approval to open for recruitment (months)</th>
<th>Time from open to 1st Patient (months)</th>
<th>Total (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
<td>0.7</td>
<td>3.4</td>
<td>0.3</td>
<td>4.9</td>
</tr>
<tr>
<td>2</td>
<td>7.8</td>
<td>2.4</td>
<td>0.0</td>
<td>1.7</td>
<td>11.9</td>
</tr>
<tr>
<td>3</td>
<td>2.8</td>
<td>0.1</td>
<td>1.1</td>
<td>0.6</td>
<td>4.5</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
<td>3.7</td>
<td>2.0</td>
<td>3.7</td>
<td>10.0</td>
</tr>
<tr>
<td>5</td>
<td>0.7</td>
<td>5.4</td>
<td>1.4</td>
<td>0.7</td>
<td>8.2</td>
</tr>
<tr>
<td>6</td>
<td>5.8</td>
<td>0.8</td>
<td>2.6</td>
<td>0.6</td>
<td>9.9</td>
</tr>
<tr>
<td>7</td>
<td>5.3</td>
<td>0.9</td>
<td>0.4</td>
<td>0.4</td>
<td>6.9</td>
</tr>
<tr>
<td>8</td>
<td>7.0</td>
<td>0.1</td>
<td>0.5</td>
<td>1.0</td>
<td>8.6</td>
</tr>
<tr>
<td>9</td>
<td>6.7</td>
<td>1.4</td>
<td>0.7</td>
<td>0.4</td>
<td>9.2</td>
</tr>
<tr>
<td>10</td>
<td>4.9</td>
<td>0.0</td>
<td>0.2</td>
<td>0.6</td>
<td>5.7</td>
</tr>
<tr>
<td>Median</td>
<td><strong>5.1</strong></td>
<td><strong>0.8</strong></td>
<td><strong>0.9</strong></td>
<td><strong>0.6</strong></td>
<td><strong>8.4</strong></td>
</tr>
</tbody>
</table>

Figure 35 Time taken for R&D approvals plotted over time

![Graph showing time taken for R&D approvals plotted over time](image)

\[ R^2 = 0.3657 \]
Figure 36 SSI Submission as a percentage of total time for local review and approval (Submission of SSI + R&D Approval)

Figure 37 Combined timescales for overall study set-up time plotted over time
Summary of final resource

Liver Disease cohort summary statistics

The original recruitment target for randomisation into the ELUCIDATE trial was 1000 participants, with 878 eventually recruited of whom 847 contributed a sample to the RTB. The summary statistics of the final liver biorepository cohort are shown in Table 79.

Table 79 Characteristics of liver disease patients

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Participants</td>
<td>847</td>
</tr>
<tr>
<td>Age (yrs)</td>
<td>Median (range) 55 (23-75)</td>
</tr>
<tr>
<td>Gender</td>
<td>Male 476 (56)</td>
</tr>
<tr>
<td></td>
<td>Female 371 (44)</td>
</tr>
<tr>
<td>Weight (Kg)</td>
<td>Median (range) 81.3 (38-166)</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>Median (range) 169 (131-203)</td>
</tr>
<tr>
<td>Cause of fibrosis</td>
<td>Non-alcoholic fatty liver disease 225 (27)</td>
</tr>
<tr>
<td></td>
<td>Viral liver disease 338 (40)</td>
</tr>
<tr>
<td></td>
<td>Alcoholic liver disease 59 (7)</td>
</tr>
<tr>
<td></td>
<td>Other/unknown 225 (27)</td>
</tr>
<tr>
<td>ELF Score</td>
<td>Registration 9.36 (8.4-17.35)</td>
</tr>
<tr>
<td></td>
<td>Randomisation 9.37 (7.13-17.84)</td>
</tr>
</tbody>
</table>

Renal Cell Carcinoma and healthy control cohort summary statistics

The recruitment target for the renal cell carcinoma study was 700 patients with suspected RCC (500 cross-sectional and 200 longitudinal) and 200 healthy controls. The final totals were 706 RCC patients and 149 healthy controls were eventually recruited. The summary statistics of the patients recruited suspected of having RCC are shown in Table 80 for all patients and 81 for the clear cell sub-group. The characteristics of the healthy controls are shown in Table 82.
Table 80 Characteristics of all patients recruited with suspected RCC

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>All patients (n=706)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n (%)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>447 (63)</td>
</tr>
<tr>
<td>Female</td>
<td>259 (37)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>median (range)</td>
<td>64 (29, 92)</td>
</tr>
<tr>
<td><strong>Body Mass Index</strong></td>
<td></td>
</tr>
<tr>
<td>median (range)</td>
<td>27.9 (15.6, 74.4)</td>
</tr>
<tr>
<td><strong>Symptoms</strong></td>
<td></td>
</tr>
<tr>
<td>Local</td>
<td>199 (28)</td>
</tr>
<tr>
<td>Systemic</td>
<td>103 (15)</td>
</tr>
<tr>
<td>Both</td>
<td>151 (21)</td>
</tr>
<tr>
<td>None</td>
<td>253 (36)</td>
</tr>
<tr>
<td><strong>Procedure</strong></td>
<td></td>
</tr>
<tr>
<td>Radical nephrectomy</td>
<td>465 (66)</td>
</tr>
<tr>
<td>Partial nephrectomy</td>
<td>162 (23)</td>
</tr>
<tr>
<td>Radiofrequency ablation</td>
<td>20 (3)</td>
</tr>
<tr>
<td>Cryoablation</td>
<td>17 (2)</td>
</tr>
<tr>
<td>Biopsy only</td>
<td>12 (2)</td>
</tr>
<tr>
<td>None</td>
<td>29 (4)</td>
</tr>
<tr>
<td>Missing</td>
<td>1 (&lt;1)</td>
</tr>
<tr>
<td><strong>Histological type</strong></td>
<td></td>
</tr>
<tr>
<td>Clear cell</td>
<td>481 (68)</td>
</tr>
<tr>
<td>Papillary</td>
<td>59 (8)</td>
</tr>
<tr>
<td>Chromophobe</td>
<td>46 (7)</td>
</tr>
<tr>
<td>Oncocytoma</td>
<td>27 (4)</td>
</tr>
<tr>
<td>Translocation Carcinoma</td>
<td>2 (&lt;1)</td>
</tr>
<tr>
<td>Undeclassified</td>
<td>12 (2)</td>
</tr>
<tr>
<td>Other</td>
<td>39 (5)</td>
</tr>
<tr>
<td>Unknown</td>
<td>39 (5)</td>
</tr>
<tr>
<td>Missing</td>
<td>1 (&lt;1)</td>
</tr>
<tr>
<td><strong>TNM Stage</strong></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>336 (48)</td>
</tr>
<tr>
<td>II</td>
<td>71 (10)</td>
</tr>
<tr>
<td>III</td>
<td>143 (20)</td>
</tr>
<tr>
<td>IV</td>
<td>56 (8)</td>
</tr>
<tr>
<td>Not applicable*</td>
<td>93 (13)</td>
</tr>
<tr>
<td>Characteristic</td>
<td>Clear cell patients</td>
</tr>
<tr>
<td>---------------------------------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td></td>
<td>(n=481)</td>
</tr>
<tr>
<td></td>
<td>n (%)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>321 (67)</td>
</tr>
<tr>
<td>Female</td>
<td>160 (33)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
</tr>
<tr>
<td>median (range)</td>
<td>64 (29, 92)</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td></td>
</tr>
<tr>
<td>median (range)</td>
<td>28.3 (16, 74.4)</td>
</tr>
<tr>
<td>Radical nephrectomy</td>
<td>345 (72)</td>
</tr>
<tr>
<td>Partial nephrectomy</td>
<td>103 (21)</td>
</tr>
<tr>
<td>Procedure</td>
<td></td>
</tr>
<tr>
<td>Radiofrequency ablation</td>
<td>17 (4)</td>
</tr>
<tr>
<td>Cryoablation</td>
<td>9 (2)</td>
</tr>
<tr>
<td>Biopsy only</td>
<td>7 (1)</td>
</tr>
<tr>
<td>Tumour size (mm)</td>
<td></td>
</tr>
<tr>
<td>median (range)</td>
<td>55 (11, 180)</td>
</tr>
<tr>
<td>Pathological T-stage</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>239 (50)</td>
</tr>
<tr>
<td>2</td>
<td>50 (10)</td>
</tr>
<tr>
<td>3</td>
<td>154 (32)</td>
</tr>
<tr>
<td>4</td>
<td>5 (1)</td>
</tr>
<tr>
<td>Not applicable*</td>
<td>33 (7)</td>
</tr>
<tr>
<td>Fuhrman Grade</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>13 (2)</td>
</tr>
<tr>
<td>2</td>
<td>156 (32)</td>
</tr>
<tr>
<td>3</td>
<td>231 (48)</td>
</tr>
<tr>
<td>4</td>
<td>77 (16)</td>
</tr>
<tr>
<td>Missing</td>
<td>4 (&lt;1)</td>
</tr>
<tr>
<td>Necrosis</td>
<td></td>
</tr>
<tr>
<td>Present</td>
<td>146 (30)</td>
</tr>
<tr>
<td>Absent</td>
<td>302 (63)</td>
</tr>
<tr>
<td>Missing</td>
<td>33 (7)</td>
</tr>
</tbody>
</table>

* Not applicable to patients found to have benign tumours or other benign conditions
### Table 82 Healthy Control Characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Participants</td>
<td>149</td>
</tr>
<tr>
<td>Age (yrs)</td>
<td>Median (range) 45 (21-88)</td>
</tr>
<tr>
<td>Gender</td>
<td>Male 50 (33.6%)</td>
</tr>
<tr>
<td></td>
<td>Female 99 (66.4%)</td>
</tr>
<tr>
<td>Weight (Kg)</td>
<td>Median (range) 71 (41-153)</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>Median (range) 168 (148-203)</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>Median (range) 25 (16-52)</td>
</tr>
<tr>
<td>Illnesses</td>
<td>none 113 (74.8%)</td>
</tr>
<tr>
<td></td>
<td>hypertension 22 (14.6%)</td>
</tr>
<tr>
<td></td>
<td>diabetes 6 (4.0%)</td>
</tr>
<tr>
<td></td>
<td>other 10 (6.9%)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>White 121 (81.2%)</td>
</tr>
<tr>
<td></td>
<td>Black Caribbean 3 (2.0%)</td>
</tr>
<tr>
<td></td>
<td>Asian Indian 6 (4.0%)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Count</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Black African</td>
<td>7</td>
</tr>
<tr>
<td>Chinese</td>
<td>1</td>
</tr>
<tr>
<td>Asian Pakistani</td>
<td>1</td>
</tr>
<tr>
<td>Other Asian</td>
<td>4</td>
</tr>
<tr>
<td>Asian Bangladeshi</td>
<td>1</td>
</tr>
<tr>
<td>Mixed</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Menopausal status (females)</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>66</td>
<td>66.7%</td>
</tr>
<tr>
<td>Post</td>
<td>25</td>
<td>25.3%</td>
</tr>
<tr>
<td>Peri</td>
<td>8</td>
<td>8.1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoking Status</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>16</td>
<td>10.7%</td>
</tr>
<tr>
<td>No</td>
<td>105</td>
<td>70.5%</td>
</tr>
<tr>
<td>Ex</td>
<td>25</td>
<td>16.8%</td>
</tr>
<tr>
<td>Passive</td>
<td>3</td>
<td>2.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alcohol consumption</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light</td>
<td>136</td>
<td>91.3%</td>
</tr>
<tr>
<td>Teetotal</td>
<td>11</td>
<td>7.4%</td>
</tr>
<tr>
<td>unknown</td>
<td>2</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Diet</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>136</td>
<td>91.3%</td>
</tr>
<tr>
<td>Vegetarian</td>
<td>9</td>
<td>6.0%</td>
</tr>
<tr>
<td>Vegan</td>
<td>1</td>
<td>0.7%</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>2.0%</td>
</tr>
</tbody>
</table>
Renal Transplant cohort summary statistics

The planned recruitment target for the renal transplant cohort was for a total of 340 transplanted patients, with a mix of around 300 deceased donors and 40 live donors. The final total number of transplanted patients was 312, of whom 214 were deceased donors and 86 were live donors. Within the cohort acute complications were observed within at least 66 patients, 10 with acute rejection and 56 with delayed graft function. In terms of the causes of ESRD, glomerulonephritis and inherited renal disease are over-represented in the cohort relative to their frequency in the renal disease population at large which often happens in renal studies unless specifically focusing on other causes such as hypertension or diabetes. The frequency of DGF is slightly lower than expected although several with missing data are known to have had dialysis in the first week after transplant and may be confirmed as having DGF later. This is also though partly a reflection of the population recruited with the live donor proportion being higher than the transplant population overall. Efforts are ongoing to obtain the missing data values also and the richness of this cohort will continue to improve over time with further data chasing and up to 5 years long term follow-up planned. The summary statistics for the cohort are shown in Table 83.

Table 83 Characteristics of the renal transplant patients

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients transplanted</td>
<td>312</td>
</tr>
<tr>
<td>Median age at transplantation (years; range) (missing 5)</td>
<td>52 (19-80)</td>
</tr>
<tr>
<td>Transplants per centre</td>
<td></td>
</tr>
<tr>
<td>Liverpool</td>
<td>21 (6.7)</td>
</tr>
<tr>
<td>Leeds</td>
<td>150 (48.1)</td>
</tr>
<tr>
<td>York</td>
<td>18 (5.8)</td>
</tr>
<tr>
<td>Newcastle</td>
<td>29 (9.3)</td>
</tr>
<tr>
<td>Hull</td>
<td>10 (3.2)</td>
</tr>
<tr>
<td>Portsmouth</td>
<td>40 (12.8)</td>
</tr>
<tr>
<td>Plymouth</td>
<td>2 (0.6)</td>
</tr>
<tr>
<td>Bristol</td>
<td>17 (5.4)</td>
</tr>
<tr>
<td>Sheffield</td>
<td>10 (3.2)</td>
</tr>
<tr>
<td>Bradford</td>
<td>15 (4.8)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>191 (61.2)</td>
</tr>
<tr>
<td>Ethicity (%)</td>
<td>Female</td>
</tr>
<tr>
<td>-------------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cause of ESRD (%)</th>
<th>Pyelonephritis/reflux nephropathy</th>
<th>28 (9.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diabetes</td>
<td>15 (4.8)</td>
</tr>
<tr>
<td></td>
<td>Glomerulonephritis</td>
<td>85 (27.2)</td>
</tr>
<tr>
<td></td>
<td>Hypertension/vascular</td>
<td>36 (11.5)</td>
</tr>
<tr>
<td></td>
<td>Obstructive</td>
<td>3 (1.0)</td>
</tr>
<tr>
<td></td>
<td>Inherited</td>
<td>70 (22.4)</td>
</tr>
<tr>
<td></td>
<td>Other/Unknown</td>
<td>61 (19.6)</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>14 (4.5)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transplant types</th>
<th>DBD</th>
<th>126 (40.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DCD</td>
<td>88 (28.2)</td>
</tr>
<tr>
<td></td>
<td>LD</td>
<td>86 (27.6)</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>12 (3.8)</td>
</tr>
</tbody>
</table>

| Median ischaemic time (range) (n=293; 17 missing CIT and n=278; 34 missing WIT) | Cold (h:min) | 12:17 (00:25 – 23:53) |
|-----------------------------------------------------------------------------|--------------|
|                                                                            | Warm (min)   | 00:40 (00:03 – 02:23) |

<table>
<thead>
<tr>
<th>Acute complications</th>
<th>DGF</th>
<th>56 (17.9)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Missing</td>
<td>17 (5.4)</td>
</tr>
<tr>
<td></td>
<td>AR</td>
<td>10 (3.2)</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>14 (4.5)</td>
</tr>
</tbody>
</table>
Learning and recommendations for future bioresources

Within this NIHR Programme Grant for Applied Research, three different patient cohorts, have been established, together collecting 5976 sets of samples from 2116 patients across 50 NHS centres.

The parallel development of these three projects, in particular the diverse expertise of the teams and individual experiences gleaned from each, not only cross-fertilised each of the projects, but has provided some good learning points worthy of disseminating to future researchers wishing to embark on similar endeavours.

- The successful delivery of these cohorts confirms that the UK has an outstanding national network of enthusiastic patients, clinical investigators and NHS centres willing to participate in research studies. It also verifies the widely shared view that the NIHR CRN’s systems for setting up national research studies work and although clearly too slow initially, are improving in both reduction of complexity and increased pace. Whilst notable improvements have been made, the NIHR must continue to streamline, simplify and speed-up the study approval and set-up process. Researchers are advised, where possible, to keep study elements as simple as possible, for unusual or complex factors may cause delays. At a local level, study/trial co-ordinators should push for SSI forms to be submitted to R&D offices quickly, to start the clock towards sign off.

- Another noteworthy highlight of the NIHR CRN is the unprecedented access it provides to the smaller or less well known research sites. The CRN specialty groups played a key role in identifying potential research centres, many of whom went on to be outstanding recruiters to all three of our cohorts. Not only is this excellent for research, it also provides NHS patients outside of the main centres of academic excellence the opportunity to participate and benefit from cutting edge research. As many regional services, including transplant and cancer, operate around a hub and spoke model, researchers need to consider where patients will be identified, consented, treated and followed up. Multiple centres may be involved in a single patient’s journey and therefore good co-ordination of and communication with sites is essential. Losing patients at follow-up through discharge back to their initial referring
hospital is certainly posing a challenge for the RCC cohort at present at some sites for example.

- The RTB ethical approval process may have several benefits over project-specific approvals for certain types of research, for example collecting surplus surgical tissue. However, for prospective multi-centre observational cohorts, such as those conducted within this programme, a project-specific approval process followed by transfer to an RTB may with hindsight have proven easier and quicker to deliver. Project-specific approval would have negated the delays encountered during NIHR CRN portfolio approval and the issues involved the Human Tissue Act (HTA), in particular the huge burden and cost associated with “storage incidental to transportation”. Within this initial phase of a multicentre research programme such as this, there appear to few benefits to using the RTB process but certainly for long-term sustainability, governance and maximising the use of samples, the RTB infrastructure will be invaluable.

- At the start of this programme, the zeitgeist surrounding patient consent to biomarker research and biobanks was very much in favour of giving patient’s choice and the opportunity to opt in or out of different aspects. Whilst this is still a long-term goal, our experience and that of others suggests that the vast majority of patients do not take up most of these options and are happy to provide broad consent. However, the administrative burden and error rate associated with managing such multi-level consent processes is high. Electronic patient management and consenting systems, will likely reduce this burden and are under development. However, until such systems are the norm and deployable across multiple centres (a major challenge), researchers are advised to only consider a multi-level consent process when the use of broad consent will hinder recruitment. A consideration here though is to what extent that is possible as in this case this aspect was raised by the ethics committee at the time as being necessary. However practice is evolving and experience of such aspects is greater and undoubtedly patients’ views will be important in this area. If multi-level consent must be used, then it is important to consult with and thoroughly test the process with patients and healthcare professionals.
Within the RCC cohort, 18% of patients did not consent to their relatives being contacted regarding any health findings, if they themselves could not be contacted. The precise reasons for this were not clear and warrant further analysis, but one might speculate that they did not wish for their families to be informed of an identified genetic susceptibility to an illness. In the case of a serious and actionable genetic finding, a healthcare professional may believe they have a duty of care to the patient’s relative(s), which may justify breaching their confidentiality with the donor patient. More research with patients is required in this area to refine patient information leaflets and consent forms, to ensure that patients adequately understand how their medical information may or may not be used with regard to their family’s care.

Pre-analytical factors are a major source of uncertainty in biomarker research, as verified in Chapter 13. We implemented standardised processes, local training, contingency planning and rigorous Quality Control for these within the cohorts. However, future researchers may benefit from going further and monitoring/auditing sites for their sample processing competency and compliance. Researchers should pay specific attention to blood collection tubes as the lack of standardisation of tube top colours makes them especially vulnerable to errors. Investigators should also budget for sample processing errors within a percentage of the final sample numbers, for power calculations etc, depending on the complexity of the study.

In comparison with single centre sample banking, there are many additional costs associated with multicentre sample banking. Of particular note are: the manufacturer and distribution of standardised sample kits, provision of local freezer storage and temperature controlled return shipping. The expense of providing high quality sample banks and associated clinical data is little appreciated and often the financial support offered by funders is low compared with what is needed – a fact which goes against the drive to improve sample quality for research purposes.

Compliance with CRFs could have been improved within these cohorts. There is no doubt that the use of hand written paper based CRFs, sent via post is complex and costly for all involved. The time delay for data entry onto the database also makes real-time monitoring and management of the study difficult. The use of electronic
data capture and the increasing use of routinely collected data from electronic patient records, in particular pathology results, should be further developed and encouraged. Testing and early auditing of CRFs with research nurses is also advisable.

- We observed a noticeable decline in the number of samples and follow-up forms being completed after the initial consent visit or discharge. Researchers could consider incentives or penalties to improve compliance with study protocols. Undoubtedly this may in part reflect the lack of “policing” in a biomarker study such as this, compared with a CTIMP where regular inspections of site files and compliance with regulatory requirements of the MHRA is the norm. Additionally the consequences are quite different.

Concluding remarks

These cohorts form an invaluable resource that will underpin research studies validating biomarkers of renal cancer, renal transplant and liver disease for many years to come. A review of the learning outcomes above suggest that whilst there is still much to do, many of the problems and challenges encountered within this programme have been or are already being addressed.
Chapter 12 - Review and prioritisation of circulating biomarkers in renal cancer and renal transplantation
This chapter describes systematic reviews of two areas to enable serum and plasma protein biomarker prioritisation for subsequent studies, namely i). biomarkers of prognosis in localised clear cell RCC (ccRCC), and ii). biomarkers for the early detection and diagnosis of delayed graft function following renal transplantation and long-term prognosis. The clinical context and rationale for these areas as being the initial focus have been presented in Chapter 10.

Renal Cancer

Literature Search Strategy

During the early part of the Programme a review of the literature was undertaken to identify a small number of biomarkers with high potential. This was done in order to focus some of the initial technical studies and in some cases generate further supporting evidence to warrant eventual investigation using the multicentre bank of samples accrued during the Programme (described in Chapter 13). Prior to final selection of the biomarkers to evaluate in the first major study involving assessment of prognosis in patients with localised ccRCC, a systematic review of the literature was carried out. Publications were searched using PubMed. Search terms used were as broad as possible to maximise coverage, with subsequent manual filtering as indicated below (Figure 38) to select the relevant references. Reasons for inclusion or exclusion at the various stages are as indicated. This strategy was selected following an iterative testing process, checking to see if selected known diverse references were detected and for this reason the term “biomarker” was not included as it was too restrictive in some cases. Exclusion during the search itself using the term “metastatic” was not used either as it was recognised that many studies involved both patients with localised and metastatic disease and it was important to include those with later manual filtering removing those solely involving patients with metastatic disease. The term “predictive” was included as although this should primarily refer to the response to treatment, it is still often used interchangeably with prognostic. Biomarkers with only single published studies were excluded as there was insufficient evidence to justify evaluation in a multicentre cohort at this stage but some may be evaluated in the future using local samples and eventually the NIHR cohorts as described in Chapter 10. To provide further background information or context for specific clinical or biological areas where the biomarkers are discussed below, additional specific references were also searched for as needed but not included in the final reference counts in Figure 38.
The systematic literature review process adopted for circulating serum or plasma biomarkers of prognosis in clear cell RCC. The search for publications relating to circulating prognostic biomarkers in RCC to allow a focus on serum and plasma was carried out on 23/06/2015 on PubMed.

<table>
<thead>
<tr>
<th>Search</th>
<th>Query</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>“renal cancer” OR “kidney cancer” OR “renal cell carcinoma” OR RCC Filters: English</td>
<td>35710</td>
</tr>
<tr>
<td>#2</td>
<td>Filters: English</td>
<td>30683</td>
</tr>
<tr>
<td>#3</td>
<td>outcome OR prognosis OR prognostic OR predictive OR recurrence OR relapse</td>
<td>2296771</td>
</tr>
<tr>
<td>#4</td>
<td>blood OR serum OR plasma OR circulating</td>
<td>4313261</td>
</tr>
<tr>
<td>#5</td>
<td>#2 AND #3 AND #4</td>
<td>2162</td>
</tr>
</tbody>
</table>

2162 abstracts reviewed

1,992 discounted
Non-biomarker studies, reviews, case studies, therapeutic predictive (therapy) studies, non-protein (e.g. circulating cells, miRNA or DNA), cohorts not applicable e.g. metastatic or dialysis-related RCC

166 papers obtained and checked

53 discounted
Predictive (therapy), tissue only, single studies

113 papers reviewed

Review of serum and plasma biomarkers of prognosis in ccRCC
A number of analytes routinely available for measurement in hospital labs have demonstrated an association with outcome in patients presenting with localised RCC although as yet they are not generally used clinically for that specific purpose. If sufficient evidence could be provided for any of these, either alone or as part of a panel, they have the advantages of being relatively easily and cheaply measured with widely available and standardised assays in many cases. Although focusing on proteins, we have also included some others in this category such as electrolytes. It is also important to bear in mind that although not within the scope of this review, several routinely measured haematological variables have shown promise also. Given their routine measurement pre-operatively, they would certainly merit inclusion in any analyses, including particularly the neutrophil lymphocyte ratio. Following this we review those protein biomarkers that are still measured only in the research lab environment currently.

**Routinely Measured Analytes**

Many such analytes, not unexpectedly, are associated with prognosis in cancers generally, particularly in the metastatic disease setting, and this is also true for metastatic RCC with serum LDH, calcium and haemoglobin, for example, being included within the MSKCC nomogram. However even for localised disease, several such factors appear to have prognostic value although often not included as variables in many studies, presumably due to the perceived likely lack of specificity, and with many studies potentially confounded by inclusion of all subtypes. For example, we first reported pre-operative sodium as being independently prognostic for DFS and OS both patients with localised RCC or when all stages were included (n=212). This was true whether treated as a continuous variable or dichotomized to above and below the median value (139 mmol/L; HR=0.44, 95% CI [0.22,0.88], p=0.014) with patients with values equal to or below the median having significantly poorer survival. No studies have yet explored this further although the association of hyponatraemia with worse survival has since been confirmed in patients with metastatic disease being treated with IL-2/IFN-based therapy or targeted therapies but whether this is prognostic or predictive is not yet clear. Interestingly in a study investigating the outcomes of patient with tumour thrombus in either the renal vein or inferior vena cava (IVC), sodium wasn’t listed amongst the large number of pre-operative laboratory variables analysed but high serum potassium was significantly associated with poor survival on multivariate analysis.
Hypercalcaemia has been reported to increase with increasing stage, even in the absence of bone metastases and in a large study of 1,707 patients with localised ccRCC, hypercalcaemia was found in 9% of patients and was significantly independently prognostic for CSS (risk ratio 1.64, p=0.002) together with anaemia and ESR. 460, 461 This was later confirmed in another large study which also demonstrated the significant association of tissue parathyroid hormone-like hormone (PTHHLH; also known as parathyroid hormone-related protein/PTHrP) mRNA levels with OS in the clear cell subtype of RCC. PTHHLH has been linked to the hypercalcaemia seen as part of the paraneoplastic syndrome seen in cancers such as RCC for example.462 An earlier study had shown a significant correlation between serum calcium and PTHrP with calcium but not PTHrP being significantly associated with stage although the authors did highlight possible issues with long-term storage and PTHrP and the use of serum rather than plasma.463 A further study demonstrated that high serum PTHrP was a significant adverse indicator in terms of survival at univariate but not multivariate level although numbers were small (n=51) with only 7 patients having hypercalcaemia.464

In a very early study, serum alkaline phosphatase (ALP) which exists in several isoforms with major sources being liver and bone although routinely used assays do not measure specific forms, was reported to be a better indicator of outcome than a bone scan.465 In a study involving 365 patients with RCC the incidence of paraneoplastic elevation of ALP was reported to be 21.1% and with no stage-related differences and a significant association with poorer survival.466 A further similar sized study examining a range of common lab variables found a significant association of ALP with PFS and when results for the non-metastatic subgroup were examined, serum ALP was an independent prognostic factor for CSS.467 However the previously described study involving 1,707 patients with localised ccRCC found elevated ALP in 8% of patients but no significant association with CSS.461 Addressing very specific clinical groups within RCC populations, two studies evaluated outcome in patients with either local recurrence or extension into the IVC.468, 469 In the former, serum ALP at the time of recurrence was prognostic for poor outcome in terms of CSS but only at univariate level although numbers were small (n=54) and in the latter, pre-operative ALP was prognostic on multivariate analysis in the whole group (n=166), and on univariate analysis only if restricted to patients with non-metastatic disease.468, 469
Recently, gamma-glutamyltransferase (GGT) has also been implicated as having independent prognostic significance in a large study involving 921 patients with RCC (all stages and subtypes) although a study involving 700 patients with non-metastatic disease only, GGT was only significant on univariate analysis.\textsuperscript{470, 471}

Obesity is a recognised risk factor for RCC and BMI is associated with poor prognosis and two recent studies have independently shown significant associations between low pre-operative serum total cholesterol and worse prognosis.\textsuperscript{472, 473} Both involved patients with metastatic and localised disease with one including only ccRCC subtype (n=364) whereas the later study included all subtypes (n=867).\textsuperscript{472, 473} In both studies, total serum was a significant independent predictor for CSS and this was also the case in one study when the only the subgroup of patients with localised disease was examined.\textsuperscript{473} Interestingly three studies examining nutritional status, employing varying indices in each but all of which included serum albumin together with varying factors such as cholesterol or lymphocyte count, found nutritional deficiency to be a significant independent prognostic factor for recurrence, or CSS and OS when restricted to patients with local or loco-regional disease.\textsuperscript{474-476}

A small number of studies have examined “conventional” tumour markers used in other cancer types. Serum $\beta_2$-microglobulin was related to both stage and grade (n=145) and was significantly inversely related to survival although only at the univariate level.\textsuperscript{477} Analysis of CEA, CA-50, CA-19-9, CA-125 and CA-15-3 found elevated serum levels for all except CEA in a cohort of 154 RCC patients, with correlations of the latter two with stage and grade and independent prognostic value of CA-125.\textsuperscript{478} In a more recent study, almost two decades later, CA-15-3, CA-125 and $\beta_2$-M were all associated with stage and grade and CSS but only CA-15-3 was significant on multivariate analysis and also for PFS.\textsuperscript{479} Whether the marked differences between the two studies regarding the relative merits of CA-125 and CA-15-3 reflect changes in assays, the patient populations or the outcome (it is not clear in the earlier study if CSS or OS was examined) is not clear. Two studies but by the same group have also examined the free $\beta$-subunit of HCG in 177 and 256 patients with RCC, possibly with overlapping patients used in both studies, and found the free $\beta$-subunit of HCG to be an independent prognostic variable for CSS.\textsuperscript{480, 481} Evaluation of neuron-specific enolase (NSE) has shown positive correlations with stage and grade in three studies, two of which were
undertaken by the same group but in their later study using a commercially assay. A significant independent prognostic association with survival was reported.\textsuperscript{482-485}

Abnormalities in coagulation occur widely in cancer patients. Initial findings of significantly higher plasma fibrinogen concentrations in RCC patients with metastatic disease have subsequently been confirmed and extended by two larger studies.\textsuperscript{486} Including all subtypes and stages (n=286), an independent association was found for fibrinogen with both DFS and OS and a larger study involving 994 patients with localised RCC of all subtypes found an association of plasma fibrinogen (measured prospectively the day before surgery) with tumour stage and grade with independent prognostic significance for MFS, CSS and OS with hazard ratios varying from 2.15 to 2.48 (p<0.001 for all).\textsuperscript{487488} A further smaller study (n=128) also found a significant association with CSS but additionally reported D-dimer to be negatively associated with OS.\textsuperscript{489} Clearly this represent an attractive possibility given the ready availability of standardised assays in clinical laboratories and these results could be relatively easily confirmed in further prospective studies on a multicenter basis.

**C-Reactive Protein (CRP)**

CRP is an acute phase protein, elevated in many inflammatory diseases and often raised in patients with cancer. With the main source generally being the liver and expression regulated by cytokines such as interleukin-6 (IL-6) and tumour necrosis factor- α (TNF-α), it has been shown that IL-6 is produced by renal cancer cells and functions as an autocrine growth factor.\textsuperscript{490} Circulating IL-6 is increased in patients with RCC, particularly in those patients with metastatic disease, and several studies have shown a significant correlation with serum CRP.\textsuperscript{491-494} High serum IL-6 has been shown to be associated with poorer survival but either not maintained or not examined in multivariate analysis.\textsuperscript{492, 495} It is now known that CRP is also produced by renal tumour cells and indeed intra-tumoural CRP staining has been significantly associated with OS.\textsuperscript{496, 497} However the extent to which this contributes to the circulating CRP pool is unclear as no significant normalization of CRP or IL-6 was seen at approximately 3 months post-surgery for RCC although an earlier study had reported significantly concentrations by 6 months-post-surgery.\textsuperscript{494, 498}
Although one of the earliest studies twenty years ago analysing several acute phase proteins including CRP found all to be significant at univariate level but only orosmucoid (α₁-acid glycoprotein) at multivariate level, a large number of studies have now shown CRP to be a significant independent prognostic biomarker in RCC and a recent meta-analysis showed that this is the case across the urological cancers.⁴⁹⁹ Many of the studies have investigated the value of pre-operative serum CRP against the standard clinic-pathological prognostic factors and demonstrated independent prognostic significance for DFS/RFS and/or CSS in patients with localised disease, although one study involving only patients with disease extending into the vena cava showed a significant association of CRP with CSS only at univariate level.⁴⁹⁵,⁵⁰⁰-⁵⁰⁵ The largest such study included 1,161 patients (including 146 with M1 disease) across all subtypes and confirmed CRP as a significant independent prognostic factor for CSS and OS.⁵⁰⁶ Interestingly, although pre-operative serum CRP was demonstrated in one study to be a significant independent prognostic factor for RFS in patients with localised disease (n=263), non-normalisation of CRP post-operatively rather than pre-operative CRP was significantly associated with OS on multivariate analysis. The 5-year survival figures were 96.9% compared with 30% in patients whose CRP normalized or failed to normalise respectively.⁵⁰⁷

Several recent studies have also explored the complementarity or additive value with existing clinic-pathological prognostic factors or scoring systems. In a study involving 83 patients with localised disease, CRP and also the UISS and SSIGN scores were all shown to be independent prognostic predictors of RFS.⁵⁰⁸ This was confirmed in a similar subsequent study (n=130) for SSIGN score and CRP but not UISS, with only CRP and platelets being independent for 1-year OS.⁵⁰⁹ In a later study by this group CRP was shown to remain significantly prognostic when lifestyle factors such as smoking and obesity were included in the model.⁵¹⁰ In a study of 313 patients (21% with metastatic disease), pre-operative CRP treated as three categorical variables (≤4.0 mg/L, 4.1-23.0 mg/L and >23.0 mg/L) was independently prognostic for CSS (p=0.003). Importantly, CRP added to the UISS prognostic model, improving its accuracy by 3.8% at 5 years (p<0.001).⁵¹¹ The TNM-C score, which is based on CRP and TNM alone, was developed based on 249 RCC patients with advanced and localised disease, dichotomizing CRP as < or ≥ 5 mg/L and combining it with TNM to generate 4 risk groups, stratifying CSS from 99% to 18% 5-year survival across the groups.⁵¹² External validation was achieved in a further 290 patients, with a c-index of
0.865 and subsequently in an additional 518 patients with ccRCC with a c-index of 0.85. More recently in a cohort of patients with localised ccRCC (n=403), pre-operative CRP was independently significantly associated with DFS and increased the prognostic accuracy of the SSIGN score. The ability to increase the prognostic accuracy in this way though will depend very much on the initial performance of the scoring system and its components and addition of CRP to a model including TNM stage, grade and Karnofsky index did not improve the model’s performance which already had a high predictive value of 88.1%.

A study from our own group, investigated whether CRP prior to nephrectomy adds to a published model which is solely based on pre-operative factors, in this case age, gender, symptoms, tumour size, clinical T stage and metastatic status. Based on 286 patients (84% clear cell subtype) and with CRP dichotomized as ≤ or >15 mg/L, 5-year survival rates of 72% (95% CI 65%-78%) and 33% (95% CI 23%-44%) were shown with CRP independently prognostic factor for OS (p<0.006) and CSS (p<0.001), and adding significantly to the pre-operative score.

Clearly serum CRP has significant prognostic potential and has the major advantage of being easily measured in hospital labs using existing assays although it is worth noting that in the more recent studies, the availability of different generations of higher sensitivity CRP assays will have allowed quantification over a wider range. Challenges moving forward will include the possible effects of co-morbidities on CRP, and the investigation of optimal cut-points, with for example the studies above using a variety including 2.5, 3, 4, 5, 7.5, 10 and 15 mg/L. Alternatively the possibility of treating CRP as a continuous variable should be explored with only three studies having adopted this and with two of these undertaking a comparison and concluding categorical was best.

**Serum amyloid A (SAA)**

Significantly elevated concentrations of the acute phase protein SAA have been reported in patients with RCC, particularly in patients with metastatic disease with patients with localised disease having values largely similar to those of healthy controls. Three studies have now shown SAA has significant independent prognostic value for CSS although only moderately sized cohorts (72-119) but including all stages. The two largest and most recent studies (including one we undertook involving only ccRCC patients) were remarkably...
similar (HR 2.46; 95% CI 1.17-5.15; p=0.017, compared with HR 2.51; 95% CI 1.09-5.78; p=0.030).\textsuperscript{517, 520} However when we also included CRP in our model, SAA was no longer independently significant.\textsuperscript{520}

**Ferritin**

Serum ferritin is another acute phase protein reported to be increased in patients with RCC, increasing with stage and significantly correlated with tumour volume and with some evidence supporting the tumour as being a possible source of some of the circulating ferritin.\textsuperscript{521-523} In a further analysis of serum ferritin in 158 RCC patients of all stages, grouping on the basis of both pre-operative and post-operative ferritin combined (n=103) relative to normal healthy control values (i.e. high or normal), and stage, were significantly independently associated with survival. Pre-operative serum ferritin alone was significant only at univariate level and this was further confirmed in a smaller study (n=52) with renal vein ferritin being higher than peripheral vein ferritin and both significantly associated with survival but not on multivariate analysis.\textsuperscript{523, 524} Interestingly in some studies patients were excluded if anaemic, transfused recently or co-morbidities included liver disease, due to possible effects on ferritin and this may preclude its usefulness.

**Erythropoietin (EPO)**

An initial study with 57 patients with RCC found increased plasma EPO in 63% of cases but no correlation with stage or grade.\textsuperscript{525} However subsequent studies (n=165 and n=195) reported associations of serum EPO with grade and stage and survival although not significant on multivariate analysis.\textsuperscript{526, 527} In a more recent study examining both serum EPO (n=138) and tissue EPO receptor (n=56) expression, the association of EPO with stage and grade was confirmed and also the association between higher EPO and lower survival but examination of the patients with localised disease only (n=110) found no association of EPO with DFS.\textsuperscript{525} However grouping the 47 patients analysed for both serum EPO and tissue EPO receptor identified a group with high levels of both and with worse CSS although the numbers are small.\textsuperscript{525}

**Vascular endothelial growth factor (VEGF)**
With the role played by VEGF in angiogenesis and its regulation by VHL and early studies reporting elevated circulating concentrations of VEGF in patients with RCC, particularly those with metastatic disease, exploring its potential as a prognostic marker was logical. However, mixed results were obtained in survival analyses with analysis of serum VEGF pre-operatively in all subtypes of RCC (n=146) or just ccRCC (n=45) finding no significant association with outcome, a slightly larger study (n=161) finding a significant association but only at univariate level and the most recent study of similar size and patient mix (n=124) finding that on multivariate analysis, VEGF was an independent marker associated with CSS and RFS. There was some indication that concentrations differed between subtypes in two of these studies but this is unlikely to have accounted for the differences in findings here. Analysis of serum VEGF in 83 patients with non-metastatic clear cell RCC found serum VEGF to be a significant independent predictor of recurrence (p=0.013). The RFS was significantly lower in the cases which stained positively for VEGF or had higher serum concentrations, although there was no significant association between VEGF staining and serum VEGF. Interestingly serum VEGF has shown to increase markedly in most patients (n=66 including 48 with distant metastases) following nephrectomy and in this study pre- to post-operative changes in serum VEGF were not significantly related to outcome.

Studies have also analysed plasma VEGF concentrations and a significant correlation has been reported between this and cytoplasmic VEGF staining. However, although tissue VEGF correlated with outcome, no such relationship was seen for circulating VEGF. Using carefully prepared citrated plasma samples to minimise release of VEGF from platelets, plasma VEGF prior to surgery has been shown to be significantly associated with CSS (n=74 patients including 67 ccRCC and 22 stage IV) but this is lost on multivariate analysis. Focussing on just the clear cell subtype but including all stages (n=102), plasma VEGF was associated with T stage and grade but not nodal or metastatic disease. However, both tissue and plasma VEGF were significantly associated with PFS and CSS. A later study analysing plasma samples in ccRCC patients only (n=68) reported higher VEGF concentrations in patients with nodal or metastatic spread but found no significant association with OS.
Clearly although VEGF appears to have prognostic potential in several studies, there are inconsistencies in the findings, probably reflecting the heterogeneity of the studies in terms of RCC subtypes, whether restricted to patients with localised disease only and possibly even more critically, whether serum or plasma has been used and how samples have been processed. It is now recognized that platelet release of VEGF is a main contributor to VEGF measured in serum but equally this may also contribute to circulating VEGF concentrations measured in plasma depending on the sample handling conditions. Platelet number and serum VEGF concentrations have been reported to be highly significantly correlated in patients with advanced cancer and hence the prognostic value of serum VEGF may actually be related to the prognostic value of platelet number or thrombocytosis. These and other pre-analytical aspects of VEGF measurement are considered further in Chapter 14.

**Carbonic anhydrase IX (CAIX)**

Interest in CAIX in RCC stemmed initially from the observation of the selective binding of a monoclonal antibody (G250) to RCC tissue but not normal proximal epithelium. The antigen was later identified as CAIX, a hypoxia-inducible protein involved in the regulation of intra- and extracellular pH and upregulated as a consequence of VHL inactivation in RCC. Studies examining the prognostic value of CAIX expression in tissue samples have largely found high expression to be related to better outcome although inconsistencies between studies have been found and it is still not clear whether CAIX has independent prognostic value, as we have recently reviewed. However, with the recognition that CAIX exists as a shed form in the blood and urine, several studies have explored soluble CAIX as a prognostic marker in RCC, with most using the same assay as the one we have used in this programme and making inter-study comparison more feasible. In patients with ccRCC (n=91; 79 with localised disease), mean serum CAIX concentrations were significantly higher in patients with metastatic disease (p=0.004), with concentrations correlating with tumour grade, size and stage. On univariate analysis, serum CAIX was significantly associated with early relapse in patients with localised disease. In a larger study involving 361 RCC patients with all subtypes, serum CAIX was related to stage and grade but failed to reach significance as a prognostic marker for CSS when dichotomised around the median in the cohort of patients with ccRCC (n=287) and was not a significant prognostic factor on multivariable analysis. Our subsequent study analysed serum CAIX, CRP and plasma osteopontin (OPN) prior to nephrectomy in 216 patients with ccRCC (24% with M1 disease).
and found CAIX to be significantly associated with CSS, DFS and OS on univariate analysis but independently prognostic for OS only on multivariate analysis. The combination of the three markers outperformed stage.

Matrix metalloproteinase-7 (MMP-7)

MMP-7 (matrilysin) is a member of a family of zinc-containing enzymes involved in proteolytic degradation of many ECM components and hence involved in many of the pathological processes in cancer, particularly invasion and angiogenesis. Expression of MMP-7 in RCC tissue has been shown to be increased relative to normal kidney and independently prognostic for CSS or OS. Following a two-dimensional PAGE study of an RCC cell line with screening of separated proteins using RCC patient sera to detect immunogenic reactivity, pro-MMP-7 was detected and subsequently reported as being elevated in serum from RCC patients. Using an assay measuring pro-MMP-7, MMP-7 and TIMP-1 complexed forms (according to the manufacturer’s information), plasma concentrations were found to be significantly elevated in RCC patients (n=97 including 45 with metastatic disease) compared with healthy controls. This was particularly the case for patients with distant metastases and on multivariate analysis, MMP-7 was found to be independently prognostic for CSS (HR 2.70; 95% CI 1.39-5.24; p=0.003).

Recent pilot data from our group also supports plasma MMP7 as being associated with tumour size and stage.

Osteopontin (OPN)

Osteopontin, also called secreted phosphoprotein 1 (SPP1), is a member of the small integrin-binding ligand N-linked glycoprotein family (SIBLING). One of the most abundant non-collagenous extracellular matrix proteins in bone, it is now known to have a widespread tissue distribution and plays a role in many processes including cell adhesion, remodeling, angiogenesis and inflammation. Increased tissue expression has been reported for many cancer types, including renal. The first study to demonstrate an independent prognostic role for OPN in RCC involved 80 patients of all stages and subtypes. Plasma OPN was significantly elevated in patients with metastatic disease and was the only factor, together with the presence of metastases, out of several examined to retain independent prognostic significance for CSS in a multivariate model. In a larger study involving 269 patients with
renal cancer of all subtypes, analysis of OPN in plasma and serum samples (n=75 and 116 respectively with ccRCC subtype), found higher concentrations in plasma (median 2.3 fold higher than serum) and independent prognostic significance for CSS in ccRCC with higher concentrations linked to poorer survival, particularly for plasma, although almost 50% of patients included had metastatic disease. The lower OPN concentrations in serum compared to plasma have been found in other studies and may be accounted for, at least in part, by the known cleavage by thrombin during clotting and this is discussed further in Chapter 13. Surprisingly, neither stage nor grade were independently prognostic in this study. Subsequently, our study focusing on ccRCC only (n=216) found pre-operative plasma OPN to be significantly prognostic for OS, CSS and DFS but only at univariate and not multivariate level. This may have been due to CRP being included in our model which was strongly prognostic and correlated with OPN and also only 24% had metastatic disease. Interestingly, a group of patients with low stage RCC who had higher OPN concentrations were identified as being at high risk of death, mainly from non-cancer-related causes. Importantly and although not relevant to the initial study proposed here with localised RCC, OPN was one of the markers identified as a strong prognostic marker for patients with metastatic RCC in the placebo arms of phase 2 and 3 trials with pazopanib, out-performing routine clinical indicators.

**Immunosuppressive acidic protein (IAP)**

Serum IAP was first described in 1986 as being elevated in patients with high stage RCC compared with low stage disease and initially confirmed in a slightly larger study of 66 patients and 133 previously untreated RCC patients where IAP had an area under the ROC curve of 0.894 for metastatic disease. IAP was also found to be associated with higher grade in 181 RCC patients with higher IAP providing a >4-fold risk of higher grade. In terms of survival, an initial study (n=143) showed higher IAP to be associated with poorer survival at 3 years although not analysed at multivariate level and a subsequent smaller study involving 92 patients of mixed stages confirmed the association with survival, although not significant on multivariate analysis. In the most recent study in 2006, IAP doubling time when measured longitudinally after nephrectomy in patients with localised disease who subsequently relapsed (n=125) have been shown to be independently prognostic for survival (p=0.0026). Using the cut-point of a doubling time of greater or less than 200 days, 3-year survival was 58.9% and 12.5% respectively. A limitation of several of these and which has
probably accounted for the restriction to Japanese studies and the long time period over which the studies span is the assay availability with all except the last (which used a nephelometric assay) having used radial immunodiffusion to measure IAP. Additionally it is not clear for most studies which histologic subtypes of RCC were included.

**Tumour M2 pyruvate kinase (TuM2-PK) and Thymidine kinase (TK1)**

Serum concentrations of the glycolytic dimeric M2 isoenzyme form of pyruvate kinase, TuM2-PK were initially reported simultaneously in two studies in 1999 to be increased in some patients with RCC compared with patients with benign diseases or healthy controls and correlating with tumour stage and grade in both studies and grade in one.\(^5^6^9,\)\(^5^7^0\) These results have been extended with more recent larger studies largely confirming these findings.\(^5^7^1,\)\(^5^7^2\)

In patients with RCC of varying subtypes (n=116), pre-operative TuM2-PK was also found to be a significant independent prognostic marker for disease recurrence (HR 7.3; CI 1.1-47.8, p=0.037) with a crude 5-year RFS of 55% for patients with elevated concentrations compared with 94% for normal concentrations (p<0.001).\(^5^7^1\) This study also examined the prognostic potential of TK1 which had previously been reported to be associated with grade, stage and size in in a small study involving 27 patients and found similar results as for TuM2PK in terms of independent prediction of disease recurrence and with crude 5-year RFS of 21% vs 90% (p=0.002) for patients with elevated or normal concentrations of TK1 respectively.\(^5^7^1,\)\(^5^7^3\)

**Soluble Il-2 receptor (sIL-2R)**

Serum concentrations of sIL-2R have been shown to be significantly increased in RCC patients with significant associations with clinical stage and higher sIL-2R being associated with poorer CSS although not examined at a multivariate level and with only 52 patients, with subtypes being unclear.\(^5^7^4\) However a recent study involving only patients with ccRCC (n=70) has confirmed these findings.\(^5^7^5\) Significant correlations with stage were demonstrated and particularly with higher concentrations in patients with stage IV disease (p<0.001) and the group with higher sIL-2R being associated with shorter CSS (p<0.05).\(^5^7^5\) Larger studies with multivariable analysis are now needed however.

**Basic fibroblast growth factor (bFGF)**

353
Following early demonstrations of elevated serum bFGF in approximately 50% of patients with RCC and a trend towards higher concentrations with increasing stage and grade, significant correlations with stage and grade were reported in a larger study involving 206 patients. However, although higher bFGF was associated with poorer survival, this was not significantly associated with outcome on multivariate analysis. In the most recent study performed ten years ago, no significant association between serum bFGF and stage was apparent (n=74) although higher concentrations were found in patients with metastatic disease.

Conclusions and Prioritisation

Clearly, several circulating markers reflecting diverse aspects of RCC biology appear promising but need further systematic evaluation including whether a multiplex panel would be most effective and the optimal combination and whether these could add value to or outperform existing clinic-pathological scoring systems. On the basis of the evidence presented above, OPN, VEGF (serum and plasma), CAIX and CRP were prioritised to take forward in the initial prognostic study. Although fibrinogen appeared to be very promising, this was not possible as it would need to be measured prospectively on freshly collected, rather than frozen, citrated plasma. MMP7 also appeared to be promising but further work needs to be carried out to determine the effects of inhibitors such as TIMP-1 on the assays, the detectability of pro-MMP7 and the relative suitability of plasma or serum. Lack of commercially available assays preclude investigating IAP and further studies are needed with larger numbers of patients to investigate TK1 and TuM2-PK. Several other markers do not seem to have been pursued for many years despite promising earlier findings and this may relate to lack of assays currently. There are also several markers which appear to have promise, although only at single-study level at present, and covering a wide range of tumour biology. These include insulin-like growth factor-1, soluble B7-H3, tumour-associated trypsin inhibitor and vitamin D. In the future it is possible that these may be explored using the RTB assembled within this Programme once further supportive evidence is available.
Renal Transplantation

We have restricted this first review to circulating plasma or serum biomarkers. Although intuitively, urine would appear to be the most obvious choice of biological fluid for the detection of biomarkers, in the early post-operative period following kidney transplantation several considerations reduce the clinical validity of urinary biomarkers. Firstly, some patients, particularly those that receive a pre-emptive (prior to starting RRT) kidney transplant maintain residual urine output from their native kidneys, which will confound the measured values of biomarkers. Secondly, patients who develop DGF and who do not have any residual urine output will have minimal urine output. Thirdly, from an analytical and scientific perspective there are a number of issues concerning the normalisation of urinary measurements of biomarkers.

Literature Search Strategy

Publications were searched using PubMed. Search terms used were as broad as possible to maximise coverage, with subsequent manual filtering as indicated below (Figure 39) to select the relevant references. Reasons for inclusion or exclusion at the various stages are as indicated. This strategy was selected following an iterative testing process, checking to see if selected known diverse references were detected and for this reason the term “biomarker” was not included as it was too restrictive in some cases. To provide further background information or context for specific clinical or biological areas where the biomarkers were discussed, additional specific references were then searched for as needed.
Figure 39 The systematic literature review process adopted for circulating biomarkers of DGF in renal transplantation. The search for publications relating to circulating biomarkers of DGF in renal transplantation to allow a focus on serum and plasma biomarkers was carried out on 11/10/2015 on PubMed.

<table>
<thead>
<tr>
<th>Search</th>
<th>Query</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>“renal transplant” OR “kidney transplant”</td>
<td>31905</td>
</tr>
<tr>
<td>#2</td>
<td>“renal transplantation” OR “kidney transplantation”</td>
<td>90010</td>
</tr>
<tr>
<td>#3</td>
<td>#1 OR #2</td>
<td>94358</td>
</tr>
<tr>
<td>#4</td>
<td>“delayed graft function” OR “DGF”</td>
<td>2839</td>
</tr>
<tr>
<td>#5</td>
<td>#3 AND #4</td>
<td>2551</td>
</tr>
<tr>
<td>#6</td>
<td>blood OR serum OR plasma OR circulating</td>
<td>4364777</td>
</tr>
<tr>
<td>#7</td>
<td>#5 AND #6</td>
<td>1271</td>
</tr>
<tr>
<td>#8</td>
<td>Filters: English</td>
<td>1205</td>
</tr>
<tr>
<td>#9</td>
<td>Filters: English; Review</td>
<td>47</td>
</tr>
<tr>
<td>#10</td>
<td>#8 NOT #9</td>
<td>1158</td>
</tr>
</tbody>
</table>

1158 abstracts reviewed

106 papers obtained and checked

30 papers included in the review

1052 discounted

urine, tissue, case studies, non-protein (e.g. circulating cells, miRNA or DNA), non-biomarker studies

76 discounted

Tissue, paediatric, non-biomarker studies
Review of serum and plasma biomarkers of DGF following renal transplantation – diagnosis and prognostic utility for long-term outcome

Although the focus of this Programme is on protein biomarkers, given the relative paucity of biomarkers with a significant level of supporting evidence currently emerging in this area of renal transplantation at present, at the end of this section we also mention some promising non-protein studies.

Creatinine

Although not a protein biomarker, creatinine is of course widely available and routinely measured. Creatinine is a product of the metabolism of creatine, which is released from muscle. Creatine is non-enzymatically dehydrated to creatinine in the liver. Creatinine is freely filtered in the glomerulus and does not undergo significant metabolism or reabsorption in the kidney. As such it is the usual clinical standard for monitoring kidney function post-transplantation, together with urinary output. Consequently it is the benchmark against which other biomarkers are compared which in itself is problematic as it is actually quite poor in several ways. For example increases or decreases in its concentration lag behind true changes in kidney function, including by a number of days, and it is influenced by a wide variety of commonly variable factors such as age, gender, muscle mass, level of nutrition (including protein intake) and liver function. In addition around 10% of creatinine is cleared by tubular secretion, which can be disrupted by particular medications e.g. trimethoprim and cimetidine. Post-transplantation, the use of serum creatinine suffers from all these limitations. Glomerular filtration rate (GFR) estimates based on serum creatinine, including the commonly used Modification of Diet in Renal Disease (MDRD) equations, and the Cockcroft and Gault equation are not useful unless a steady state is present, which of course is not the case post-renal transplant. At lower GFR, secretion of creatinine increases as a proportion of creatinine clearance, causing these equations to overestimate GFR.

Cystatin C

Cystatin C is a 13.4 kDa cysteine protease inhibitor produced by all nucleated cells. It is an endogenous marker of glomerular filtration rate as it is freely filtered by the glomerulus and is not reabsorbed into the circulation or secreted (Bicik, Z 2005). In an early study of serum cystatin C 41 consecutive deceased donor kidney transplant patients had concentrations
measured before and 1, 3, 6 and 10 days post surgery. The study demonstrated that in patients with DGF the serum cystatin C did not fall. It is unclear from the paper how many patients developed DGF or what the ROC was. Serum cystatin C and other biomarkers (serum NGAL and Interleukin 18) were measured for the first 3 days in 78 recipients of deceased donor kidney transplants, of which 26 had DGF. Serum cystatin C values were effective in distinguishing DGF from immediate graft function with an AUC of 0.83 on day 1 post operatively.

Fonseca et al investigated the potential of uNGAL to predict DGF and one year kidney transplant function in comparison with cystatin C. This prospective study measured serum cystatin C at day 0, 1, 2, 4 and 7 post transplant in 20 consecutive patients, of whom 18 developed DGF. Day 1 cystatin C predicted DGF with an AUC 0.95. Cystatin C was further investigated as a comparative biomarker in a prospective study of malonaldehyde (MDA) as a predictor of DGF. Plasma concentrations were measured pre-operatively (day 0) and post-operatively (days 1, 2, 4, and 7) in 40 consecutive recipients of kidney transplants of whom 18 developed DGF. Day 1 serum cystatin concentrations accurately predicted DGF with an AUC of 0.91. The same research group utilised a multiple biomarker approach was used to detect DGF which included uNGAL, serum leptin, serum MDA and serum cystatin C. It was a prospective cohort study of 40 consecutive patients including deceased donors and living donors of whom 18 developed DGF. Serum cystatin C had an AUC of 0.914 at 8-12 hours post-surgery. The most informative combination was a triple biomarker approach with an AUC of 0.96 that included serum creatinine, MDA and serum cystatin C. A further analysis of the data demonstrated there was a trend for cystatin C values on day 1 post-transplant to correlate with kidney transplant function at 3 months when divided into upper, middle and lower tertiles. Serum cystatin C demonstrates very good utility as an early predictive biomarker of DGF following kidney transplantation.

Neutrophil Gelatinase Associated Lipocalin (NGAL)

In a study of 41 consecutive deceased donor kidney transplants serum NGAL (sNGAL) was measured before and 1, 3, 6 and 10 days post surgery. The study demonstrated that in patients with DGF the sNGAL did not fall. It is unclear from the paper how many patients developed DGF or what the receiver operator characteristic curve was. However it concluded that sNGAL needed to be investigated further as a potential marker of DGF. A prospective
observational study measured plasma NGAL (pNGAL) in 41 patients receiving a deceased (n=39) or living donor kidney transplant (n=2). DGF developed in 15 patients, all of whom had received a deceased donor kidney. The pNGAL receiver-operator characteristic curve at 12h for predicting DGF demonstrated an AUC 0.97. Serum NGAL and other biomarkers (serum cystatin C and Interleukin 18) were measured for the first 3 days in 78 recipients of deceased donor kidney transplants, of which 26 had DGF. Serum NGAL values were ineffective in predicting DGF. In contrast, sNGAL at 24 hours post-transplantation was shown to be an accurate predictor of DGF which affected 6 of the 33 transplant patients, with an AUC 0.82. Of these 33 patients 20 received a deceased donor kidney and 13 received a living donor kidney.

A retrospective study measured sNGAL (and serum IL-18) in 59 recipients pre-operatively and days 1,5 and 14 post-operatively. Day 1 sNGAL had an AUC 0.86 in the 14 patients who developed DGF. A further retrospective study analysed sNGAL in 67 patients of whom 27 received kidneys donated after a circulatory death (DCD). The function in the DCD kidneys never recovered which may account for the AUC of 0.99 to predict DGF on the first day post-operatively. Hollmen et al (2014) measured sNGAL in 176 consecutive deceased donor kidney transplant recipients utilising two different methods. Sixty six patients developed DGF and sNGAL was significantly higher in this group with an AUC of 0.853 when measured on day 1. It also predicted DGF lasting longer than 14 days with an AUC of 0.825. In a further study of 97 patients (17 living donors and 80 deceased donors), of whom 20 developed DGF, there was no correlation between donor pNGAL (and uNGAL) and post-transplant DGF. In this study pNGAL predicted DGF with an AUC of 0.73 at 6 hours, 0.80 at 12 hours and 0.85 at 24 hours post transplant.

A different approach has been taken in two studies that investigated the serum concentrations of NGAL in the kidney donors prior to surgery. Hollmen et al (2011) collected serum and urine samples from 99 consecutive deceased kidney donors prior to the operation and their 176 recipients. Serum NGAL failed to predict DGF. Muller et al performed a prospective, multicentre observational study that included 146 brain-dead donors leading to 243 transplants, of which 56 developed DGF. The concentrations of sNGAL in the donors failed to predict DGF or normal transplant function in recipients.
There is now a significant body of evidence suggesting that serum NGAL (less so for plasma NGAL) has good potential for use as early biomarker for predicting DGF when measured in the recipient but not in the donor.

**Aminoacylase-1**

Serum aminoacylase-1 (ACY-1) was identified as a potential outcome biomarker following mass spectrometry analysis of serum samples, before and on day 2 post transplant from five patients with DGF and five with immediate transplant function. Following development of an ELISA for ACY-1, analysis of results from an initial validation cohort (n=55 patients) showed moderate predictive value for ACY-1 on day 1 or 2 post-transplant, complementing cystatin C. A further validation cohort 194 patients (54 with DGF) confirmed this association with a day 1 AUC of 0.74 for ACY-1, 0.9 for cystatin and 0.93 in combination. Importantly however, analysis of long-term follow-up for 54 patients with DGF showed a highly significant association between day 1 or 3 serum ACY-1 and dialysis-free survival, mainly associated with kidney donation after brain death, and offering the potential of use in stratification of follow-up.

**Other Promising biomarkers**

A number of additional biomarkers have been evaluated but either only at single study level or in small patient numbers. The most promising of these, all of which would require further validation before consideration for multicentre evaluation, are described briefly below.

**Complement**

Ischemia/reperfusion injury in the kidney results in the activation of the complement cascade. A terminal panel of complement molecules (C3a, C5a, and C5b-9/membrane attack complex) was analysed following kidney transplant reperfusion. Seventy-five kidney transplant recipients were divided into early, slow, and delayed graft function groups. Blood samples were collected from the renal vein during reperfusion. Analysis revealed that C5b-9/membrane attack complex (C5b-9mac) concentrations were two to three times higher in DGF patients than patients with early and slow graft function (P<0.005). In addition, C5b-9/mac concentrations had a relatively high clinical sensitivity and specificity (70%-87.5%) for the prediction of early and 1 year kidney transplant function. Receiver-operator characteristic curves were not calculated which limits the interpretation of this study.
**C-Terminal Agrin Fragment (CAF)**

The C-terminal agrin fragment (CAF) is a cleavage product of agrin, the major proteoglycan of the glomerular basement membrane. It has been proposed that elevated CAF values may be related to reduced glomerular filtration and clearance. Serum CAF and creatinine concentrations were measured in 96 healthy individuals and in 110 patients undergoing kidney transplantation before and after transplantation.\(^{601}\) Serum CAF concentrations at day 1 and 3 were significantly associated with DGF (40 patients) with an AUC of 0.81. This small study demonstrates moderate clinical utility for CAF in the early prediction of DGF.

**Fms-Like Tyrosine Kinase (Flt-1)**

Ischemia/reperfusion induces tubular epithelial and endothelial cell damage in the kidney transplant and contribute to the development of DGF. Chapal prospectively assessed the kinetics of the soluble VEGF receptor, soluble Fms-like tyrosine kinase-1 (sFlt-1) in 136 consecutive kidney transplant patients.\(^{602}\) Patients with DGF had higher sFlt-1 concentrations at all time points during the first 7 days following kidney transplantation. Multivariate analysis demonstrated that a peak plasma sFlt-1 of 250 pg/mL or higher was associated with 2.5-fold increase in the risk of DGF (P=0.04). Receiver-operator characteristic curves were not calculated which limits the interpretation of this study.

**IgA Antibodies to β2-Glycoprotein I**

The prevalence of IgA anti-β2-glycoprotein I antibodies (IgA-a β2GPI-ab) in patients on dialysis is elevated (>30%), and these antibodies correlate with mortality and cardiovascular morbidity. Isolated IgA-a β2GPI antibodies are associated with thrombosis. A single centre prospective study to evaluate the effect of IgA-a β2GPI-antibodies in 269 patients following kidney transplantation.\(^{603}\) Presence of IgA-a β2GPI-ab in pre-transplant serum was examined retrospectively. Eighty nine patients were positive for IgA-a β2GPI-ab, and the remaining patients were negative. Multivariate analysis showed that the presence of IgA-a β2GPI-ab was an independent risk factor for early graft loss (P=0.04) and DGF (P=0.04). Receiver-operator characteristic curves were not calculated which limits the interpretation of this study.

**Pregnancy – associated plasma protein (PAPP-A)**

Pregnancy-associated plasma protein A (PAPP-A) has been shown to be a marker of acute coronary syndrome and cardiovascular pathology. Blood samples were taken from 178
patients prior to receiving their first deceased donor kidney transplant. Sixty-one patients subsequently developed DGF. Pre-transplant PAPP-A values were significantly elevated in the group of patients with DGF. Multivariate analysis showed that PAPP-A was an independent risk factor for DGF although ROC curves were not examined.

**Interleukin-16 (IL-16)**

Alachkar et al analysed serum and urine samples from 61 patients 48 hours following kidney transplant for a panel of 23 cytokines including IL-16 which has been implicated in ischaemia-reperfusion injury. Six patients developed DGF. The AUC was 0.74 for serum IL-16 in this small study and therefore does not provide any compelling evidence to support the use of IL-16 as an early predictor of DGF.

**Interleukin-18 (IL-18)**

Interleukin-18 is a cytokine that mediates inflammation and ischaemic tissue injury in many organs including the proximal tubules in the kidney. Hall et al compared alternative serum biomarkers with creatinine for predicting DGF. Interleukin-18 and other biomarkers (serum cystatin C and serum NGAL) were measured prospectively for the first 3 days in 78 recipients of deceased donor kidney transplants, of which 26 had DGF. Serum IL-18 was unable to distinguish DGF from slow graft function or immediate graft function. A retrospective study measured sIL-18 (and serum NGAL) in 59 recipients pre-operatively and days 1, 5 and 14 post-operatively. Day 1 sIL-18 had an AUC 0.63 in the 14 patients who developed DGF and therefore had limited value. Serum IL-18 does not appear to be useful in predicting DGF.

**Leptin**

Leptin is removed from circulation primarily by the kidney and could be considered surrogate marker for kidney function. A prospective study was performed to measure the concentrations of leptin in 40 consecutive patients at day-0, 1, 2, 4 and 7 following kidney transplantation. Median leptin concentrations were significantly higher in patients developing DGF (n=18) at all times points. The leptin reduction rate between pre-transplant and one-day post operatively moderately predicted DGF with an AUC of 0.73. Day-1, serum leptin predicted DGF with an AUC of 0.76 Separating the analysis by gender showed improved the performance of leptin in predicting DGF for male gender with an AUC of 0.86.
A further prospective cohort study of 40 consecutive kidney transplant patients utilised a multiple biomarker approach including serum leptin. Both deceased donors and living donors were included in the study of whom 18 developed DGF. Serum leptin had an AUC of 0.76 at 8-12 hours post-surgery. Serum leptin appears to be only moderately useful in predicting DGF.

**Resistin**

Brain death triggers a complex cascade of molecular and cellular events resulting in the release of inflammatory mediators. Resistin increases during several inflammatory diseases and after intracerebral bleeding or head trauma. It promotes endothelial activation and may initiate an inflammatory response. The potential role of plasma resistin values in the brain dead kidney donors in predicting DGF in the recipient was analyzed in 63 kidney transplant patients. Twenty-six recipients of kidneys from living donors were used as controls. Donor resistin values in the recipients of kidneys from brain dead donors correlated with DGF with an AUC of 0.765. Donor resistin values appears to be of only moderate clinical utility in predicting DGF.

**ST6Gall**

Galbeta1-4GlcNAcalpha2-6 sialyltransferase (ST6Gall) is an acute phase reactant whose release from cells can be induced by proinflammatory cytokines. It has been hypothesised that patients with chronic kidney disease may have circulating concentrations of ST6Gall, which might increase the risk of DGF. Serum concentrations of ST6Gall were measured in 70 patients immediately before receiving a kidney transplant. The mean serum level of ST6Gall was significantly higher in the patients than in 19 controls. Twenty patients developed DGF and had significantly higher concentrations of ST6Gall pre-transplant than 50 patients who had immediate graft function. In a multivariate analysis the ST6Gall level was found to be independent risk factors for the development of delayed graft function. Receiver-operator characteristic curves were not calculated which limits the interpretation of this study.

**Stem Cell Factor (SCF)**

Alachkar et al investigated whether a panel of serum and urine cytokines could act as early biomarkers for predicting DGF and slow graft function (SGF). Serum and urine samples
from 61 patients were collected 48 hours following kidney transplantation and analysed using a multiplex enzyme-linked immunosorbent assay (ELISA) technique to measure concentrations of 23 cytokines. One of the cytokines included SCF which is been implicated in early inflammation and tissue fibrosis. Six patients developed DGF and 8 developed SGF. The AUC was 0.88 for serum SCF, however sampling was performed at 48 hours following kidney transplant which would reduce its clinical utility as an early predicted biomarker of DGF.

**Hydroxyeicosatetraenoic Acids**

Eicosanoids are the active metabolites of arachidonic acid and have been implicated in the pathogenesis of ischaemia/reperfusion injury in the kidney. 20-Hydroxyeicosatetraenoic (20-HETE) acid is one such active metabolite and is generated via cytochrome P450 enzymes. To assess the potential roles of eicosanoids the concentrations of lipoxigenase (LOX)-derived 5-, 12- and 15-hydroxyeicosatetraenoic (HETE) acids concentrations were measured in 69 kidney transplant recipients.\(^6\) The kidney transplant recipients were divided into early, slow and delayed graft function groups. Blood was taken directly before, and immediately following kidney transplant reperfusion. Application of newly proposed cut-off limits for 5-HETE, 12-HETE and for 15-HETE resulted in 72.5-81.5% sensitivity and 50-54% specificity for slow graft/DGF prediction. A mixed model analysis revealed that recipients classified according to results of 5-HETE and 15-HETE cut offs were able to predict 1-year kidney transplant function. A further study of 20-HETE measured concentrations during the first 5 min of kidney transplant reperfusion, and analysed whether the concentrations were associated with post-transplant kidney function.\(^7\) Sixty-nine kidney transplant recipients were divided, according to their outcome, into early, slow and DGF groups. The sensitivity, specificity, positive and negative predictive value of 20-HETE in discriminating early and slow graft function from DGF were 69%, 54%, 74% and 48% respectively. Both the studies were performed in relatively small groups of patients and did not calculate the receiver-operator characteristic curves, which limits the interpretation of these studies.

**Malondialdehyde**

Ischaemia/reperfusion injury results in cellular death mediated by a number of different pathways. Oxidative stress is one such pathway and leads to the generation of reactive oxygen species. Malondialdehyde (MDA) is a marker of oxidative stress and has been
investigated as a potential biomarker of DGF and transplant function at 1 year in a prospective study of 40 consecutive kidney transplant patients. Plasma concentrations of MDA were measured pre-operatively (day 0) and post-operatively (days 1, 2, 4, and 7). At all time points after transplantation mean MDA concentrations were significantly higher in patients developing DGF (n=18). Day 1 MDA concentrations accurately predicted DGF with an AUC of 0.90, with a performance higher than SCr (AUC of 0.73) and similar to cystatin C (AUC of 0.91). Multivariable analysis revealed that MDA concentrations on day 7 represented an independent predictor of 1-year graft function. Another prospective cohort study of 40 consecutive patients including deceased donors and living donors utilised a multiple biomarker approach to detect DGF which included uNGAL, serum leptin, serum MDA and serum cystatin C. Serum MDA had an AUC of 0.90 at 8-12 hours post-surgery. The most informative combination was a triple biomarker approach with an AUC of 0.96 that included serum creatinine, malondialdehyde and serum cystatin C. Malondialdehyde used either alone or in combination with other biomarkers demonstrates good potential as an early biomarker of DGF.

**Neutrophil-Lymphocyte Ratio**

The neutrophil-lymphocyte ratio (NLR) is an indicator of inflammatory status and has been used to assess outcome in critically ill surgical patients. A retrospective study was performed to investigate the effect of pre-operative elevated NLR on the kidney transplant recipient in relation to the risk of DGF. The pre-operative white blood cell count of 398 kidney transplant recipients was analyzed. Two hundred and forty nine patients received kidneys from donors after brain death (DBD), 61 from donors after circulatory death (DCD), and 88 from living donors. An NLR >3.5 was considered to be elevated. One hundred three patients developed DGF, of which 67 had NLRs >3.5. Multivariate analysis demonstrated that NLR > 3.5 had a Hazard ratio of 10.673 with CI = 6.151-18.518. Receiver-operator characteristic curves were not calculated which limits the interpretation of this study.

**Regulatory T Cells**

Regulatory T (Treg) cells have been shown to be protective in models of acute kidney injury (AKI) and their suppressive function predictive of AKI following kidney transplantation. The role of Treg cells as a biomarker of DGF has been explored in a prospective observational cohort study. Fifty three deceased donor kidney transplant recipients were divided into
those who developed AKI (n=37), including DGF and slow graft function (SGF) and immediate graft function (n=16) groups. Pre-transplantation peripheral blood CD4CD25FoxP3 Treg frequency was quantified by flow cytometry. Regulatory T-cell suppressive function was measured by suppression of autologous effector T-cell proliferation by Treg in co-culture. In univariate and multivariate analyses accounting for the effects of cold ischemic time and donor age, Treg suppressive function accurately predicted AKI (DGF and SGF) with an AUC of 0.82. The same group also performed a prospective observational cohort study utilising flow cytometry to measure pre-transplant recipient circulating CD4+CD25+CD127lo/- and CD4+CD127lo/- tumour necrosis factor receptor 2 (TNFR2)+ Treg cells in 76 deceased donor kidney transplant recipients of whom 18 patients developed DGF.614 The receiver operating characteristic curves demonstrated an AUC of 0.75 and 0.77, respectively for the percentage and absolute number of CD4+CD127lo/-TNFR2+ Treg cell in predicting DGF. Neither of these studies demonstrate good clinical utility in predicting DGF. The first study had a good AUC but combined DGF with SGF.

Summary

The data presented in the systematic review demonstrates that there have been a number of small studies that have investigated a range of potential serum/plasma biomarkers to enable the early detection of DGF but relatively few that have looked at longer-term outcomes. The studies have been heterogeneous in terms of the population studied and definitions of DGF, although most have applied the definition of receipt of haemodialysis in the first week following kidney transplantation. The most promising serum biomarkers for the early detection of DGF appears to be NGAL and serum cystatin C with ACY-1 validated but only in a single centre to date. The use of cystatin C is becoming more widespread throughout the health care system and with a growing familiarity it may well prove to be utilised in the future as an earlier biomarker of DGF. In the case of NGAL there remain issues of standardisation with respect to which cut-off values to recommend.

There are a number of emerging biomarkers that have been studied in acute kidney injury outside of kidney transplantation with the most recent candidate being TIMP-2 and IGFBP7. It is only natural to assume that studies of these biomarkers will follow in the setting of kidney transplantation and predicting outcomes. The most recent study investigating urinary
TIMP-2 and IGFBP7 was unfortunately rather unimpressive. The most promising urinary biomarker to date is NGAL and there have been a number of studies that have demonstrated its potential utility. However, as discussed earlier there are significant issues surrounding the use of urinary biomarkers post-kidney transplant.

There is a distinct lack of studies in the literature that have investigated the clinical use of biomarkers in predicting longer–term outcomes. It has been proposed by many experts that panels of multiple biomarkers may be able to improve the predictive value but again such studies are lacking. The most obvious panel to utilise would include serum cystatin C and serum NGAL which have both demonstrated very good utility in predicting DGF in the early phase following kidney transplantation, and ACY-1 should also be explored further, particularly in view of its promising prognostic performance. There is now an excellent opportunity to validate existing biomarkers and investigate novel biomarkers in a cohort of kidney transplant patients with a well described phenotype.
Chapter 13 - Exploring Technical Aspects of Biomarker Assays

– Verification, Validation and Pre-Analytical Variables
The evaluation of the validity and performance of assays is fundamental to their introduction in to clinical practice. This chapter describes some of the aspects, concepts and guidelines for the technical evaluation of assay performance together with pre-analytical considerations, before describing the practices developed and results generated in a series of such studies undertaken within this Programme, in preparation for the analysis of biomarkers in specific research studies.

Appraisal of Assay Performance

General Concepts

Ensuring assays are appropriately validated is critical in terms of ensuring that measurements of biomarkers are accurate and reproducible, both across time and between laboratories. It is important within this to ensure that they are “fit for purpose” and deliver the level of performance needed for the study phase or clinical situation, avoiding a dogmatic approach to guidelines which may not be completely relevant at some stages. The level of validation may vary depending on whether the assay is for research use only (RUO) or requires CE-marking or FDA-approval and is for use in a hospital lab or pharmaceutical lab for example. However many immunoassays are relatively easy to purchase and use and many studies are undertaken by researchers with the assumption that because they are commercially available, they will be “fit for purpose”. It is increasingly recognised that this is not the case and examples affecting specific analytes include recent studies with an ELISA (USCN Life Science) for CUB and zona pellucida-like domains protein 1 (CUZD1) which actually measured CA125 and an ELISA for soluble hemojuvelin from the same company which did not detect the specified target but some unknown protein. A key issue which has been highlighted is the plethora of biotech companies which have sprung up, marketing a very wide range of immunoassays and antibodies, with several companies often using the same reagents although this is not always clear, and with very little apparent validation. The extent of this problem was really made apparent with the testing of over 5,000 commercially available antibodies in immunohistochemistry and Western blotting where almost 50% failed although importantly not all had been certified by the manufacturers for these specific applications and only a generic protocol was used for all with a limited range of sample types examined. This type of problem is not restricted to antibodies, with other laboratory biological reagents also posing problems and with limited access to information to resolve such issues due to commercial sensitivities.
as with a clinically used biomarker, nevertheless such issues when encountered in research labs are very costly in terms of inappropriate conclusions regarding clinical utility and utility, the waste of money, samples and time. Suggestions of how to minimise the impact of antibody-based problems have been made, covering the various stakeholders.337, 340 Clinical labs are not immune to such problems either with pitfalls and lack of consistent results across gastrin measurements, growth hormone and insulin-like growth factor, cardiac troponin and even serum creatinine for example, depending on the assays used.342, 620-622

Guidelines

Numerous guidelines exist to provide a framework and set consistent standards for assay validation. The Clinical Laboratory Standards Institute (CLSI; www.clsi.org) produces some of the most widely used and accepted standards and guidelines for clinical laboratory measurements, some of which are mandatory for certain regulatory bodies or accreditation. These range from methods and performance standards for specific procedures through to safety, laboratory and quality management system standards and include more than 25 Evaluation Protocols (EPs) for assay evaluation. These vary depending on the particular stage or aspect of the assay being examined. For example EP05 is an extensive validation protocol covering the initial establishment de novo of the precision of an assay (i.e. agreement, although it is actually imprecision which is measured), for example when first developing or significantly changing an assay, whereas EP15 is a shorter verification protocol aimed at confirming stated imprecision results such as those provided by the manufacturer with a commercially available assay.623 Both within-run (repeatability) and between-run assessment of precision needs to be determined, with the latter covering both within-lab precision and overall reproducibility where assessment involves the changed conditions across multiple labs. A list of the available CLSI EPs (as at December 2015) showing the breadth of coverage is shown below (Table 84).

Table 84 List of available CLSI Evaluation Protocols (December 2015)

<table>
<thead>
<tr>
<th>Code</th>
<th>Title of CLSI EPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>EP06-A</td>
<td>Evaluation of the Linearity of Quantitative Measurement</td>
</tr>
<tr>
<td>Documentation Code</td>
<td>Title</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>EP15-A3</td>
<td>User Verification of Precision and Estimation of Bias; Approved Guideline—Third Edition</td>
</tr>
<tr>
<td>EP18-A2</td>
<td>Risk Management Techniques to Identify and Control Laboratory Error Sources; Approved Guideline—Second Edition</td>
</tr>
<tr>
<td>EP18-A2/EP23-A</td>
<td>Sources of Failure Template</td>
</tr>
<tr>
<td>EP21-A</td>
<td>Estimation of Total Analytical Error for Clinical Laboratory Methods; Approved Guideline</td>
</tr>
<tr>
<td>EP23-A™</td>
<td>Laboratory Quality Control Based on Risk Management; Approved Guideline</td>
</tr>
<tr>
<td>EP23-A WB</td>
<td>A Practical Guide for Laboratory Quality Control Based on Risk Management; Workbook</td>
</tr>
<tr>
<td>EP23-A WS</td>
<td>A Sample Form for Laboratory Quality Control Based on Risk Management; Worksheet</td>
</tr>
<tr>
<td>EP24-A2</td>
<td>Assessment of the Diagnostic Accuracy of Laboratory Tests Using Receiver Operating Characteristic Curves; Approved Guideline—</td>
</tr>
<tr>
<td>Code</td>
<td>Title</td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>EP25-A</td>
<td>Evaluation of Stability of In Vitro Diagnostic Reagents; Approved</td>
</tr>
<tr>
<td></td>
<td>Guideline</td>
</tr>
<tr>
<td>EP26-A</td>
<td>User Evaluation of Between-Reagent Lot Variation; Approved Guideline</td>
</tr>
<tr>
<td>EP27-A</td>
<td>How to Construct and Interpret an Error Grid for Quantitative</td>
</tr>
<tr>
<td></td>
<td>Diagnostic Assays; Approved Guideline</td>
</tr>
<tr>
<td>EP28-A3c</td>
<td>Defining, Establishing, and Verifying Reference Intervals in the</td>
</tr>
<tr>
<td></td>
<td>Clinical Laboratory; Approved Guideline—Third Edition</td>
</tr>
<tr>
<td>EP29-A</td>
<td>Expression of Measurement Uncertainty in Laboratory Medicine;</td>
</tr>
<tr>
<td></td>
<td>Approved Guideline</td>
</tr>
<tr>
<td>EP30-A</td>
<td>Characterization and Qualification of Commutable Reference Materials</td>
</tr>
<tr>
<td></td>
<td>for Laboratory Medicine; Approved Guideline</td>
</tr>
<tr>
<td>EP31-A-IR</td>
<td>Verification of Comparability of Patient Results Within One Health</td>
</tr>
<tr>
<td></td>
<td>Care System; Approved Guideline (Interim Revision)</td>
</tr>
<tr>
<td>EP32-R</td>
<td>Metrological Traceability and Its Implementation; A Report</td>
</tr>
<tr>
<td>EP36-Ed1</td>
<td>Harmonization of Symbology and Equations, 1st Edition</td>
</tr>
</tbody>
</table>

Additionally insights and educational resources can also be provided through bodies such as the European Federation for Clinical Chemistry and Laboratory Medicine (EFLM), exemplified by a recent paper concerning the assessment of quality of the analytical methods and the various parameters to investigate. Internationally recognised standards also exist such as ISO 15189 for medical laboratories, providing particular requirements for quality and competence and accreditation of laboratories will be based on meeting such standards and guidelines as stipulated.

**Pre-Analytical Errors and Variation**

In addition to appropriate analytical validation of the assay for any specific biomarker, considering and minimising the impact of pre-analytical variables is arguably just as important to ensure clinically valid results. In the hospital clinical chemistry labs where the total testing process (TTP) can be broken down into pre-, intra- and post-analytical phases, it is generally considered that the majority of errors, possibly up to 75%, occur in the pre-analytical phase. There are many anecdotal examples including falsely elevated urinary
Amylase due to contamination with salivary amylase from the nurses collecting the samples, and haemolysed blood samples due to contamination with rainwater but essentially all aspects of the pre-analytical phase can be affected either by errors or by variability in the processes adopted.  

Potential errors include requesting of the incorrect test and failure to comply with any defined requirements, for example time of day or after fasting, through to obtaining the specimen from the correct patient in inappropriate tubes, labelled incorrectly, and then failure to process the sample according to agreed protocols, ready for analysis. Although historically the focus has often been on the analytical phase, this is now regarded as the “tip of the iceberg” and more efforts in the future will be directed towards addressing quality aspects of the pre-analytical phase. Within the EFLM, a Working Group for Preanalytical Phase (WG-PRE) has been established to promote this, recently publishing their findings with regard to non-compliance with CLSI guidelines for phlebotomy.  

Similarly the Laboratory Errors and Patient Safety (LEPS) working group of the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) has included 35 quality indicators relating to the pre-analytical phase in its model of quality management for the TTP. In the UK a Preanalytical Working Group within the Association for Clinical Biochemistry and Laboratory Medicine has also begun to address this, with a recent UK-wide survey indicating wide variation in the recording of pre-analytical issues (for example recording of haemolysis, icterus and lipaemia in 80% of labs but sample mislabeling recorded in only 56.9%) but support for quality assurance schemes going forwards and a UK NEQAS pilot scheme has been set up. In the research community, the International Society for Biological and Environmental Repositories (ISBER) have a Biospecimen Science Working Group looking at aspects of defining and coding critical pre-analytical factors in biobanking (SPREC – Sample PREanalytical Code) and possible quality control procedures and the Biospecimen Reporting for Improved Study Quality (BRISQ) recommendations describe critical elements of the pre-analytical phase which should be included in publications to aid transparency. The importance of the development of appropriate ICT tools has also been raised.

Wherever possible in the hospital environment, automated procedures are used or are being introduced to reduce the chance of error. To reduce variability, standards or guidelines have been produced by the CLSI covering all procedures from venepuncture through patient and sample identification, tube specifications, to sample processing. Using blood as an example, technical sources of such variability include factors such as type of blood collection
tube and components (even the presence of gel activators can markedly affect results), inadequate fill compromising sample:anticoagulant ratio, haemolysis, elapsed time between venepuncture and centrifugation, centrifugation speed and temperature, elapsed time between centrifugation and analysis and, for samples not analysed immediately, the duration and temperature of storage. These last two factors are particularly important for banking of samples for biomarker research studies, storage duration and temperature. Recent reviews discuss this in greater detail, both from the perspective of the biomarker discovery/validation studies in the research environment where often the impact is poorly understood with many proteins detected in proteomic studies having been little studied, and the hospital labs where it is often better understood for the specific panel of tests in routine use. Where the potential impacts can’t be controlled for, it is essential that they are at least appreciated and factored into study protocols and analysis to avoid potential confounding of the results. Such effects need to be considered in terms of understanding the pre-analytical uncertainty of measurements and any potential changes, for example the impact of venepuncture or tube type on measurement of a panel of 15 routinely measured serum analytes or a comparison of five tube types for serum preparation where clinically significant differences were seen depending on tube type for several analytes including creatinine, amylase and phosphate. The consequences of pre-analytical errors or variation in research studies are generally not appreciated but include loss of time and wasted resources and further indirect effects through repetition and follow-on studies based on these results, all leading to massive financial consequences. More broadly, the irreproducibility in pre-clinical research in the life sciences generally has been estimated to cost >$28 million per year in the USA alone. In the clinical environment, failure to standardise pre-analytical conditions can have a critical impact on patient safety, for example in terms of wrong or delayed diagnosis or inappropriate treatment and of course there are also financial implications. Within the development of a health economic model assessing the impact of pre-analytical errors, various clinical case study scenarios were used and the average cost of a pre-analytical error was remarkably similar in North America and Europe at just over $200 and accounting for up to 1.2% of hospital operating costs.

Biological variability must also be taken into account when considering the pre-analytical phase. Factors such as an individual’s age, gender, diet, time of day, co-morbidities, effects of drugs, smoking status or alcohol consumption, body mass and for females, menstrual cycle stage, pregnancy and menopausal status can all potentially have an impact on biological
markers. In addition, the acceptability criteria of an assay for use in a particular clinical context may vary depending on the physiological or pathological variability. For example, a higher level of assay imprecision may be acceptable if a large effect size is being sought and given sufficient sample size. Recently and analogous to the STARD criteria, a checklist has been produced by the EFLM which specifies the key data which need to be reported to allow studies of biological variation to be interpreted and used effectively.\textsuperscript{647} A databank of intra-subject and inter-subject variation for over 300 routinely measured analytes has been compiled from the literature and is available online (http://www.westgard.com/biodatabase1.htm) together with two guest essays describing how such data is important in clinical chemistry in defining quality specifications and desirable performance characteristics such as total error, imprecision and bias (http://westgard.com/guest17.htm) and (http://www.westgard.com/guest12.htm).\textsuperscript{648, 649} Although the concept of this database is recognised as being very valuable, recently issues have been highlighted which limit this value. These include dubious quality of some of the studies from which the data was derived, the age of some of the studies resulting in the use of methodology which is now outdated, the use of different units of measurement in different studies of the same analyte, the reliance in some cases on data from only a single study and the limitation of many studies to healthy individuals.\textsuperscript{650} The importance of biological variability lies in its requirement for the calculation of the reference change value which enables the interpretation of the significance of a difference in successive results. It also gives information as to whether reference intervals can be used to interpret an individual’s marker results, for example creatinine has a small within-subject variability and large between-subject variability so that any changes should be interpreted in light of that individual’s reference interval. An important consideration to bear in mind is how the clinician interprets assay results also and to what extent the assay performance impacts on that.\textsuperscript{651} There is often confusion about the terminology used and the “unfinished symphony” of the evolution, meaning and relative merits of various reference values such as decision limits, reference change values and reference intervals has been discussed recently.\textsuperscript{652}
Development of Verification Protocols for Commercially Available Immunoassays

Key Technical Validation (Verification) Elements and Criteria

This was developed within this Programme based on a consideration of the various guidelines discussed above and the level of implementation of the assays i.e. in a GCLP-level research lab rather than the diagnostic environment. Key elements to be considered and criteria of acceptability when validating existing immunoassays, often referred to as verification in that context compared with when establishing a new assay, were decided upon although the need for flexibility was also recognised. This is illustrated in the examples provided of studies undertaken within the Programme where a range of problems were encountered requiring some unique investigations which would not be needed generically. A similar flexible approach allowing adjustments or modifications as needed has been described for a range of pharmacodynamics assays from their development within NCI through to deployment in multiple centres.\textsuperscript{653} Similarly the assumption is that samples have been handled under ideal standardised conditions and are therefore suitable for assessment of clinical validity utility with relaxation of this only being possible once pre-analytical aspects have been investigated. If this is not the case, evaluation of some pre-analytical aspects may have to occur earlier in the process.

Essentially the assay protocol specified by the manufacturer is followed, with any required adjustments as necessary recorded, for example alterations in timings or settings of any specific step. The aspects described below are investigated as standard with some being dictated by the performance requirements of the assay, based on its intended clinical application and available information at the time e.g. type of sample matrix intended for use and range of analyte concentrations.

Once the specific matrix has been decided upon, a familiarisation and range-finding run can be undertaken to determine if the initial assay protocol is satisfactory and to determine selection of samples for use in the validation studies such as those with high or low analyte concentrations. Appropriate samples from the assembled bank of samples ("surplus
diagnostic samples”) with known high concentration of endogenous interferents are also needed together with recombinant/purified proteins (analyte and proteins with known homology). The following aspects are then assessed for each matrix.

Standard Assessments

1. Analytical Range

- Limit of Detection (LoD): ~20 reps of zero standard over multiple plates (mean blank + 3 SD)
- Lower Limit of Quantification (LLoQ): serial dilutions of low standard to ~LoD, analysed over multiple plates (n≥3) to generate a precision profile, LLoQ <20% imprecision & inaccuracy
- Evaluation of hook effect: analyte spiked into sample at 100-1000 times highest standard and the reported value should be >highest standard. Spiked sample diluted back into assay range and check recovery.

It is worth noting that a significant element of confusion is the variable use of terminology and methodology employed by users and manufacturers to determine the performance of assays, particularly at the lower concentrations. So values variably quoted include those for functional sensitivity, limit or lower limit of quantification (LoQ/LLoQ), limit of detection (LoD), limit of blank (LoB), minimal detectable concentration (MDC) and sensitivity, making comparisons and interpretation difficult. In particular LoD and LLoQ or LoQ are often used interchangeably and it is critical that a distinction is made between these values and that they are both assessed. This is illustrated for LoB, LoQ and LoD. 654

2. Imprecision

- Intra-assay
- Inter-assay

Assessed with minimum of 2 QC samples (pooled or independent) with a high and a low analyte concentration ≥5 independent determinations for each, over each of five days i.e. 25 determinations in total minimum. Intra-Assay CV should be ≤10% and Inter-Assay should be ≤15% (20% at LLoQ)
3. Accuracy

- Recovery - spike of recombinant/purified analyte is added to ≥3 independent pools or base material of appropriate matrix at 3 different concentrations; acceptable recovery is 80-120%.
- Evaluation of suitable reference materials, if available (≥5 determinations over 3 concentrations; <20% imprecision & inaccuracy)
- Comparison to a reference method if available

4. Analytical Specificity

- Cross-reactivity: identified proteins with homology to the analyte spiked in (recombinant/purified forms) into independent samples (n≥2), at two concentrations spanning the pathophysiological cross-reactant range (if known)
- Parallelism/Dilution linearity (normal working dilution and ≥3 serial dilutions of a minimum of 3 samples). Assessed by back calculating the diluted concentration of the 4 dilutions to the actual concentration with an acceptability limit of ≤15%
- Common interferents: (e.g. rheumatoid factor, lipids, bilirubin, complement, haemolysate)
  - Recombinant analyte spiked into surplus diagnostic samples (n≥3) with known moderate and high interferent concentrations and recovery calculated. For example 150 and 300 µmol/L total bilirubin or 10 and 25 mmol/L triglycerides. Alternatively stock interferents can be purchased and spiked into samples with known amounts of analytes. Final concentrations would be 50 or 150 µg/mL bilirubin (conjugated and unconjugated respectively) or 30 mg/mL for triglycerides. For testing the effects of haemolysis, samples containing known concentrations of analyte can be spiked with haemolysate to produce +5 mg/mL Hb for serum and plasma samples or 5 mg/mL and 2.25 µg/mL (this latter equates to +++ on urine dipstick as determined experimentally) for urine samples.
• Dilutional linearity also assessed for at least 1 spiked test sample.
• If recovery is outside 80-120% or significantly different to previous samples, a dose-response series can be undertaken by spiking ≥5 concentrations of purified/recombinant/synthetic interferent into samples to further assess the effects (n≥3).

5. Evaluation of curve fitting model (≥5 determinations over multiple runs)
   • Imprecision (<10%; 20% at LLoQ)
   • Inaccuracy (<10%; 20% at LLoQ)
     (>80% of non-zero standards, including highest and lowest must pass)

Additional Assessments

These may be necessary depending on the phase of the study.
1. Inter-laboratory imprecision (reproducibility)
2. Reference Ranges
3. Analyte Stability
   • Freeze thaw
   • Short term bench stability
   • Long term storage stability (length and temperature)
4. Pre-analytical variables
   • Biological: e.g. within-subject variability, stress, exercise, diet and alcohol, smoking status, drugs, pregnancy, age, gender, co-morbidities, race, sample timing etc.
   • Technical: e.g. phlebotomy technique, blood collection systems, blood collection tubes, sample preparation procedures, transportation conditions etc.

Specific Biomarker Technical (Assay and Pre-Analytical) Studies Undertaken Within the Programme

The assays selected for validation/verification were for biomarkers which we had prioritised for potential analysis in either the RCC prognostic study undertaken within the timeframe of
this Programme or in future planned studies in renal transplantation in the case of NGAL. Selection was based on published studies at that time and this has been reconfirmed through more recent studies, also as reviewed in the previous chapter. Samples used for these initial studies were obtained with fully informed consent as part of the Leeds Multidisciplinary RTB and had been collected and processed according to stringent in-house SOPs. For example, the Vacuette® system was used (Greiner Bio-One, Frickenhausen), with Z/serum clot-activator tubes (coated with micronised silica particles) and EDTA plasma tubes (K$_2$EDTA). Samples were processed within 45-60 minutes of venepuncture with centrifugation at 2000 x g at 20ºC for 10 minutes. Serum and plasma were aliquoted and stored at -80ºC until used. The exceptions to this were samples obtained for interference studies with high concentrations of factors such as bilirubin and lipid, which were obtained from Leeds Teaching Hospitals Blood Sciences lab as anonymised surplus diagnostic samples under project-specific ethics approval (REC 10/H1313/12) and without patient consent. The specific technical studies we undertook and the results, many of which have been published as indicated, are described in the following sections.

**Osteopontin**

**Introduction**

Osteopontin (OPN) is a member of the SIBLING (small integrin binding ligand, N-linked glycoprotein) family which includes bone sialoprotein. An extracellular matrix glycoprotein, OPN is produced by many cell types and is predominantly secreted although intracellular and other forms of OPN with varying post-translational modifications have been reported.$^{655-657}$ With roles in cell adhesion through binding to integrins and CD44 splice variants, OPN can differentially affect adhesion and migration through cleavage by thrombin with subsequent separation of the integrin and CD44 binding domains. Together with effects on proliferation, apoptosis and differentiation, OPN has been implicated in many processes including tissue remodelling, inflammation and tumorigenesis and is associated with tumour aggressiveness in several cancer types.$^{655-657}$ This includes renal cancer where tissue expression of OPN and plasma OPN are associated with several prognostic clinical variables and are prognostic themselves although only plasma OPN has independent prognostic significance.$^{557, 558}$ Given the promising nature of those findings we selected OPN to evaluate further prognostically and consequently we undertook an evaluation of an OPN ELISA prior to analysis of samples from a cohort of RCC patients within our Leeds Multidisciplinary RTB. The results from that
study confirming the prognostic utility of OPN have now been published and this has been reviewed together with other relevant studies in the previous chapter and has led to OPN being one of the prioritised biomarkers for evaluation in this NIHR Programme as described in the following chapter.\textsuperscript{436}

**Methodology**

The OPN ELISA kit used was the Quantikine ELISA for human osteopontin from R&D Systems. All samples were assayed in duplicate and according to the manufacturer’s protocol. Assessment of assay performance was based upon our in-house protocol described earlier including inter- and intra-assay imprecision, parallelism, and recovery (using recombinant OPN from Abcam) and interference. The main assessment was carried out using EDTA plasma samples as during the clotting process, OPN is proteolytically cleaved by thrombin which is widely thought to lead to the lower OPN concentrations found in serum.\textsuperscript{560} EDTA plasma samples from patients with RCC were used for the main studies with matched serum and EDTA plasma being used for the plasma-serum comparison we undertook to confirm the plasma-serum differences previously reported. In addition as one of the future local uses of this assay was for analysis of samples from patients in the Leeds Melanoma Cohort where samples were shipped by post, we included a small study examining stability over up to 4 days post-venepuncture as at that time there was no published data for the stability of OPN. For this purpose, blood samples were collected into EDTA tubes from 5 melanoma patients and 4 healthy volunteers with informed consent. For each individual, two 4-ml tubes of blood were collected with one being processed immediately after venepuncture and one being processed after 4 days at room temperature. In each case plasma was then removed, aliquoted and stored at $-80^\circ$C until analysis.

**Results**

*Serum versus EDTA plasma*

Results from a matched comparison of serum and EDTA plasma from 6 RCC patients clearly showed the marked differences between serum and plasma OPN concentrations (Table 85). Most serum values ranged between 40 and 65\% of those obtained for plasma with one exception where a value of 93.1\% was seen. These differences are presumed to be due to the thrombin cleavage previously described and accordingly we used EDTA plasma for the assay evaluation studies.
Table 85 Comparison of plasma and serum osteopontin concentrations in matched samples

<table>
<thead>
<tr>
<th></th>
<th>Patient 1</th>
<th>Patient 2</th>
<th>Patient 3</th>
<th>Patient 4</th>
<th>Patient 5</th>
<th>Patient 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasma OPN (ng/mL)</td>
<td>45.11</td>
<td>71.79</td>
<td>36.73</td>
<td>286.65</td>
<td>269.8</td>
<td>99.7</td>
</tr>
<tr>
<td>Serum OPN (ng/mL)</td>
<td>21.45</td>
<td>31.30</td>
<td>18.75</td>
<td>137.03</td>
<td>251.2</td>
<td>63.5</td>
</tr>
</tbody>
</table>

**Imprecision**

Overall intra-assay imprecision was 2.9% with values of 3.1, 2.8 and 2.8% for low, medium and high QC samples respectively with each being assayed 5 times in duplicate in a single assay run. Inter-assay imprecision was also <10%.

**Parallelism**

Samples (n=5 RCC) titrated out in parallel with serial dilutions from 1 in 12.5 to 1 in 100 and parallelism CVs ranged from 6.8 to 12.9% which is within our limits of acceptability (<15%).

**Recovery, Interference and Hook Effect**

As shown below (Table 86), recovery was acceptable for all EDTA plasma samples (n=6; 3 melanoma and 3 RCC) with low spikes of +217.5 ng/mL and high spikes of +389.5 ng/mL with the exception of one sample with the low spike where the replicate CV was >10% and so the result wasn’t available.

Table 86 The recovery of OPN spiked into EDTA plasma samples. Spiking was done at two concentrations using rOPN.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Clinical group</th>
<th>% recovery (low spike)</th>
<th>% recovery (high spike)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Melanoma</td>
<td>102.3</td>
<td>93.1</td>
</tr>
<tr>
<td>2</td>
<td>Melanoma</td>
<td>-</td>
<td>96.0</td>
</tr>
<tr>
<td>3</td>
<td>Melanoma</td>
<td>104.9</td>
<td>109.6</td>
</tr>
<tr>
<td>4</td>
<td>RCC</td>
<td>86.5</td>
<td>87.1</td>
</tr>
</tbody>
</table>
No evidence of a hook effect was seen and bilirubin, haemolysis, triglycerides and rheumatoid factor showed no appreciable signs of interference in terms of measured OPN concentration or dilution linearity.

**LLOQ**
This was assessed as being less than 78 pg/mL equating to 1.95 ng/mL in a sample diluted 25-fold. Further evaluation of this using lower OPN concentrations to determine an exact value was not carried out as all samples measured were well above this value.

**Sample Stability**
As shown below (Figure 40), there was no significant difference in OPN concentrations in EDTA plasma between blood samples (4 healthy controls and 5 patients with melanoma) processed immediately compared with four days later (Wilcoxon matched-pairs signed rank test; p=0.07) with the majority of samples differing by <3% between the two timepoints.

*Figure 40 The effects of immediate versus delayed processing on plasma OPN concentrations (n=9).* Delayed equals 4 days between venepuncture and centrifugation of the EDTA blood samples.
Discussion
The assay performed satisfactorily in all aspects evaluated, briefly referred to in subsequent published studies and confirming and extending the manufacturer’s documentation.\textsuperscript{436, 658} The main issues to consider though going forwards really relate to the forms of OPN measured in the various assays. Although we confirmed in a small number of samples the marked serum-plasma differences in OPN concentration which is commented on in several studies and in the manufacturer’s kit insert as being due to thrombin cleavage of OPN, this is actually not clearly evidenced as definitely being the underlying mechanism when the various studies are reviewed in more detail as here. At the time we undertook this evaluation we clearly had not looked into this in as much detail as would have been ideal and had taken at face value papers stating this as being based on earlier solid evidence.

Certainly OPN is cleaved by thrombin experimentally and analysis of the amino acid sequence shows such a cleavage site to be present.\textsuperscript{560} For the R&D Systems kit used here there is no data about the epitopes recognised by the antibodies and so it is impossible to say for certain that this assay only detects the intact form of OPN, although the lower concentrations of OPN in serum may support this. Similar findings using this assay (or R&D Systems reagents and protocol in the first study cited below) have been reported before by other groups, for example with samples from healthy controls and RCC patients with 3.8-4.8 fold higher or 2.34 fold higher median OPN concentrations respectively in EDTA plasma compared with serum.\textsuperscript{559, 659} Using Western blotting to explore the differences in the serum and plasma forms recognised by one of the antibodies was not conclusive although it appeared that the monoclonal coating antibody used recognised multiple forms of OPN including the cleaved form.\textsuperscript{659} However this would not be quantified if the detection antibody was unable to recognise that fragment also which is not clear, but may result in fragments binding to the coating antibody and competing with the intact form and hence affecting measurements indirectly. An alternative commercially available assay kit from Immuno-Biological Laboratories documents the antibody specificity with the kit insert describing the coating antibody as recognising an epitope in the N-terminal region and the detection antibody binding to an epitope at the C-terminal side of the thrombin cleavage site. This therefore shouldn’t detect the cleaved form of OPN but it is very possible that the N-terminal fragment of cleaved OPN would bind to the coating antibody and competitively inhibit binding of intact OPN with unknown effects quantitatively as discussed above, certainly precluding the use of serum as indicated by the manufacturer. For plasma this may not be an
issue assuming that there is no circulating endogenous thrombin-cleaved OPN. A definitive assay for intact OPN only and not suffering from any competitive effects of the thrombin-cleaved form would have to incorporate a coating antibody recognising an epitope which spans the region of the thrombin cleavage site and includes amino acids either side and which is accessible within the 3D conformational structure of OPN under assay conditions. The production of antibodies to various domains in the OPN protein have been used in combination to design several ELISAs with different specificities for the different isoforms and truncated forms (although not as per the above suggestion) and used to characterise OPN in cell supernatants and urine samples, demonstrating that in the latter few if any cleaved forms of OPN were present. One, possibly two, of these antibodies have been used in the IBL ELISA above.

Interestingly OPN has also been reported to bind complement factor H and detection of OPN in serum samples using an in-house competitive ELISA was only possible once such complexes had been disrupted through heating in a chaotropic buffer with a reducing agent. Clearly the effects of such complex formation on OPN concentrations measured will be dependent on the antibodies used in the assay in the same way as for detection of intact versus cleaved OPN. Whether this could also contribute to the differences in serum versus plasma is not clear but clearly spiking studies using complement factor H would be worth pursuing in future studies. In addition to the presence/absence of calcium regulating thrombin activity and affecting OPN cleavage, OPN also binds calcium and it is not inconceivable that the presence/absence of calcium in serum and plasma respectively could influence the detection of OPN depending on calcium-dependent epitopes recognised by antibodies as we have described later in this chapter for NGAL. If calcium is part of the underlying mechanism in this way, a possible explanation may lie in the fact that OPN undergoes a calcium-dependent polymerisation mediated by transglutaminase II, which may affect accessibility of antibodies to binding sites and has been proposed to potentially contribute to differences in serum and EDTA plasma. Alternatively direct sequestration by the clot during formation of serum may account for the findings.

In terms of stability, multiple freeze-thawing of plasma was not found in other studies to affect OPN concentrations measured using a new multiplex assay or the R&D Systems assay, at least in the latter case until the 5th freeze-thaw cycle. However a study using the IBL ELISA reported effects of freeze-thawing for both serum and to a lesser extent plasma,
and even decreased OPN after storage at -80°C for 1 to 4 weeks although no details are provided as to the extent of these differences and whether these exceed the variability of the assay. Different storage conditions of plasma or serum prior to freezing have been reported to have little impact. Whereas delaying centrifugation of whole blood prior to serum or plasma removal, for up to an hour at room temperature or 6 hours at 4°C, did not affect plasma OPN concentrations, similar to our results over much longer time periods, decreased OPN was detected in serum measured after 15 or 60 minutes at RT, which is presumably during the clotting process.

It is difficult to understand whether serum or plasma measurements of OPN provide the optimal information clinically, particularly given the differences in fluid types used across studies and also with direct comparison of different assays yielding very different results, whether due to forms of OPN present or measured or standardisation differences. Both serum and plasma OPN have been shown to be prognostic in advanced non-small cell lung cancer for example but in mesothelioma with a matched sample comparison, plasma was superior to serum. In a study of RCC published since ours, both serum and plasma samples were used for OPN measurement depending on what was available for each patient and with independent prognostic value shown for both, although this included multiple RCC subtypes. Interestingly significantly higher OPN concentrations were found in men compared with women, whether plasma or serum, although it is not clear whether this was corrected for clinical factors such as the stage and grade mix in the two groups for example. This is under investigation in our RCC cohort, but we did not find any significant gender-specific differences in normal healthy controls and a study determining the reference range for OPN in 300 healthy individuals also found no significant effects of age or gender and a low biological within-subject variation of 8.2%. This study also found similar precision results for the R&D Systems to those in our study.

Clearly although OPN has shown value in many studies across various disease areas and findings have been shown to be reproducible in terms of clinical validity, further characterisation of the available immunoassays in terms of the possible effects proposed here would be of value and could be relatively easily achieved using spiking experiments for example with complement factor H, cleaved OPN or of serum with EDTA. This may address some of the variability across studies and provide a clear way forward as to the optimal way of determining OPN and importantly provide clarity as to which form(s) is the most relevant.
clinically in the different disease areas, allowing the potential of OPN as a biomarker to be more fully and robustly realised.

**Carbonic Anhydrase IX**

**Introduction**

Carbonic anhydrase IX (CAIX) is a 46 kDa membrane protein, shown to exist in several forms as a result of alternative splicing, proteolytic cleavage, glycosylation and phosphorylation. Playing an important role in regulating intracellular pH, allowing cell survival in hypoxic conditions for example, increased expression in renal cancer was initially recognised through positive reactivity of a monoclonal antibody (clone G250) with the antigen later being identified as CAIX/MN protein. It is now known that this upregulation is also present in other cancers and is mediated through hypoxia inducible factor-1α (HIF-1α), which in RCC is increased as a downstream consequence of the VHL gene alterations as described in Chapter 10. The use of CAIX has been investigated both in diagnostic imaging and as a therapeutic target in antibody, vaccine or small molecule inhibitor-based strategies. At the time we commenced our studies relating to CAIX, several studies had shown tissue expression of CAIX to be prognostic and soluble CAIX in serum or plasma from patients with RCC and also other cancers appeared to have both prognostic and predictive potential, although there was inconsistency across studies. Given these findings and the biological relevance related to VHL, we selected CAIX as a potential prognostic biomarker to explore further, initially in the Leeds Multidisciplinary RTB cohort which has now been published and now with planned extension to the NIHR cohort as described in the following chapter. We describe here the analytical verification of two commercially available immunoassays for CAIX and the issues encountered, which have important consequences for some published studies and highlight the importance of adequate validation in the first place and later verification. Our findings have been published in detail and are summarised here.

**Methodology**

A total of 17 sets of matched serum and EDTA plasma samples from patients with RCC of varying age, gender, stage and grade were used for the various parts of the verification study. The assays used were both commercially available sandwich ELISAs: the human MN/CAIX
ELISA kit from Oncogene Science/Siemens Healthcare Diagnostics Inc. and the Quantikine human carbonic anhydrase IX/CA9 ELISA kit from R&D Systems (USA). The initial work was carried out using the Siemens assay only, but once it became apparent that there were some problems with this assay, the R&D assay was also examined to help resolve the nature of the problems seen and also as a possible alternative to use in future studies. Samples were assayed in duplicate and the manufacturers’ protocols followed in each case. The limit of detection of the Siemens and R&D assays as quoted by the manufacturers are 2.5 and 2.28 pg/mL respectively. Assay verification was based upon our in-house protocol described earlier including elements such as imprecision, parallelism, specificity, and recovery. Following demonstration of specific issues with the Siemens assay, additional ad hoc investigations were undertaken, including direct comparisons of serum and plasma and the effects of metal ions, to determine the nature of the problems as described.

Results

Standardisation
Differences in standardisation were apparent with analysis of recombinant CAIX (rCAIX) showing a ratio of Siemens: R&D assay values of 2.8. Reciprocal analysis of standards between the assays showed average ratios of 2.51 and 3.13 supporting this, with the slight differences presumably due to differing matrices of the standards between the two assays. However analysis of 15 plasma samples showed Siemens: R&D ratios varying from 1.95 to 17.3 with figures for 15 matched serum samples varying from 0 to 7.1 indicating that with clinical samples additional factors were affecting the results.

Imprecision
Intra- and inter-assay imprecision for both assays were <10% at concentrations of CAIX spanning the standard curve, although CVs of 14.5% and 18.6% respectively were found for the lowest concentration controls analysed on the Siemens assay. For any sample duplicate, the CV was generally <5% of the mean and this was the same for both assays.

Parallelism
Issues were seen with the Siemens assay with 1 of 5 samples passing compared with 5 out of 5 in the R&D assay as shown (Figure 41).
Figure 41 Results from the assessment of parallelism of the CAIX assays. A) Siemens and B) R&D ELISAs. Open symbols = serum and closed symbols = plasma. Back-calculated concentrations are plotted against serial doubling dilutions of the samples with arrows indicating the normal working dilution used in each assay. Reproduced with permission.

Recovery and Specificity
As shown in Table 87 below, recovery was acceptable for all EDTA plasma samples tested on both the Siemens and R&D assays. Serum wasn’t tested in the Siemens assay due to emerging problems with this assay. For the R&D assay, one of five serum samples showed poor recovery of both high and low spikes. Cross-reactivity with CA II and CA XII was minimal or non-existent for both assays.

Table 87 Recovery of rCAIX spiked into EDTA plasma and serum samples. Results are shown for the Siemens and R&D assays. Reproduced with permission.

<table>
<thead>
<tr>
<th>Assay</th>
<th>Sample</th>
<th>EDTA plasma</th>
<th></th>
<th></th>
<th>Serum</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Initial CAIX (pg/mL)</td>
<td>% recovery (low spike)</td>
<td>% recovery (high spike)</td>
<td>Initial CAIX (pg/mL)</td>
<td>% recovery (low spike)</td>
</tr>
<tr>
<td>Siemens</td>
<td>1</td>
<td>516.5</td>
<td>102.2</td>
<td>88.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>830.9</td>
<td>93.3</td>
<td>81.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Measurement of CAIX and effects of metal ions

No significant difference was seen between results obtained for matched EDTA plasma and serum samples in the R&D assay (Figure 42), with a significant correlation (p<0.001; $r^2=0.998$) and a slope of 0.905. Using the R&D assay EDTA plasma concentrations ranged from 17.7 pg/mL to 482.9 pg/mL and serum concentrations were 18.2 pg/mL to 436.6 pg/mL. However, with the Siemens assay significantly higher CAIX concentrations were found in the EDTA plasma samples compared with serum (p<0.001) and although significantly correlated, (p<0.001 ; $r^2=0.961$), the slope of the line was only 0.538 with several outlying samples. EDTA plasma concentrations ranged from 34.5 pg/mL to 1476.4 pg/mL compared with <2.5 pg/mL to 770.6 pg/mL for the matched serum samples using this assay.

These differences between serum and plasma in the Siemens assay were originally assumed to be due to clotting events causing a generic reduction in measurable CAIX in serum rather than this being assay-specific as at that time we hadn’t used the R&D assays as a comparator. Addition of excess calcium to EDTA plasma to promote clot formation and effectively generating serum did reduce the CAIX concentrations measured, almost to those resulting from analysis of serum directly as shown below and supporting this possibility. However this was reversible with a marked elevation of CAIX detected when EDTA was added to the serum. Importantly this was also seen with rCAIX. These results, together with the fact that the R&D assays showed no difference between serum and plasma, clearly indicated that clotting per se was not responsible for the effects seen in the Siemens assay but that metal ions may be the cause. Magnesium and calcium were interchangeable in terms of causing
this effect and no effects were seen with calcium or EDTA addition in the R&D assay indicating an assay-specific effect.

**Figure 42 The relationship between concentrations of CAIX in EDTA plasma or serum.** Results are shown for 15 matched sample pairs assessed using either the A). Siemens or B). R&D assays. The dotted line shows the line of equivalence with a slope of 1. C). shows the effects of adding 20 mmol/L CaCl$_2$ to EDTA plasma or of adding 1.8 mg/mL EDTA to serum on the CAIX concentrations as measured using the Siemens assay. D). The reversibility of the effect is shown by sequential addition of CaCl$_2$ and EDTA to EDTA plasma or recombinant CAIX. Reproduced with permission$^{686}$. 

Introducing the Siemens capture antibody into the R&D assay did not change the R&D pattern of results in terms of serum and plasma having similar measured CAIX concentrations. However, using the R&D capture antibody in combination with the Siemens detection antibody resulted in marked differences in the values generated for serum and plasma (Figure 43), indicating that Siemens detection antibody (M75 clone) is responsible for the effects seen.$^{687}$ The most likely explanation is that this antibody recognises a metal ion-dependent epitope.
Figure 43 The effect of crossing over antibodies between assays to determine which antibody-antigen interaction accounts for the metal-ion dependent effects seen in the measurement of CAIX. Results are shown for 4 matched pairs of EDTA plasma and serum samples assayed using either the R&D assay, the Siemens assay or each of these but with the capture antibodies swapped between the assays. Absorbances were measured at 450 nm in both cases but background subtraction at 540 nm carried out in the case of the R&D systems assay. Reproduced with permission.

Discussion

No studies have been published which have undertaken verification of the available assays for soluble CAIX, despite there now being considerable interest in its potential clinical utility. Early studies, predominantly focussed on RCC, have reported elevated serum or plasma concentrations decreasing in some cases post-surgery and variable associations with tumour size, stage and prognosis. There has also been interest in other cancers and even since we undertook this validation study, particularly promising results have been obtained for CAIX within a diagnostic urinary biomarker panel for bladder cancer and further studies in RCC have been published, as reviewed in the previous chapter. Clearly our results show the need to undertake such studies so that results can be generated robustly and that the Siemens assay has not undergone adequate validation prior to becoming commercially available. In contrast the R&D assays performed well although confirmation of
the manufacturer’s sensitivity data is needed and potentially the effect of other pre-analytical factors.

The explanation of the metal ion effect we have described on the binding of CAIX to antibodies within the Siemens assay is not completely clear. Cleavage of the extracellular region of CAIX generates at least two soluble forms (50 and 54 kDa). This region contains the carbonic anhydrase (CA) catalytic domain which has 3 metal binding regions important for the catalytic activity and the V10 capture antibody recognises a conformational epitope in this domain. However our evidence supports the phenomenon seen in our studies as arising through the M75 detection antibody which recognises a linear epitope on the proteoglycan (PG)-like domain also present in the extracellular region. This could be explained though by the finding that the catalytic activity of CAIX is also regulated by binding of multiple divalent cations to this negatively charged PG domain and it is conceivable therefore that the M75 antibody binds a metal ion-dependent epitope with metal ions then either directly inhibiting binding of M75 to CAIX competitively or through induction of a conformational change. Although this markedly affects the serum results in particular, effects on plasma measurements where a chelating agent is present may also be affected to variable extents depending on the variability of the final concentrations of the EDTA for example.

This issue was not apparent from the data given in the Siemens assay kit insert, but it did show a possible trend towards lower concentrations in serum although slight and only samples from healthy volunteers were used with lower CAIX concentrations. Many published studies have used the R&D systems assay but several studies in renal, ovarian, lung and bladder cancers have been undertaken with the Siemens assay or assays involving the M75 antibody and the results may therefore be questionable. As the M75 antibody clone is also used in many immunohistochemical studies, it is possible that results obtained may depend on the buffers used and the presence/absence of metal ions. This effect of metal ions on epitope availability and analyte measurements is relatively rare with possible examples, although not necessarily through the same mechanism, including the calcium dependence of calretinin measurements and S100A12. Our results clearly demonstrate the need for careful independent verification of commercially available immunoassays, even from large diagnostics companies. As far as we are aware the Siemens assay used in this study is still available although from a different source, Nuclea Biotechnologies, as around
the time of this study Wilex acquired the assets of Oncogene Science from Siemens Healthcare Diagnostics and in 2013 Nuclea Biotechnologies Inc acquired Wilex Inc. It is not readily apparent that the assay available currently has been modified in any way from the one we used and it certainly uses the M75 antibody. Nuclea have a major interest in CAIX and have recently announced that their CAIX automated IHC kit has been granted FDA CLASS I IVD status which will be of potential utility in several cancer types although particularly in the determination of aggressiveness of head and neck squamous cell carcinomas. The question of whether it is the responsibility of the manufacturer or end-user to ensure assays are validated and verified more robustly was raised in an editorial focussing on our study and it may be that the answer lies in more open partnerships between the manufacturers and researchers at early stages of biomarker research studies.696

Neutrophil Gelatinase-Associated Lipocalin (NGAL)

Introduction
One of the most promising emerging kidney-related biomarkers is neutrophil gelatinase-associated lipocalin (NGAL), in particular in acute kidney injury (AKI) which is now recognised as a major health problem.697, 698 Originally isolated from neutrophils in 1993, NGAL was proposed as a novel urinary biomarker of ischaemic kidney injury produced predominantly by proximal tubule cells ten years later, although subsequently an NGAL reporter mouse model has supported a distal tubule cellular origin.699-701 Now known to be expressed by several tissues, the predominant form released by renal tubules is a 25kDa monomer whereas the main neutrophil-derived form is a 45 kDa homodimer.702 In addition, NGAL exists as a 125 kDa covalently complexed heterodimer with matrix metalloproteinase-9 (MMP-9) and other forms have been found.699, 703-705 Key functions are in protecting against MMP-9-mediated degradation, a role in bacteriostasis through mediating shuttling of iron through binding siderophores, and multiple effects on cell proliferation, differentiation and apoptosis.702 Our interest in NGAL was stimulated by the rapidly growing level of interest surrounding NGAL in renal diseases with promising results of urinary, plasma or serum concentrations of NGAL being superior to creatinine diagnostically or prognostically in AKI in several studies when reviewed at that time, and also with urinary NGAL outperforming kidney injury molecule-1 (KIM-1), cystatin C, interleukin-18 and liver-type fatty acid binding protein in a large prospective study in the emergency department setting.706-709 Of
particular interest was emerging evidence of an association with delayed graft function (DGF) following renal transplantation.\textsuperscript{710, 711} However it was also apparent even then that to provide definitive answers to its potential clinical utility, bigger prospective studies needed to be pursued and the impact of factors such as background levels of chronic kidney disease (CKD) on NGAL determined. Additionally, there was a lot of uncertainty regarding the assays used and form of NGAL being measured, with only a small number of limited studies having been undertaken.\textsuperscript{712-715} To contribute to the evidence, we undertook a study independently examining 5 of the commercially available assays for NGAL, two CE-marked \textit{in vitro} diagnostics (IVD) and three research-use only (RUO) ELISAs, using urine samples as most assays had only been validated for that matrix. This study has been published \textsuperscript{716} and the findings are described in brief below.\textsuperscript{716}

\textbf{Methodology}

Mid-stream urine samples from patients with AKI, RCC), renal stones, recurrent urinary tract infection (UTI), diabetic albuminuria or healthy controls (to ensure a variety of matrix backgrounds and NGAL concentrations; \textit{n}=78 in total) were banked for this study after processing according to local SOP. The CE-marked assays evaluated were the NGAL Test\textsuperscript{TM} from BioPorto Diagnostics A/S (Denmark), a particle-enhanced turbidimetric immunoassay which we used on the Siemens ADVIA\textsuperscript{®} 1800 platform, and the ARCHITECT\textsuperscript{®} Urine NGAL assay from Abbott Laboratories (USA), a two-step chemiluminescent microparticle assay which we used on the ARCHITECT\textsuperscript{®} i2000SR analyser. The RUO assays evaluated were all sandwich ELISA format for human NGAL and were the NGAL ELISA (HK330) from Hycult Biotech (Holland), the NGAL ELISA Kit 036 from BioPorto Diagnostics and the Quantikine Lipocalin-2/NGAL (DLCN20) Immunoassay from R&D Systems (USA). All assays were performed according to manufacturers’ instructions. Assay verification was based upon our in-house protocol described earlier with assessment of imprecision, parallelism, recovery, selectivity, limit of quantitation, haemoglobin interference, and high-dose hook effect, as described in detail.\textsuperscript{716} Measurements were performed in singlicate on the ARCHITECT\textsuperscript{®} and ADVIA\textsuperscript{®} platforms and in duplicate on all ELISAs and CVs of <10% within replicates were considered acceptable.
Results
Our detailed findings are summarised below and in Table 88. As we found a much higher LLOQ with the BioPorto/ADVIA® assay than expected from the manufacturer’s specification, we were unable to investigate recovery and effects of NGAL/MMP-9 complex and haemoglobin with this assay as the endogenous NGAL concentrations were below the LLOQ determined in practice and so the baseline values could not be used.

Imprecision
Intra-assay imprecision was generally acceptable across all platforms with one exception of a CV of 34.4% with the low NGAL QC urine sample in the Hycult assay. Problems with the Hycult assay only were also seen for inter-assay imprecision which was unacceptable and this assay also showed poor agreement between replicates.

Parallelism
As shown in Figure 44, parallelism was demonstrated for the Abbott and R&D Systems assays but not for the Hycult assay and for 2/3 samples in the BioPorto ELISA. An issue with one sample was also seen on the BioPorto/ADVIA® assay, although the issues relating to the LLOQ meant this was inconclusive.

LLOQ
The LLOQ values for the Abbott assay and R&D Systems, BioPorto and Hycult ELISAs were 5, 2, 5 and 20 ng/mL respectively, including any dilution factors. The LLOQ was 150 ng/mL for the BioPorto/ADVIA® assay which was much higher than expected given the manufacturer’s figure of 25 ng/mL for the lower end of the measuring range.

Recovery
A stock solution of recombinant NGAL (rNGAL; 1 mg/mL with ~80-90% monomeric form) gave quite different results across the assays with near-quantitative results for the BioPorto ELISA but 67-75% of this with all other assays and only 31% with the Hycult assay. This may reflect differences between the assays in specificity for the different NGAL forms or in standardisation. To overcome this, the recovery of spiked rNGAL material in urine samples was related to the assay-specific assigned concentrations for the rNGAL stock, essentially allowing the determination of relative recoveries and allowing assays to be compared.
Acceptable recoveries were found with the Abbott assay and R&D Systems, BioPorto and Hycult ELISAs (with the exception of a recovery of 73.6% for one sample in the latter).

**Figure 44** Results from the assessment of parallelism for each of the five assays, comparing dilution-adjusted NGAL concentrations (log scale) against serial double dilutions of each of three samples represented by different colours. The initial dilution factor used is indicated next to each sample dilution. Fine dashed lines illustrate ±15% of the mean for each sample. “<LLOQ” highlights sample dilutions that fall below the LLoQ determined by this study. This represents Figure 1 from our published results reproduced with permission.
Table 88  Summary of the performance data for the five NGAL assays evaluated using urine samples. This represents an adaptation of Table 1 in our published results where full details of all results and spike concentrations can also be found in Supplemental Data Table 1.

<table>
<thead>
<tr>
<th>Test</th>
<th>BioPorto/ADVIA®</th>
<th>Abbott</th>
<th>R&amp;D Systems</th>
<th>BioPorto ELISA</th>
<th>Hycult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assay standards range (ng/mL)</td>
<td>150-5000</td>
<td>10-1500</td>
<td>0.156-10</td>
<td>0.01-1</td>
<td>0.4-100</td>
</tr>
<tr>
<td>Dilution factor range</td>
<td>N/A</td>
<td>N/A</td>
<td>1/25-1/800</td>
<td>1/500-1/4000</td>
<td>1/20</td>
</tr>
<tr>
<td>Imprecision %CV Range (median)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intra-Assay (n=5 for each of 4 QC samples)</td>
<td>0.7 – 3.0 (2.0)</td>
<td>0.8 - 10.8 (1.4)</td>
<td>3.2 – 4.4 (3.6)</td>
<td>0.6 – 5.1 (3.4)</td>
<td>5.8 - 34.4 (8.8)</td>
</tr>
<tr>
<td>Inter-Assay (n=10 for each of 4 QC samples)</td>
<td>1.9 - 7.9 (3.7)</td>
<td>4.8 - 9.9 (7.6)</td>
<td>3.2 - 10.1 (7.1)</td>
<td>6.4 - 15.8 (12.6)</td>
<td>26.1 - 33.3 (30.2)</td>
</tr>
<tr>
<td>% Parallelism Range (median; n)</td>
<td>1.9a</td>
<td>2.2 - 5.8 (3.0; n=3)</td>
<td>1.9 - 7.9 (3.6; n=3)</td>
<td>3.2 - 49.8 (44.6; n=3)</td>
<td>17.8 - 30.2 (21.5; n=3)</td>
</tr>
<tr>
<td>% Recovery Range (median; n)</td>
<td>Not determineda</td>
<td>88.6 - 99.1 (95.6; n=8)</td>
<td>93.5 - 106.7 (98.9; n=9)</td>
<td>100.6 - 113.4 (104.1; n=8)</td>
<td>73.6 - 95.2 (88.1; n=8)</td>
</tr>
<tr>
<td>Specificity</td>
<td>Not determineda</td>
<td>No Effect</td>
<td>No Effect</td>
<td>No Effect</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>+ Complex</td>
<td>Not determineda</td>
<td>No Effect</td>
<td>No Effect</td>
<td>No Effect</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>LOQ (including sample dilution factor)</td>
<td>150 ng/mL</td>
<td>5 ng/mL</td>
<td>0.078 (2.0) ng/mL</td>
<td>0.01 (5.0) ng/mL</td>
<td>1 (20) ng/mL</td>
</tr>
<tr>
<td>Haemoglobin Interference</td>
<td>0.75, 1.125, 2.25 µg/mL (+/++/+++)</td>
<td>Not determineda</td>
<td>No Interference</td>
<td>No Interference</td>
<td>Interference</td>
</tr>
<tr>
<td>5.0 mg/mL</td>
<td>Not determineda</td>
<td>Interference</td>
<td>Interference</td>
<td>Interference</td>
<td>Interference</td>
</tr>
<tr>
<td>Hook Analysis</td>
<td>Hook Effect</td>
<td>No Effect</td>
<td>No Effect</td>
<td>No Effect</td>
<td>Inconclusive</td>
</tr>
</tbody>
</table>

aNot determined or determined in limited samples due to endogenous NGAL concentrations being <LLOQ as determined by this study.
**Selectivity**
None of the assays detected the rNGAL/MMP-9 complex and neither rMMP-9 nor the complex affected the urinary NGAL results except for the Hycult assay where variable effects were seen.

**Haemoglobin Interference**
The Hycult assay was affected by haemoglobin at all concentrations tested but all other assays were only affected at the highest concentration of haemoglobin of 5 mg/mL.

**Hook Effect**
A hook effect was absent from the Abbott assay and the R&D Systems and BioPorto ELISAs. However, using the BioPorto/ADVIA® assay, a typical high dose hook effect was seen, although only at NGAL >~70,000 ng/mL. With the Hycult assay an atypical effect was seen where the assay appeared to plateau at NGAL concentrations of 20-25 ng/mL, although the range of the standard was up to 100 ng/mL (Figure 45).

**Inter-assay NGAL comparison**
Using the BioPorto/ADVIA® and Hycult assays, over 50% of the urine samples were below our determined LLoQ values and so couldn’t be included in comparison results involving those assays. From the modified Bland-Altman plots (Figure 46), it can be seen that there was generally good agreement between the Abbott and R&D Systems assays with only two samples falling outside 95% of overall bias, and a mean bias of 14%. A Passing-Bablok analysis for comparison of these two assays only, demonstrated that data lay on the line of equality. Good agreement between the BioPorto ELISA and the Abbott and the BioPorto/ADVIA® assays was also seen with mean % bias of 13% and 23% respectively (the latter case only included 22 samples). Results with the two BioPorto assays tended to be higher than with other assays. The Hycult assay showed a marked negative bias, with 29 of the 32 samples above the LLoQ showing NGAL concentrations <50% of those measured on the Abbott assay with some values even being <20%. With the Hycult assay, no urine samples demonstrated NGAL concentrations >20 ng/mL (400 ng/mL corrected for dilution) which mirrors the plateau effect of this assay shown in the hook effect studies.
**Figure 45 Results from the hook effect analysis.** A: Point-to-point line illustrates high dose hook effect seen in the BioPorto (IVD) assay with dotted line showing the upper limit of the assay range; B. Hycult assay data illustrating plateauing/saturation effect with dotted line showing the upper limit of the assay range. This represents Figure 2 from our published results 715.

![Graph A: BioPorto/Advia](image1)

![Graph B: Hycult](image2)

**Figure 46 Modified Bland-Altman Plots for various comparisons of the NGAL assays.** Samples <LLoQ have been omitted from the relevant plots. A). comparisons of the Abbott assay, the BioPorto/ADVIA® assay and the R&D Systems and BioPorto ELISAs, and b). the comparison of the Hycult ELISA assay with each assay, with an expanded y axis to incorporate the differences seen. The solid red line indicates no bias, with the dotted lines indicating the mean bias and the limits of agreement (mean difference +/- 1.96 SD of the differences). In each case the difference referred to in the y axis is the second assay in the plot title subtracted from the first named assay. This represents Figure 3 from our published results 716.
Discussion

Our independent verification findings support the Abbott and R&D Systems assays as having been validated adequately by the manufacturers and producing comparable results. A single published study evaluating the Abbott NGAL assay and sample stability had previously reported excellent reproducibility and precision although not examining other aspects such as recovery, linearity and selectivity, and found functional sensitivity (based on imprecision of <20% alone in this case) to be <2 ng/mL which is similar to our determination of an LLoQ of <5 ng/mL. Acceptable variability of the Abbott assay was also reported in a further small study. Excellent performance characteristics had been previously reported for the BioPorto ELISA in an extensive verification study with both urine and plasma, although issues were reported for inter-batch variability and some evidence for non-parallelism. The latter was not as pronounced as that seen in our study and this may reflect study-specific differences such as the urinary matrices (paediatric versus adult), dilution factors or NGAL forms present. The fundamental reasons for the saturation of the Hycult assay and essentially rendering it unusable are not clear but again may reflect a combination of differences in standardisation and forms of NGAL recognised and potentially interfering.

Although broadly comparable with other assays, we found the BioPorto assays were biased towards higher concentrations, although not as marked as the 65% bias previously reported in a comparison of the BioPorto assay on the Beckman Coulter platform with the Abbott assay. Marked biases were also seen in a study comparing two Bioporto assays with the Abbott Diagnostics assay. Whether such results arise due to differences in standardisation, assay design or NGAL forms measured (including possibly glycoforms) is not clear but we saw this with both recombinant and endogenous NGAL and even with assays with good agreement, differences were apparent in some samples. Clearly the forms of NGAL detected will depend on the antibodies used in the various assays but our results demonstrated that the NGAL/MMP-9 complex was not detected by and didn’t interfere with measurement of NGAL in the Abbott, BioPorto and R&D Systems assays. The relative specificities of the assays for the monomeric and dimeric forms of NGAL were not determined, although a study reporting a significant
association of monomeric forms with tubulointerstitial fibrosis in CKD found a significant correlation of urinary NGAL results obtained using the Abbott assay with the monomeric form shown by Western blotting. However, it doesn’t appear that a systematic analysis of the absolute specificity for this form had been undertaken.\textsuperscript{704} Assays able to robustly differentiate between the NGAL forms would be useful in terms of determining potential contributory sources and providing further clinical insight and this is being addressed by several companies currently. For example, predominantly homodimeric NGAL is found in urine from patients with urinary tract infection compared with monomeric NGAL in patients with AKI and mainly elevations in homodimeric NGAL in urine in patients following cardiac surgery in the absence of AKI suggesting a predominant activated neutrophil source.\textsuperscript{705, 718} Indeed in a cohort of 5599 individuals from the general population recruited into the Copenhagen Heart Study, plasma NGAL measured using an in-house assay, was significantly associated with several inflammatory indices, particularly neutrophil count and CRP, in addition to showing an inverse association with eGFR, and was independently associated with outcome irrespective of eGFR.\textsuperscript{721} Interestingly, similarly elevated plasma NGAL concentrations compared with healthy individuals were found in anephric and anuric patients on dialysis, providing support for the elevated NGAL in CKD being predominantly extra-renal although the balance of increased production against decreased clearance to account for the elevated concentrations is not known.\textsuperscript{722}

The most surprising finding was the marked disparity between the manufacturer’s reported measuring range with the CE-marked BioPorto/ADVIA\textsuperscript{®} assay and our LLoQ findings which essentially placed the LLoQ at the suggested optimal cut-off value for NGAL in diagnosing AKI, a situation which is far from ideal given the inherent greater variability in that area.\textsuperscript{707} This may well be platform-specific as a previous study had reported performance of this assay on the Beckman Coulter AU5822 platform to be acceptable, although LLoQ, hook effect and recovery were not investigated and also on the Roche Cobas 6000 and Hitachi 917 platforms, although consistently higher results were obtained for EDTA plasma on the former.\textsuperscript{713, 723} Following our study, BioPorto issued a new version of the application note for the Advia platform with revised measuring range and LLoQ included amongst the changes.
The significance of the interference of very high concentrations of haemoglobin in all assays (although a previous study had not found this for the Abbot assay) isn’t clear and the extent to which such high concentrations are found clinically in urine needs to be investigated.\(^{712}\) Certainly haemolysis of whole blood is known to affect plasma NGAL measurements, although this could potentially reflect neutrophil-derived NGAL or interference by haemoglobin directly.\(^{714}\) The importance of such technical and biological pre-analytical effects are beginning to be recognised for NGAL with an association of leukocyturia with higher NGAL concentrations, much higher concentrations of NGAL being present in serum compared with matched EDTA plasma samples most likely due to release of NGAL from neutrophils which is of particular importance when historical studies are compared, and age- and gender-related effects on urinary NGAL being reported.\(^{724-728}\) Stability during processing and storage do not seem to be an issue under the conditions examined.\(^{729, 730}\)

The aspects of NGAL measurement highlighted in this Chapter are crucial to an interpretation of studies examining the clinical validity and utility of NGAL. However the extent to which studies or reviews examining the clinical potential of NGAL consider such aspects are highly variable with, for example, a recent review of the broader clinical applications of NGAL only briefly mentioning such aspects, an earlier meta-analysis of studies examining the potential of NGAL in AKI diagnosis including a consideration of the assays used and a recent review focusing on NGAL in predicting AKI providing a much more comprehensive overview including a tabulated summary of the assays used in each study.\(^{697, 698, 707}\) This latter review, whilst supporting the promise of NGAL, also flagged up the issue that many of the studies had inadequate study design and failed to follow guidelines such as the STARD criteria, issues to address if NGAL is to fulfil its potential.\(^{322}\) Indeed whilst agreeing that NGAL is a promising biomarker the idea that it is the “troponin of the kidney” has recently been used as an example of one of the “false myths and legends” in laboratory diagnostics and further analytical and biological insights including understanding of the impact of comorbidities and specific assays for kidney-derived NGAL are needed to allow rational evaluation and optimal use.\(^{702, 708, 731}\)
Vascular Endothelial Growth Factor (VEGF) - Relative value of serum or plasma and QC aspects

Introduction
VEGF-A, often referred to as VEGF although there are several family members, is a major angiogenic cytokine. Existing as several isoforms generated by alternative splicing, there has been considerable interest in VEGF in cancer particularly, given its pivotal role in regulating angiogenesis. With increased understanding of the underlying mechanisms involving hypoxia and HIF-1α in upregulating VEGF, numerous therapies have been developed targeting this pathway and are in use across many cancer types. This has particularly been the case in clear cell RCC given the widespread inactivation of the VHL gene and consequent stabilisation of HIFs and increased expression of VEGF and where conventional chemotherapy is ineffective. Accompanying this has been a raft of studies exploring the possible use of VEGF as a biomarker, in particular for either prognostic use or in terms of predicting response to VEGF-related therapies. In renal cancer, several studies have reported predictive uses for VEGF or its receptors although studies are small and heterogeneous and require further confirmation. As reviewed in the previous Chapter, VEGF has also shown promise in RCC prognostically although complicated by studies using either serum or plasma where concentrations vary markedly. When we and others first reported such serum-plasma differences in VEGF, this was proposed to be due to platelets containing and releasing VEGF and subsequently confirmed in several studies. Based on the evidence reviewed in Chapter 12, VEGF was included as one of the prioritised biomarkers for the RCC prognostic study undertaken as part of this Programme and we included both plasma and serum VEGF in order to determine which if any provides clinically useful information, particularly given the impact of processing as reviewed extensively below. The results of the prognostic study are described in Chapter 14 but we describe here some of the technical aspects of the measurements.

Methodology
As described in the following Chapter, VEGF concentrations were determined for the RCC patients in the prognostic cohort. Matched serum and EDTA plasma samples were analysed for each patient as available with a total of 430 patients having both sample types with detectable VEGF in both. Samples were analysed using the Human VEGF Quantikine kit from R&D Systems which is a sandwich ELISA specifically measuring VEGF-A. This assay is one of the most widely used commercially available assays for VEGF and we have previously carried out validation studies of this assay some years ago, both when evaluating the importance of blood sample handling and also when describing the existence of a novel soluble VEGF receptor variant in amniotic fluid.  

All aspects evaluated including parallel dilution, within- and between-run precision and recovery were acceptable and similar findings were reported in subsequent studies using this assay also. It was apparent that the assay measured the free form of VEGF and not VEGF complexed with the receptors and although specific for VEGF-A, both VEGF121 and VEGF165 isoforms are detected. The manufacturer’s protocol was followed with all samples being analysed in duplicate and low, medium and high rVEGF controls were also analysed. On inspection of the data some results appeared anomalous in terms of almost no difference in values between serum and plasma. To investigate this further, additional data analysis was carried out and following this some of the samples were subjected to additional measurements of calcium and potassium undertaken using routine assays in the Leeds Teaching Hospitals Clinical Chemistry laboratories to confirm the presence of potassium EDTA in the purported plasma aliquots and its absence in the serum aliquots. In addition to assess stability during a freeze-thaw cycle, matched aliquots of serum and plasma from 20 patients with RCC were thawed at room temperature and after approximately an hour were refrozen. These 40 aliquots were then analysed together with 40 matched aliquots which had been stored frozen without any additional freeze-thaw step.

Results

As expected given the platelet-derived VEGF being released during clotting, serum VEGF concentrations were higher than matched plasma samples in most cases. The range for plasma VEGF was 5.2-1480.9 pg/mL (median 67.8) with corresponding values for
serum VEGF being 9.9-4283.3 pg/mL (median 348.4), with one patient providing the sample with the unusually low values for both serum and plasma VEGF (i.e. 9.9 and 5.2 pg/mL) which was extremely low and almost at the limits of the assay (manufacturer’s LoD 9 pg/mL and 10.9-12.7 pg/mL) and will be reanalyzed and the functional sensitivity of the assay determined in our hands. What was striking however was a number of samples where there was very little difference between plasma and serum VEGF and a small number where the plasma concentrations were higher than serum. This is depicted below in Figure 47.

**Figure 47 Frequency distribution of plasma VEGF concentrations as a percentage of the serum VEGF for 426 patients with RCC.** X axis labels indicate the centre of each column. Results are shown for n=426 rather than 430 as 4 of the most extreme values (210% to 474%) were omitted to be able to show the spread of the majority of samples optimally.

When the results were examined further, there were several apparent patterns. For 4 patients (not shown in Figure 47) it appeared that the plasma and serum aliquots may have been switched either at the time of processing at the four specific sites involved or
during analysis as there were marked differences between them but in the opposite direction to that expected., for example 1480.9 pg/mL versus 312.2 pg/mL for plasma and serum respectively. For a further group of 48 patients (from the column centred around 80% upwards), plasma and serum samples were much more similar and in many cases differ by <30 pg/mL (the range of the standard curve in this assay covers from 15.6 to 1,000 pg/mL). This, together with i). the relatively low values for the serum VEGF concentrations in this group (27.5 to 579 pg/mL, median 98.6 pg/mL) compared with the group as a whole as above; ii). the lack of significant difference between serum VEGF concentrations in this group of 48 patients (27.5 to 579 pg/mL, median 98.6 pg/mL) and their matched plasma samples (25.0 to 621 pg/mL, median 92.4 pg/mL; p=0.169); iii). the lack of recorded deviations from the protocol in terms of processing causing possible spurious results, and iv). the fact that 44 of these 48 patients had been recruited at one specific site, led to the hypothesis that most if not all of these had actually had no clotted sample collected for serum at recruitment, but two sets of anticoagulated blood for plasma. This was further supported by the Figure shown below. By taking the average of the two measurements (i.e. “serum” and plasma) for each patient and subtracting it from the “serum” result (in this case the “serum” but the same result would be obtained for plasma), and expressing this difference as a % of the average, this could be examined in relation to the known variability of the assay. In Figure 48, essentially all these 48 samples make up the columns with centres from 10% down to -20% as indicated, the majority of which therefore are within the variability of the assay (9.1-9.5% inter-run CVs in our study) and therefore likely to have been determined from two aliquots of the same sample type i.e. plasma in this case. The most extreme columns from -40 to -70% represent the 4 samples where the most likely explanation is direct switching of serum and plasma at some point.

**Figure 48** Frequency distribution of the difference between serum VEGF and the average of the matched serum and plasma VEGF results for each of 430 RCC patients expressed as a % of that average. Where serum and plasma results are the same, the value on the x axis would be 0. X axis labels indicate the centre of each column. The red arrows indicate the bins containing the 48 samples where both samples may be plasma.
Biochemical analysis of some of the samples identified as having potential problems showed an absence of calcium in the “serum” samples and high levels of potassium, consistent with the samples actually being EDTA plasma. Subsequent biochemical analysis of 235 serum samples from RCC patients for the prognostic study (run to provide some missing routinely measured analytes) found such results for calcium and potassium in 40 samples, all of which were from this recruitment site. Enquires at the site where the 44 patients had been recruited from established that they had been unaware of any issues and had used Greiner tubes with purple and red tops for EDTA plasma and serum respectively. However on close inspection of the tubes used at that site, it was apparent that the purple top was a K₂EDTA tube as expected but the red top used was actually K₃EDTA rather than a red top z serum clot activator tube, differing purely by a purple band around the top of the patient label compared with a red band and the small print on the label showing K₃EDTA as shown (Figure 49).

Figure 49 Greiner blood collection tubes showing the differences and overlap in colour closures and labelling (top) and the printed indication of additives (bottom).
1). $K_2$EDTA for crossmatch, 2). $K_3$EDTA, 3). $K_3$EDTA, 4). $K_3$EDTA, 5). Z serum clot activator. Tubes 3 and 5 were the intended tubes to be used for EDTA plasma and serum respectively but one centre had inadvertently used Tube 4 instead of Tube 5 for “serum” collection.

Although we only have definitive evidence for this being the case in 40 samples, a total of 46 patients recruited at this site had been included in the VEGF sample analysis for the prognostic study. In addition to the 44 samples known or suspected to be EDTA plasma rather than serum, one sample appeared to have actually been taken into the correct tube types with values of 1118.0 and 42.5 pg/mL for serum and plasma respectively (this was actually the first patient recruited at the site) and one sample was the one with almost undetectable VEGF in both serum and plasma. At this stage until all 46 samples are checked through analysis of calcium and potassium, all “serum” VEGF results from patients recruited at this site have been removed from the prognostic analysis described in the following chapter, as a precaution. The four samples from other sites which were part of the group of 48 as described above have remained within the study at present as in all cases serum VEGF was higher than plasma values with plasma representing 76-93% of serum and therefore there may not be an issue and in fact measurement of potassium and
calcium concentrations for the sample with VEGF results of 64.3 and 68.9 pg/mL for plasma and serum respectively, confirmed the correctness of the plasma and serum attribution. The other three will also be checked.

For the 380 patients with no apparent issues with the plasma and serum VEGF measurements and matched results available, serum VEGF ranged from 61.6 to 4283.3 pg/mL (median 377.2 pg/mL) and was significantly different (p<0.0001) to the plasma VEGF with values of 9.2 to 419.5 pg/mL (median 67.5 pg/mL). As shown in Figure 50 below, serum VEGF was significantly correlated with plasma VEGF (Spearman r=0.594; p<0.0001) and markedly so with serum minus plasma VEGF (Spearman r=0.977; p<0.0001) which may not be completely unexpected given that in 50% of the patients, the plasma VEGF component is 20% or less of the total i.e. the serum VEGF.

**Figure 50 Relationship between serum VEGF and plasma or serum minus plasma VEGF concentrations (n=380).** Results are shown for all samples (left hand plots) or omitting 8 samples with serum VEGF >1500 pg/mL (right hand plots) to allow expansion of the x axis and more detail to be seen.
For 282 of these patients, platelet counts were available either on the day of the biomarker blood sample or within 2 days of this. Platelet counts were highly significantly correlated with VEGF concentrations, particularly the serum or serum minus plasma VEGF, with correlation coefficients of 0.57 and 0.56 respectively. Assuming the serum minus plasma VEGF to be derived from platelets, the calculated VEGF content per platelet is highly variable, ranging from 0.02 to 5.77 pg/10^6 plts with a median value of 1.14. Of note this wasn’t corrected for the hematocrit which we have previously advocated as this data wasn’t available at this time.\textsuperscript{572}

For the stability study, there was a significant difference between the frozen and freeze-thawed plasma samples analysed using the Wilcoxon matched pairs signed rank test (range 15-270 pg/mL compared with 11.0 to 231.0 pg/mL respectively, median 77.0 pg/mL for both; p=0.004), but not the serum (p=0.065) or the serum minus plasma results (p=0.447) as shown in Figure 51 below. However although this was statistically significant, many plasma samples had concentrations of VEGF at the bottom end of the standard curve where variability was higher and only in 6 cases did the decrease exceed 10%. In 6/20 cases for plasma and 7/20 for serum the results were either the same or higher in the freeze-thaw samples. This requires further investigation in a larger number of samples and with more samples with higher VEGF concentrations.

\textit{Figure 51} Comparison of VEGF concentrations of matched plasma and serum samples stored frozen and thawed immediately prior to analysis, with paired aliquots which had been subjected to an additional freeze-thaw cycle. Samples from 20 patients with RCC were used with storage at -80\textdegree C. Aliquots were thawed at room temperature.
Discussion

Pre-analytical considerations are of pivotal importance in clinical chemistry labs and an area of growing awareness and concern as reviewed earlier in the Chapter. With many biomarker studies being undertaken in research labs, such aspects are often overlooked and not considered either in the study design phase or during evaluation of the biomarkers and this will increasingly contribute to the lack of consistency between published reports and potentially be a major barrier to progress. A major finding here was the inadvertent use of the wrong blood tubes at one site which resulted in no serum being banked but two samples of EDTA plasma. This was detected both through VEGF measurements not fitting the usual pattern of serum versus plasma and also the low concentration/absence of calcium and elevated potassium levels in those samples. A further three samples may also have been inadvertently switched and this will be investigated further. Although many analytes can be measured equally in plasma or serum, the findings here with VEGF illustrate the importance of the correct tube type and specimen type if reliable biomarker results are to be achieved. If VEGF analysis of both
serum and plasma samples had not been undertaken and if none of the clinical chemistry analysis had been necessary, this may not have been apparent and the resultant data analysis could have led to additional inconsistent results in the literature. This occurred despite specific efforts with site inductions to ensure the correct tubes were used and indeed it appears as though the first sample taken at this site may have actually be taken into the correct tubes. Whether or not measurement of simple analytes such as potassium or calcium or fibrinogen should be undertaken routinely in samples associated with research biomarker studies in clinical trial for example is a consideration. This would confirm the sample type as plasma or serum but unfortunately would still not control for deviations from sample processing protocols in terms of processing time delays or storage conditions for example. CLSI standards and guidelines exist for many pre-analytical areas including identification of patients and samples, venepuncture, sample processing. There is also one covering all aspects of the blood tubes including construction material, additives and labels and although earlier versions also included aspects relating to colour of the tube closures, this has been omitted from the current version. Clearly similar colour tube caps for different tube types and with the nature of the tube additives only indicated otherwise in small print on the label and by a small different coloured band can easily lead to errors. Calls for harmonization, given this heterogeneity, have been made by the EFLM and although difficult given the multiple manufacturers, it is hoped this will be achieved in the future.

The occurrence of thrombocytosis in many cancers has been recognised for over a hundred years and its association often with shorter survival has been reviewed. However, whether this is an epiphenomenon reflecting systemic elevation of cytokines such as IL-6 or other tumour-derived thrombopoietic factors, or through a direct involvement in cancer progression, for example through physical interactions with tumour cells or the production of platelet-derived cytokines or growth factors, is still not clear. However over the last few years it has become increasingly apparent that the sequestration of VEGF by platelets is a major determinant of the results seen in many studies examining VEGF as a potential biomarker. We found serum VEGF to be significantly correlated with platelet number in line with many other studies, for example. In a meta-analysis of the studies it has been estimated that the total
platelet concentration of VEGF far exceeds the circulating amount in plasma in cancer patients by almost 30-fold with leukocyte-associated VEGF accounting for a much smaller amount and tumour tissue being one of the main sources.\textsuperscript{733} Interestingly however, skeletal muscle was calculated as having the largest reservoir of VEGF. In an animal study investigating this further similar findings were reported to the importance of platelets in terms of their accumulation of VEGF but also VEGF-impregnated pellets implanted sub-cutaneously or microscopic xenografts resulted in increased platelet VEGF but not plasma VEGF supporting a role as actively sequestering such angiogenic factors with implications for underlying biology and therapies.\textsuperscript{742}

The literature regarding VEGF as a biomarker is very mixed with differences in whether serum or plasma have been used and which type of plasma, the blood collection methods and sample processing/storage and therefore it is not surprising that there is a lack of consistency generally about the utility of VEGF as a biomarker.\textsuperscript{733, 743} Importantly however, the majority of studies have used the same ELISA kit as in our studies allowing studies to be more usefully compared. The consideration of which anticoagulant to use if circulating endogenous levels of VEGF are to be determined accurately is important. Our initial studies comparing serum and plasma were based on only four healthy volunteers where in terms of measuring endogenous circulating VEGF we found citrate plasma to be optimal with lowest VEGF concentrations, presumably due to least platelet activation and this has been confirmed subsequently.\textsuperscript{542, 738} In most cases little or no difference was seen comparing citrate with EDTA plasma if samples were processed within 30 minutes or with similar delays but samples kept on ice.\textsuperscript{542, 543, 739} However, we reported that EDTA anticoagulated blood appeared less stable over time with delayed processing up to 4 hours leading to markedly increased plasma VEGF in 2/4 cases for EDTA and to a lesser extent in 1/4 cases for citrate plasma. This may represent the worst case scenario with blood being taken into a syringe before distribution into anticoagulant-containing tubes due to the volumes needed and hence possibly leading to more platelet activation than would have occurred with blood collected directly into anticoagulant-containing tubes.\textsuperscript{542} In our biobanked RCC blood samples overall, 97\% were centrifuged within 2 hours of venipuncture with a median time of 1 hour 11 minutes. Average increases of 28-34\% have been reported for EDTA blood delayed in processing at room temperature
for 1 hour and 64-80% for 2 hours although no changes were seen if left at 4°C for prolonged periods.\textsuperscript{738, 739} Centrifugation speeds have also been highlighted as being important presumably in terms of generating platelet-poor plasma, together with avoiding sampling the plasma immediately above the buffy coat which we also adopt in our SOPs.\textsuperscript{738, 739}

However, it is now apparent that even with citrate, platelet activation occurs and VEGF is released as studies adopting a very stringent protocol to avoid any platelet activation in vitro, monitored by concurrent measurement of PF4, have found that plasma from blood collected into tubes containing either CTAD (sodium citrate, theophylline, adenosine, dipyridamole) or Edinburgh mixture (EDTA, PGE1, theophylline) with rapid processing at 4°C contained much lower VEGF concentrations, and with even more platelet activation if the citrated tubes were maintained at room temperature compared with 4°C.\textsuperscript{743} This study also reported that serum VEGF did not plateau until at least 2 hours after venepuncture although plain glass tubes were used rather than clot activator tubes as in our study and where we have shown clotting in terms of peptide fragmentation is essentially complete within 60 minutes of venepuncture.\textsuperscript{743, 744} A study comparing PECT tubes containing a mixture similar to that of the Edinburgh mixture above and with samples collected without tourniquet and at 4°C with citrated plasma collected with tourniquet and at room temperature comparing healthy controls and patients with metastatic RCC or other cancers found significantly higher VEGF levels in citrated plasma compared with PECT plasma.\textsuperscript{745} Additionally using citrated plasma the VEGF was higher in both RCC (and similar to the EDTA values in our study) and non-RCC groups compared with controls whereas with PECT samples only the RCC patients were higher.\textsuperscript{745} PF4 concentrations were higher in all citrate samples compared with PECT supporting the fact as above that even with citrate samples some platelet activation is occurring \textit{in vitro}.\textsuperscript{745} Interestingly we selected processing at ambient temperature within our biobanking protocols over many years to avoid detrimental effects of cold temperatures on platelets which are normally stored unrefrigerated and yet clearly activation and degranulation appear to be inhibited if cold.\textsuperscript{746}
Comparing serum from four patients with rheumatoid arthritis analysed prior to freezing with results following subsequent freeze-thaw cycles found a dramatic change (average 67% with almost total degradation seen in two samples) after just one F/T cycle and continued decline with subsequent F/T cycles. In absolute terms this meant a change from a mean of 352 pg/mL (166 SD) to 134 (184) after 1 cycle and 49 (18) after 6. Although we did not assay samples prior to freezing, clearly it is unlikely that our data examining samples after freezing or one further cycle of F/T support this given the high values still present in all serum samples but this needs exploring further. One notable difference was the thawing of samples at 37°C in that study. Although EDTA plasma VEGF has been reported to decrease in samples (10 healthy controls and 10 patients with rheumatoid arthritis) thawed more than once compared with once only (mean difference 20%) but thereafter no further decline until after 10 thaws, serum VEGF only declined after 10 thaws (mean difference 18%). We did not find such changes and another study has reported no changes in EDTA plasma VEGF until after 7 and 9 F/T cycles. Further study using a multiplex chip also found no effect of freeze-thawing up to at least 10 times compared with immediate analysis using serum or heparinised plasma. Similar inconsistent results for storage have been reported with stability for up to at least 2 years at -80°C for EDTA plasma or serum compared with findings of significant degradation for periods longer than 3 months at -75°C for serum but this was using accelerated stability tests at elevated temperatures and given the possible effect of higher temperatures on VEGF during thawing, this needs to be revisited.

Systematic studies of biological variability of VEGF are few but consistent. In the first reported study comparing serum and plasma (citrated) we reported no effects of age, gender or menopausal status on plasma or serum VEGF and this has been confirmed since in a large (n=306) reference range study using EDTA plasma and serum. Some evidence of diurnal variation was reported and examination of intra-individual variability showed median CVs of 39% to 56% examined at several points in a month and repeated at 6 months and 1 year later and with exercise having a marked although short-term effect. Similar relatively large intra-individual variability has been reported in a further study using just EDTA plasma with CVs of 69% and 57% for short- and long-term biological variation respectively, compared with 51% inter-individual.
biological variation may contribute to the lack of differences seen in studies using plasma between different groups of patients with breast diseases which we reported and which has also been reported for patients with colorectal cancer, benign adenoma and disease-free groups.\textsuperscript{572, 739} Interestingly platelet-associated VEGF has been reported to show only low levels of intra-individual variability over time with CVs of only 17\% and inter-subject CVs of 44\% although interestingly in this study the intra-subject variability of citrated platelet-poor plasma processed at ambient temperature was only 19\% although inter-subject was markedly higher at 148\%.\textsuperscript{749}

Clearly the clinical utility of VEGF as a biomarker is far from established and given the critical effects of anticoagulant used on “circulating endogenous VEGF” it appears likely that studies employing EDTA or even citrated plasma are not truly representative of that VEGF component and may include variable contributions from platelet-derived VEGF fraction which is largely covered by serum VEGF measurements. Similar to the findings here in RCC and with the caveat that background endogenous circulating VEGF was from measurement of EDTA plasma, there was a wide variation in calculated platelet content of VEGF in breast cancer patients and healthy controls with no significant difference between them although in a study examining patients with a range of advanced cancers, platelet VEGF was significantly increased compared with healthy controls.\textsuperscript{572, 750} There was no relationship between either plasma or serum VEGF and clinicopathological parameters, although plasma VEGF discriminated more between controls and the various breast disease groups.\textsuperscript{572} Although several groups did have elevated VEGF compared with normal controls, we did not find a clear trend in breast cancer in plasma or serum VEGF with local disease, remission or metastatic disease but this was likely to be due to a possible effect of tamoxifen on VEGF, both circulating and platelet-derived.\textsuperscript{572} The question of whether under standard consistent processing conditions EDTA or citrated plasma can act as a surrogate of cancer behavior through essentially integrating platelet VEGF content with the activatability of platelets or whether serum provides a more stable indication of any cancer-associated properties remains to be determined and under the conditions recorded in our biobanking activity this should be possible to determine in RCC at least and is illustrated in the following Chapter.
Interestingly many of the published findings above regarding plasma VEGF and anticoagulant appear to have impacted little on the measurement of VEGF in clinical research studies with variable procedures still being adopted and it is important that this is highlighted in a large study. This may be in part due to the publication of many in clinical biochemical type journals and a greater impact may result if published in a cancer journal for example, as many of the studies are led by people working in cancer research. This issue of continued publication of studies ignoring such facts has also been seen for other proteins, for example MMP9 where despite the importance of serum versus plasma being highlighted and the effect on the interpretation of results, studies continue to measure MMP9 inappropriately and neglect pre-analytical considerations.\textsuperscript{751} We intend to publish the VEGF results from our studies as a separate paper also highlighting such aspects and additionally exploring further some of the discrepancies such as the stability of VEGF to freeze-thawing.

**Overall Conclusions**

This chapter has shown several examples of the importance of verifying the performance of commercially available assays prior to use in biomarker studies, and assessing the potential impact of pre-analytical factors. This illustrates how inconsistent results across studies can easily arise and with studies employing certified assays on clinical chemistry platforms and research grade immunoassays being used at various stages in the biomarker pipeline, it is often difficult to interpret data across studies. Without assay characterisation and validation featuring at an early phase of the biomarker translational pathway, progress in biomarker translation and adoption will continue to be slow and result in wasted resources. The value of the biobank can clearly be seen in this Chapter and in the initial prognostic study in RCC described in the following chapter. Collection of samples in multicentre studies has to be pragmatic and take into account resource availability and considerations of cost and logistics when deciding on the possible sample types and frequencies and of course consider future developments so maximising the value as long-term resource.\textsuperscript{327, 442} Of most importance is consistency of processes adopted and recording of relevant information both at the biobank level and in publications as proposed in the BRISQ recommendations.\textsuperscript{329, 636} Deviations can then be
factored in, but the more complex the protocols then the less likely that most centres will comply. Thought needs to be given to possible quality assurance checks on compliance with the sample types, processing timing steps and storage and this is being pursued by several groups, for example ISBER, although as yet there is no universal panel which can be used and provide the information needed. However there will be no uniform protocol for all biomarkers and all fluid types as pre-analytical factors impact on different proteins in different ways. If a specific biomarker is the focus of the studies then protocols can be evidence-based and comply with any known necessary pre-analytical specifications to ensure measurements are valid in just the same way that the assays have to achieve the required technical performance criteria. By adopting high standard evidence-based protocols with accurate record-keeping and quality systems, it is likely that biobanks such as the one here will be of value for many years and many questions can be answered and the suitability for any specific biomarker determined based on the processes adopted and knowledge of pre-analytical impacts as studies evolve.
Chapter 14 - Circulating Prognostic Biomarkers in Renal Cancer – Clinical Validation Study of Promising Candidates
**Introduction**

As outlined in Chapter 10 of this report, there remains an urgent clinical need for the identification and validation of biomarkers that provide prognostic information for patients with localised renal cell carcinoma (RCC). It is recognised that following surgery to remove the primary tumour, around a third of patients will relapse with distant metastatic disease. Accurately differentiating these patients from those that are likely to be cured by surgery alone allows for more rational use of finite NHS resources, in terms of intensity of follow-up, and stratifies patients for entry into on-going trials of adjuvant therapy. Such treatments are likely to be both costly to the NHS and potentially toxic for patients, further highlighting the need to identify and target high-risk groups.

It is critical that for a prognostic biomarker (or panel of markers) to be adopted into clinical practice, it must be shown to be superior, or add value, to currently employed prognostic scoring systems, which for RCC are based on standard clinicopathologic criteria alone. Such nomograms fail to adequately reflect individual tumour biology and the identification of molecular markers in RCC to improve risk stratification and delivery of more personalised medicine is recognised as a research priority by both the European Association of Urology and the European Society for Medical Oncology.

The current study represents the culmination of the RCC-related work, to date, within Workstream 2. It focuses on the clinical validation of candidate circulating biomarkers detectable in serum and/or plasma, collected pre-nephrectomy/ablation, that have been reported in the literature by ourselves and/or others to carry prognostic value in patients with localised clear cell RCC (ccRCC), using the assembled multicentre prospective observational cohort and research tissue bank described in Chapter 11. Markers were shortlisted based on the level of existing published evidence of their prognostic potential, in addition to the availability of suitable and robust assays. On this basis, the following proteins were selected for validation: i) vascular endothelial growth factor (VEGF-A referred to as VEGF) ii) osteopontin (OPN), iii) carbonic anhydrase IX (CAIX) and iv) C-reactive protein (CRP). A number of routine laboratory variables were also included in the analysis, again based on existing supporting literature.
The aim of this study was to validate the prognostic utility of the selected markers individually or combined as a panel or index in a large multi-centre cohort of UK patients with localised ccRCC. Furthermore, the ability of the markers to add value to the widely employed post-operative Leibovich score was examined, in particular amongst those patients deemed at high or intermediate-risk of relapse by the score alone.388

Methods

Patient population

Patients were identified retrospectively from the whole RCC cohort (described in full in Chapter 11). Inclusion criteria for the study were broad and included patients with i) ccRCC ii) radical/partial nephrectomy or tumour ablation iii) localised disease (stage I-III) iv) pre-operative serum/plasma sample availability and v) pre-operative clinical biochemistry/haematology measurement availability. All patients who fulfilled these criteria were included except for patients with Von Hippel-Lindau (VHL) disease (an exclusion criterion to the overall study) and co-existent other active cancers. For comparative descriptive purposes only, a subset of patients presenting with metastatic disease, who may or may not have undergone nephrectomy, but who otherwise met the inclusion/exclusion criteria above, were also included.

VEGF, OPN and CAIX measurement

A full description, including details on assay validation, is provided in Chapter 13 of this report. Similarly, full details of sample collection, processing and storage are described in Chapter 11. Briefly, osteopontin (OPN) and CAIX concentrations were quantified in EDTA plasma using commercially available enzyme-linked immunosorbent assays (ELISA) Quantikine® kits (R&D systems, Minneapolis). Both serum and EDTA plasma VEGF were analysed, using the Human VEGF Quantikine kit (R&D Systems, Minneapolis), a sandwich ELISA specifically measuring VEGF-A. The difference between serum and plasma VEGF concentrations was also examined for its value as a prognostic variable, calculated as serum minus plasma VEGF concentration. All samples
were measured blinded, in duplicate and manufacturer’s QC samples were included on each plate. Assay runs not passing QC standards i.e. QC samples not meeting the specifications supplied by the manufacturer, were rerun (this applied to two plates only). Similarly samples where replicate CVs exceeded 10% were reanalysed. Other serum analytes, including CRP, were measured by the NHS clinical biochemistry lab at each participating centre with those with CRP concentrations <10 mg/L being reanalysed using the high sensitivity CRP assay in the Leeds Teaching Hospitals Blood Sciences labs.

Clinico-pathological variables

Clinical factors examined included gender, age at diagnostic procedure, smoking history, alcohol consumption, body mass index (BMI), symptoms (local/systemic/absent), and Eastern Co-operative Oncology Group performance status (ECOG PS). Pathological factors recorded were tumour size, tumour-node-metastasis (TNM) stage, Fuhrman grade, Leibovich score, and presence or absence (if not commented, assumed to be absent) of histological necrosis, sarcomatoid change and microvascular invasion (MVI). Routine laboratory variables included haemoglobin, white cell count (WBC), neutrophils, lymphocytes, platelet count, serum sodium, potassium, urea, ALT, calcium, albumin and C-reactive protein (CRP). The derived parameter, neutrophil: lymphocyte ratio (NLR), was also examined on the basis of previous findings. 449, 452, 453

Statistical methods

Baseline concentrations of each marker were explored in terms of differences within demographic and clinical factors amongst all patients (stage I-IV) using Spearman rank correlation coefficient and p-value if a single continuous variable, or median (range) and p-value from Wilcoxon–Mann–Whitney test or Kruskal–Wallis test if comparing two or more subgroups. Correlations between the markers were investigated using a correlogram, based on simple linear regression and the Spearman rank correlation coefficient.
Survival analyses were conducted exclusively in patients with stage I-III disease (i.e. non-metastatic). Metastasis-free survival (MFS) formed the principal time-to-event endpoint, calculated as the period from date of procedure to the date of distant metastases, the definition also used in developing the Leibovich score. Any patients without disease recurrence were censored at the date they were last known to be recurrence-free (for patients who died without recurrence this was the date of death). Secondary endpoints were overall survival (OS) and cancer-specific survival (CSS), defined as the period from date of procedure to the date of i) death from any cause (OS) or ii) death from RCC (CSS). Patients still alive at the time of analysis were censored at the date last known alive (or at the date of non-cancer related death when considering CSS).

Cox proportional hazards (PH) models were used to determine the prognostic potential of the markers, survival functions were estimated using the Kaplan-Meier method and compared using the log-rank test. Markers were initially examined as continuous variables and then as dichotomised variables. The latter was accomplished by considering all possible cut-points within the range of each marker, and selecting the one which maximised Harrell’s concordance index (C-index). For each Cox PH model constructed, the PH assumption was tested by assessing Schoenfeld residuals.

The ability of the shortlisted markers to add prognostic utility to the Leibovich score was explored, particularly with respect to those patients in the intermediate- and high-risk scoring groups (score 3-5, score ≥ 6 respectively). This was performed by sequentially including each marker into a Cox PH model with the Leibovich score as an existing predictor variable and MFS as the response variable.

Statistical analysis was carried out in the R Environment for Statistical Computing (R Core Team, Vienna) and reported according to REMARK (REporting recommendations for tumor MARKer prognostic studies) criteria. In making inferences, significance levels were adjusted for multiple testing, where appropriate. Tests for significance were two-sided and p-values < 0.05 were considered significant.
Sample size
Sample size calculations were based on upper and lower extremes of relapse rate at 2-years for patients with localised ccRCC of 27.5%-12.5%. Using the higher end of the relapse rate, we assumed a separation at 2-years of 15% between survival curves to be required for each marker (given a dichotomised split around a given point) to justify its inclusion in a multiplex marker model. This equates to a hazard ratio of approximately 0.5. Table 89 shows the sample sizes required to identify a hazard ratio of 0.5 with 80-95% power given a significance level (α) of 5% in a 5 year study at the higher and lower relapse rate and at an intermediate rate of 20%. Samples sizes are shown for unadjusted and adjusted (using Bonferroni correction) significance levels.

Table 89 Sample size calculations Sample sizes required to obtain 80-95% power, when identifying a hazard ratio of 0.5 with 5% significance level (α₁ unadjusted, α² adjusted for 5 markers and α³ adjusted for 10 markers) assuming a relapse rate of a) 27.5%, b) 20% and c) 12.5%.

<table>
<thead>
<tr>
<th>Relapse rate = 27.5%</th>
<th>Relapse rate = 20%</th>
<th>Relapse rate = 12.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>α¹</td>
<td>α²</td>
</tr>
<tr>
<td>0.80</td>
<td>168</td>
<td>248</td>
</tr>
<tr>
<td>0.85</td>
<td>192</td>
<td>276</td>
</tr>
<tr>
<td>0.90</td>
<td>224</td>
<td>316</td>
</tr>
<tr>
<td>0.95</td>
<td>276</td>
<td>380</td>
</tr>
</tbody>
</table>

Results
In total, 706 patients were recruited into the full study, between July 2011 and June 2014, across 11 UK centres. Amongst the 629 patients with a confirmed RCC, 481 (76.4%) had ccRCC, 59 (9.4%) papillary, 46 (7.3%) chromophobe, 27 (4.3%) oncocytoma, 12 (1.9%) unclassified and the remaining four cases made up of 2 translocation tumours, 1 cystic mixed chromophobe and ccRCC and 1 mucinous tubular and spindle cell RCC.

Amongst patients with localised (stage I-III) ccRCC, the majority (n=406; 94.2%) met the inclusion criteria for the current study. In addition, a subset of 30 out of the 50 patients with ccRCC presenting with stage IV disease were examined in parallel, selected
to represent recruiting centres and distribution of metastatic site. These patients were included purely for descriptive purposes of the selected biomarkers, rather than assessment of their prognostic ability within this group. At the time of analysis, the median length of follow-up from diagnosis amongst patients still alive was 28.9 months (range 0.6-48.3). Amongst those presenting with localised RCC, 33 patients had relapsed with distant metastatic disease and 21 patients had died, of which 6 were directly attributed to their cancer. The small number of CSS events (n=6) precluded the inclusion of this endpoint in the current analysis.

**Patient and tumour characteristics**

Patient characteristics for the 406 patients included in the current study are shown in Table 90. The male:female ratio was 1.9:1, confirming the known male preponderance of this tumour type. Based on BMI, approximately three-quarters of patients were classified as either obese (36%) or overweight (40%). The majority of patients were ECOG PS 0/1 (97%). One third of patients were asymptomatic at the time of presentation. Amongst those patients reporting symptoms, 30% reported local symptoms (such as haematuria or flank pain), 14% reported systemic symptoms (such as weight loss or fatigue) and 22% reported both local and systemic symptoms.

At the time of diagnosis, over half of tumours (55%) were pathological stage T1, of which 30% were pT1a, whilst 28% were locally advanced (pT3). Amongst the 381 patients with localised ccRCC who underwent a radical or partial nephrectomy with an evaluable Leibovich score, 39%, 43% and 18% were classified as low, intermediate and high risk, respectively.

**Biomarker associations / correlations**

Correlation of the markers with each other was examined. The correlation was strongest between plasma and serum VEGF values ($r=0.59$, $p<0.001$). In general, the markers were significantly correlated, with the exception that CAIX showed no correlation with either serum or plasma VEGF values ($r=-0.02$, $p=0.655$ and $r=0.08$, $p=0.112$, respectively).
All four markers showed associations with several clinicopathological parameters. Increased pre-operative plasma VEGF concentrations were consistently associated with poor prognostic tumour factors, such as increased pathological tumour size ($r=0.14$, $p=0.005$), stage IV disease ($p<0.001$), grade 4 tumours ($p=0.025$), high-risk Leibovich score ($p=0.031$), the presence of sarcomatoid change ($p<0.001$) and MVI ($p=0.046$).

**Table 90 Characteristics of the patients with localised (stage I-III) ccRCC**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Study patients (n=406)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>267 (66)</td>
</tr>
<tr>
<td>Female</td>
<td>139 (34)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;18.5 (underweight)</td>
<td>4 (1)</td>
</tr>
<tr>
<td>18.5-24.9 (healthy)</td>
<td>91 (22)</td>
</tr>
<tr>
<td>25-29.9 (overweight)</td>
<td>147 (36)</td>
</tr>
<tr>
<td>&gt;30 (obese)</td>
<td>161 (40)</td>
</tr>
<tr>
<td>Missing</td>
<td>3 (1)</td>
</tr>
<tr>
<td><strong>Body Mass Index</strong></td>
<td></td>
</tr>
<tr>
<td>ECOG Performance Status</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>319 (79)</td>
</tr>
<tr>
<td>1</td>
<td>74 (18)</td>
</tr>
<tr>
<td>2</td>
<td>11 (3)</td>
</tr>
<tr>
<td>3</td>
<td>1 (&lt;1)</td>
</tr>
<tr>
<td>4</td>
<td>1 (&lt;1)</td>
</tr>
<tr>
<td>Symptoms</td>
<td></td>
</tr>
<tr>
<td>Local</td>
<td>120 (30)</td>
</tr>
<tr>
<td>Systemic</td>
<td>57 (14)</td>
</tr>
<tr>
<td>Both</td>
<td>88 (22)</td>
</tr>
<tr>
<td>None</td>
<td>141 (34)</td>
</tr>
<tr>
<td>Pathological T-stage</td>
<td></td>
</tr>
<tr>
<td>1a</td>
<td>122 (30)</td>
</tr>
<tr>
<td>1b</td>
<td>100 (25)</td>
</tr>
<tr>
<td>2a</td>
<td>38 (9)</td>
</tr>
<tr>
<td>2b</td>
<td>9 (2)</td>
</tr>
<tr>
<td>3a</td>
<td>106 (26)</td>
</tr>
<tr>
<td>3b</td>
<td>9 (2)</td>
</tr>
<tr>
<td>3c</td>
<td>1 (&lt;1)</td>
</tr>
<tr>
<td>Not applicable*</td>
<td>21 (5)</td>
</tr>
<tr>
<td>Tumour size (mm)</td>
<td></td>
</tr>
<tr>
<td>median (range)</td>
<td>52 (11, 180)</td>
</tr>
<tr>
<td>Fuhrman Grade</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>11 (3)</td>
</tr>
<tr>
<td>2</td>
<td>142 (35)</td>
</tr>
<tr>
<td>3</td>
<td>203 (50)</td>
</tr>
<tr>
<td>4</td>
<td>49 (12)</td>
</tr>
</tbody>
</table>

431
<table>
<thead>
<tr>
<th></th>
<th>Missing</th>
<th>1 (&lt;1)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Necrosis</strong></td>
<td>Present</td>
<td>102 (25)</td>
</tr>
<tr>
<td></td>
<td>Absent</td>
<td>283 (70)</td>
</tr>
<tr>
<td></td>
<td>Not applicable *</td>
<td>21 (5)</td>
</tr>
<tr>
<td><strong>Microvascular invasion</strong></td>
<td>Present</td>
<td>65 (16)</td>
</tr>
<tr>
<td></td>
<td>Absent</td>
<td>320 (79)</td>
</tr>
<tr>
<td></td>
<td>Not applicable *</td>
<td>21 (5)</td>
</tr>
<tr>
<td><strong>Sarcomatoid change</strong></td>
<td>Present</td>
<td>18 (5)</td>
</tr>
<tr>
<td></td>
<td>Absent</td>
<td>367 (90)</td>
</tr>
<tr>
<td></td>
<td>Not applicable *</td>
<td>21 (5)</td>
</tr>
<tr>
<td><strong>Leibovich Risk Group</strong></td>
<td>Low</td>
<td>147 (36)</td>
</tr>
<tr>
<td></td>
<td>Intermediate</td>
<td>163 (40)</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>71 (18)</td>
</tr>
<tr>
<td></td>
<td>Not applicable **</td>
<td>25 (6)</td>
</tr>
<tr>
<td><strong>TNM Stage</strong></td>
<td>I</td>
<td>240 (59)</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>44 (11)</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>116 (29)</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>6 (1)</td>
</tr>
<tr>
<td><strong>Procedure</strong></td>
<td>Radical nephrectomy</td>
<td>286 (70)</td>
</tr>
<tr>
<td></td>
<td>Partial nephrectomy</td>
<td>99 (24)</td>
</tr>
<tr>
<td></td>
<td>Radiofrequency Ablation</td>
<td>13 (3)</td>
</tr>
<tr>
<td></td>
<td>Cryoablation</td>
<td>8 (2)</td>
</tr>
<tr>
<td><strong>Relapsed</strong></td>
<td>Yes</td>
<td>33 (8)</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>352 (87)</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>21 (5)</td>
</tr>
</tbody>
</table>

*Not available in patients undergoing tumour ablation
**Includes 4 patients initially thought to be stage IV and therefore no risk score assigned

Amongst patients with localised disease, a trend towards increased plasma VEGF in relapsers vs non-relapsers was observed (84.8 vs 66.5 pg/mL; p=0.054). Serum VEGF concentrations showed fewer significant associations, limited to Leibovich score (p=0.008), presence of necrosis (p=0.041) and overall TNM stage (p=0.018). Serum minus plasma VEGF showed an association with Leibovich score and presence of necrosis only. Box and whisker plots are shown for the markers according to TNM stage (Figure 52) and the Leibovich score (Figure 53).

Similarly, higher baseline circulating concentrations of OPN, CRP and CAIX were consistently associated with poor prognostic features, including higher stage and grade.
increased tumour size, and presence of necrosis, MVI and sarcomatoid change. An association with Leibovich score was again observed (Figure 53). Both CRP ($p=0.002$) and OPN ($p=0.001$) were elevated in relapsers vs non-relapsers. In terms of clinical associations, higher circulating levels of all three biomarkers were associated with poorer ECOG PS. Circulating concentrations of both CRP and CAIX showed a significant association with BMI, with patients with a BMI $<18.5$ (underweight) having higher median serum/plasma levels of each biomarker. However, only six patients made up this group, two of whom had metastatic disease.
Figure 53 Box and whisker plots for pre-operative circulating CRP, OPN, CAIX and VEGF according to Leibovich risk classification (low, intermediate or high): boxes correspond to the first and third quartiles (the 25th and 75th percentiles) with median concentration also shown, whiskers extend to ± 1.5 * interquartile range, with data points beyond classed as outliers.

Univariate analysis of biomarkers and time-to-event endpoints

Metastasis-Free Survival
The following were statistically significantly associated with MFS (p<0.05): gender, WBC, neutrophils, NLR, platelet count, potassium, albumin, CRP, OPN, serum VEGF, plasma VEGF, serum minus plasma VEGF, pathological (p) tumour size, pT stage, Fuhrman grade, necrosis, MVI, sarcomatoid change, Leibovich score, CT tumour size, CT T stage, and overall TNM stage (Table 91). Figure 54 shows Kaplan-Meier survival curves for Fuhrman grade and TNM stage as two of the most clinically relevant variables.
found to be significant in univariate analysis, in addition to the Leibovich score as a main focus of the study. Chi-squared and associated p-values from the log-rank test to compare survival curves are also shown and are significant in all three cases. In addition there was weak evidence of an association with MFS for serum sodium concentration (HR=0.89; 95% CI [0.79-1.01]; p=0.064).

**Figure 54 Kaplan-Meier survival curves showing MFS by TNM stage, Fuhrman grade and Leibovich score**
### Table 91 Significant (p<0.05) univariate Cox PH results of metastasis-free survival (MFS) and / or overall survival (OS) (markers considered as continuous variables)

<table>
<thead>
<tr>
<th></th>
<th>MFS</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR (95% CI)</td>
<td>p-value</td>
</tr>
<tr>
<td>Gender</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>Male</td>
<td>0.31 (0.11, 0.90)</td>
<td>0.031</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WBC</td>
<td>1.18 (1.02, 1.37)</td>
<td>0.026</td>
</tr>
<tr>
<td>Neutrophils</td>
<td>1.23 (1.05, 1.44)</td>
<td>0.009</td>
</tr>
<tr>
<td>NLR</td>
<td>1.13 (1.01, 1.26)</td>
<td>0.038</td>
</tr>
<tr>
<td>Platelet count</td>
<td>1.00 (1.00, 1.01)</td>
<td>0.002</td>
</tr>
<tr>
<td>Sodium</td>
<td>0.89 (0.79, 1.01)</td>
<td>0.064</td>
</tr>
<tr>
<td>Potassium</td>
<td>2.41 (1.11, 5.24)</td>
<td>0.026</td>
</tr>
<tr>
<td>Albumin</td>
<td>0.94 (0.90, 0.98)</td>
<td>0.006</td>
</tr>
<tr>
<td>CRP</td>
<td>1.01 (1.00, 1.01)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>OPN</td>
<td>1.01 (1.00, 1.01)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CAIX</td>
<td>1.00 (0.99, 1.00)</td>
<td>0.729</td>
</tr>
<tr>
<td>Serum VEGF</td>
<td>1.00 (1.00, 1.00)</td>
<td>0.012</td>
</tr>
<tr>
<td>Plasma VEGF</td>
<td>1.00 (1.00, 1.01)</td>
<td>0.032</td>
</tr>
<tr>
<td>VEGF (serum-plasma)</td>
<td>1.00 (1.00, 1.00)</td>
<td>0.028</td>
</tr>
<tr>
<td>Tumour size</td>
<td>1.01 (1.01, 1.02)</td>
<td>0.002</td>
</tr>
<tr>
<td>pT stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>T2</td>
<td>3.15 (0.89, 11.16)</td>
<td>0.076</td>
</tr>
<tr>
<td>T3/T4</td>
<td>6.22 (2.47, 15.68)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/2</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>3</td>
<td>2.62 (0.85, 8.04)</td>
<td>0.092</td>
</tr>
<tr>
<td>4</td>
<td>11.39 (3.59, 36.12)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Necrosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>Yes</td>
<td>5.29 (2.44, 11.51)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Microvascular Invasion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>Yes</td>
<td>2.31 (1.02, 5.24)</td>
<td>0.046</td>
</tr>
<tr>
<td>Sarcomatoid change</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>Yes</td>
<td>7.08 (2.67, 18.77)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Leibovich risk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>Intermediate</td>
<td>1.29 (0.36, 4.59)</td>
<td>0.691</td>
</tr>
<tr>
<td>High</td>
<td>11.18 (3.78, 33.04)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CT size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>1.18 (1.06, 1.31)</td>
<td>0.003</td>
</tr>
<tr>
<td>T2</td>
<td>2.62 (1.01, 6.80)</td>
<td>0.047</td>
</tr>
<tr>
<td>T3/T4</td>
<td>4.02 (1.55, 10.41)</td>
<td>0.004</td>
</tr>
<tr>
<td>TNM Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>II</td>
<td>3.61 (1.02, 12.82)</td>
<td>0.047</td>
</tr>
<tr>
<td>III</td>
<td>6.83 (2.71, 17.22)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Optimally discriminative cut-points in terms of maximised C-index were derived as: CRP (cut-point 14.1 mg/L), OPN (cut-point 120.8 ng/mL), CAIX (cut-point 60.4 pg/mL), platelet count (cut-point 333 x10^9/L) and serum sodium (cut-point 141 mmol/L), were all significantly associated with MFS (p<0.05) (Table 92). Corresponding Kaplan-Meier survival curves are shown in Figure 55.

**Figure 55** Kaplan-Meier survival curves showing MFS for dichotomised markers, serum sodium and platelet count (only markers found significant in univariate analysis shown)

**Overall Survival**

The following, when considered as continuous variables, were statistically significantly associated with OS (p<0.05): Hb, serum sodium, potassium, albumin, CAIX, Fuhrman grade, sarcomatoid change and Leibovich score. When the markers, serum sodium and platelet count were considered as dichotomised variables, CAIX (cut-point 112.1 pg/mL),
plasma VEGF (cut-point 132 pg/mL), platelet count (cut-point 300 x10^9/L) and serum sodium (cut-point 137 mmol/L) were statistically significant (Table 92).
Table 92 Univariate Cox PH analysis of metastases-free survival (MFS) and overall survival (OS) by optimal cut-points

<table>
<thead>
<tr>
<th></th>
<th>Optimised cut-point</th>
<th>N</th>
<th>Number of events</th>
<th>HR (95% CI)</th>
<th>p-value</th>
<th>Optimised cut-point</th>
<th>N</th>
<th>Number of events</th>
<th>HR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≤14.1</td>
<td>294</td>
<td>11</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
<td>≤3.9</td>
<td>174</td>
<td>5</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
</tr>
<tr>
<td>CRP</td>
<td>&gt;14.1</td>
<td>72</td>
<td>17</td>
<td>6.62 (3.10, 14.16)</td>
<td>&lt;0.001</td>
<td>&gt;3.9</td>
<td>199</td>
<td>15</td>
<td>2.65 (0.96, 7.29)</td>
<td>0.059</td>
</tr>
<tr>
<td>OPN</td>
<td>≤120.8</td>
<td>257</td>
<td>11</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
<td>≤95.4</td>
<td>199</td>
<td>6</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td>&gt;120.8</td>
<td>86</td>
<td>15</td>
<td>4.28 (1.96, 9.31)</td>
<td>&lt;0.001</td>
<td>&gt;95.4</td>
<td>149</td>
<td>11</td>
<td>2.67 (0.99, 7.22)</td>
<td>0.053</td>
</tr>
<tr>
<td>CAIX</td>
<td>≤60.4</td>
<td>176</td>
<td>8</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
<td>≤112.1</td>
<td>289</td>
<td>10</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td>&gt;60.4</td>
<td>199</td>
<td>20</td>
<td>2.37 (1.04, 5.38)</td>
<td>0.040</td>
<td>&gt;112.1</td>
<td>93</td>
<td>11</td>
<td>3.51 (1.48, 8.28)</td>
<td>0.004</td>
</tr>
<tr>
<td>Serum VEGF</td>
<td>≤268.7</td>
<td>106</td>
<td>4</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
<td>≤356.5</td>
<td>169</td>
<td>7</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td>&gt;268.7</td>
<td>223</td>
<td>20</td>
<td>2.56 (0.87, 7.50)</td>
<td>0.087</td>
<td>&gt;356.5</td>
<td>167</td>
<td>11</td>
<td>1.72 (0.67, 4.44)</td>
<td>0.262</td>
</tr>
<tr>
<td>Plasma VEGF</td>
<td>≤60.0</td>
<td>147</td>
<td>7</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
<td>≤132.1</td>
<td>322</td>
<td>14</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td>&gt;60.0</td>
<td>226</td>
<td>21</td>
<td>2.12 (0.90, 5.01)</td>
<td>0.086</td>
<td>&gt;132.1</td>
<td>58</td>
<td>7</td>
<td>3.08 (1.24, 7.66)</td>
<td>0.015</td>
</tr>
<tr>
<td>VEGF (serum-plasma)</td>
<td>≤344.9</td>
<td>194</td>
<td>10</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
<td>≤222.4</td>
<td>122</td>
<td>5</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td>&gt;344.9</td>
<td>135</td>
<td>14</td>
<td>2.12 (0.94, 4.78)</td>
<td>0.069</td>
<td>&gt;222.4</td>
<td>214</td>
<td>13</td>
<td>1.68 (0.6, 4.74)</td>
<td>0.326</td>
</tr>
<tr>
<td>Platelets</td>
<td>≤333</td>
<td>315</td>
<td>17</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
<td>≤300</td>
<td>280</td>
<td>10</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td>&gt;333</td>
<td>61</td>
<td>10</td>
<td>3.16 (1.44, 6.93)</td>
<td>0.004</td>
<td>&gt;300</td>
<td>103</td>
<td>11</td>
<td>2.84 (1.2, 6.71)</td>
<td>0.017</td>
</tr>
<tr>
<td>Serum sodium</td>
<td>≤141</td>
<td>276</td>
<td>27</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
<td>≤137</td>
<td>71</td>
<td>9</td>
<td>1.00 (1.00)</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td>&gt;141</td>
<td>102</td>
<td>1</td>
<td>0.09 (0.01, 0.66)</td>
<td>0.018</td>
<td>&gt;137</td>
<td>314</td>
<td>12</td>
<td>0.23 (0.1, 0.56)</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Multivariable Analysis of MFS

The prognostic ability of the markers to sub-stratify intermediate and/or high-risk patients by the Leibovich score was examined by sequentially adding markers found to be significant at the univariate level, in addition to platelet count and serum sodium as predictor variables, into a Cox PH model with the Leibovich score and MFS as the response variable. In the multivariable setting for each of the markers considered, the intermediate and high Leibovich score patients had a greater risk of relapse than the low Leibovich score patients, although only the high-risk group were significantly different (Table 93). When markers were considered as continuous variables none were significant (results omitted). When dichotomised, the only biomarker significant in the multivariable models was CRP (HR=3.22, 95% CI (1.39, 7.49), p-value 0.007). Corresponding Kaplan-Meier survival curves are shown in Figure 56.

Table 93 Multivariable Cox PH analysis of metastasis-free survival (MFS)

<table>
<thead>
<tr>
<th></th>
<th>HR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CRP</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤14.1 mg/L</td>
<td>1.00</td>
<td>(-)</td>
</tr>
<tr>
<td>&gt;14.1 mg/L</td>
<td>3.22 (1.39, 7.49)</td>
<td>0.007</td>
</tr>
<tr>
<td>Low</td>
<td>1.00</td>
<td>(-)</td>
</tr>
<tr>
<td><strong>Leibovich risk</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>1.21 (0.34, 4.29)</td>
<td>0.770</td>
</tr>
<tr>
<td>High</td>
<td>6.44 (2.03, 20.41)</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>OPN</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤120.8 ng/mL</td>
<td>1.00</td>
<td>(-)</td>
</tr>
<tr>
<td>&gt;120.8 ng/mL</td>
<td>1.7 (0.73, 3.97)</td>
<td>0.223</td>
</tr>
<tr>
<td>Low</td>
<td>1.00</td>
<td>(-)</td>
</tr>
<tr>
<td><strong>Leibovich risk</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>1.35 (0.32, 5.69)</td>
<td>0.680</td>
</tr>
<tr>
<td>High</td>
<td>11.29 (3.09, 41.21)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>CAIX</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤60.4 pg/mL</td>
<td>1.00</td>
<td>(-)</td>
</tr>
<tr>
<td>&gt;60.4 pg/mL</td>
<td>1.73 (0.74, 4.05)</td>
<td>0.205</td>
</tr>
<tr>
<td>Low</td>
<td>1.00</td>
<td>(-)</td>
</tr>
<tr>
<td><strong>Leibovich risk</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>1.11 (0.31, 4.02)</td>
<td>0.873</td>
</tr>
<tr>
<td>High</td>
<td>9.31 (3.04, 28.48)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Serum sodium</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤141 mmol/L</td>
<td>1.00</td>
<td>(-)</td>
</tr>
<tr>
<td>&gt;141 mmol/L</td>
<td>0.15 (0.02, 1.14)</td>
<td>0.067</td>
</tr>
<tr>
<td>Low</td>
<td>1.00</td>
<td>(-)</td>
</tr>
<tr>
<td><strong>Leibovich risk</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>1.34 (0.38, 4.75)</td>
<td>0.651</td>
</tr>
<tr>
<td>High</td>
<td>9.17 (3.09, 27.18)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Platelet count</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤333</td>
<td>1.00</td>
<td>(-)</td>
</tr>
<tr>
<td>&gt;333</td>
<td>1.60 (0.70, 3.65)</td>
<td>0.265</td>
</tr>
<tr>
<td>Low</td>
<td>1.00</td>
<td>(-)</td>
</tr>
<tr>
<td><strong>Leibovich risk</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>1.71 (0.43, 6.83)</td>
<td>0.449</td>
</tr>
<tr>
<td>High</td>
<td>13.12 (3.78, 45.52)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Figure 56 Kaplan-Meier survival curves showing MFS for dichotomised markers, serum sodium and platelet count (only markers found significant in univariate analysis shown) with the Leibovich score included as a further predictor variable. Optimised marker cut points as follows; CRP=14.1 mg/L, OPN=120.8 ng/mL, CAIX=60.4 pg/mL, optimised sodium cut point=141 mmol/L and optimised platelet count cut point=333.

Discussion

The multi-centre RCC biobank established within this Programme represents a unique resource for biomarker validation in the UK. In this initial study, we have sought to validate a number of proposed circulating prognostic biomarkers within a large cohort of patients with localised ccRCC.

The characteristics of the current study population are in keeping with previous series. For example, in a Leeds cohort of 140 patients presenting with localised ccRCC between 1998 and 2005, 58%, 7% and 35% had a pT1, pT2 and pT3 tumour, respectively, compared to 58%, 12% and 40% in the current series. The prognostic nomogram proposed by Leibovich
and colleagues in 2003 was developed in a US study of 1671 patients with localised ccRCC. The tool classifies patients into three risk groups, with 41%, 37% and 22% classified as low, intermediate or high risk for distant relapse, respectively. In the current UK cohort, equivalent figures were 39%, 43% and 18%, suggesting this distribution has changed little over the past decade.

Almost a third of patients were diagnosed with a small renal mass (SRM), defined as <4cm in maximal dimension, in the current study. The incidental detection of patients with SRMs has been rising in recent years, due to the more widespread use of cross-sectional imaging. Management of these small, typically low risk, tumours poses a significant challenge to clinicians, with the need to balance the risks of treatment against the chances of the tumour progressing within the lifetime of the patient. Biomarkers to allow stratification of these tumours by risk of progression remains a major unmet clinical need, which the current cohort of samples is well-placed to help address in the future as the data matures.

Factors such as increasing tumour stage and grade, as well as the presence of necrosis and sarcomatoid change, are accepted poor prognosis factors in RCC, and were associated with worse MFS in the current study. Reporting of MVI is variable and, in particular, is thought to be dependent on the meticulousness of the reviewing pathologist. MVI is currently not recommended for inclusion in the TNM staging of RCC. In a small study of 48 patients with T1/T2 RCCs (90% ccRCC), MVI was reported in 17% of patients and was an independent prognostic factor for DFS. MVI was found in 29% of centrally reviewed cases, in another study of 255 patients with pT1-pT3bN0M0 tumours (93% ccRCC), and was shown to have independent prognostic ability in terms of CSS and OS. Furthermore, in a recent study of 1754 patients with localised ccRCC, addition of MVI status was shown to improve the predictive accuracy of the Leibovich score by 1.4%. Conversely, however, in a study of 2078 patients with ccRCC, although MVI (seen in 19.8%) was significantly associated with a worse CSS amongst localised disease patients on univariate analysis, this was lost on multivariable testing. In the current study, MVI was reported as present in 16% of cases, in keeping with previous series without central slide review, and showed a significant association with MFS (p=0.046) on univariate analysis. The independent predictive ability of MVI was not examined in the current study due to the current small number of events limiting the power to detect differences in outcome, but will be looked at once the data has matured.
The current study confirms our previous finding from a smaller (n=216), single centre, cohort of patients, that higher circulating concentrations of OPN, CRP and CAIX are significantly associated with known poor prognostic factors such as higher stage and grade, and, in addition, extends these findings to include associations with other features such as presence of tumour necrosis, sarcomatoid change and MVI, at a univariate level. Both pre-operative CRP and OPN, but not CAIX, were associated with MFS when considered as continuous variables. Conversely, only CAIX was associated with OS, on univariate analysis. We previously reported an association of all three markers with disease-free survival (DFS), CSS and OS. MFS was not examined in the previous study although each of these endpoints is an expected surrogate of survival. The differences here are likely due to the small number of events observed in the short period of follow-up.

Only CRP was found to be independently prognostic when considered in a multivariable model including the Leibovich risk classification (p=0.007). Strikingly, the data suggest that patients in the intermediate risk group can be sub-stratified, with patients with a pre-operative CRP ≤ 14.1 mg/L associated with an excellent outcome, equivalent to a low-risk Leibovich group. Such patients could, therefore, be spared intensive follow-up and a necessity for consideration of adjuvant therapies. It must be acknowledged, however, that the relatively small number of events at the time of analysis limited the number of variables that could be included in multivariable modelling and means that the current analysis should be regarded as exploratory. Categorisation of continuous variables is more clinically applicable although does come at a cost, since information is lost, reducing statistical power. How best to dichotomise the data is also debated, with studies variably using the median value or, as in the current study, by determining an optimal cut-point that gives the minimum p-value. Such data-driven approaches have been criticised, since they can lead to overfitting and optimistic model performance. Again, this is acknowledged by the authors, and in subsequent analyses we will use additional methods for categorisation, such as our previously described simulation-based method. Validation of our previously described optimal cut-point for CRP (15 mg/L), as well as cut-offs defined by others (5 mg/L, 7.5 mg/L), will also be undertaken, although such cross-study comparisons must be performed with care. Issues such as differences in the particular assay used and, when comparing older studies, the more recent availability of high-sensitivity CRP assays, may confound results. Ultimately, whatever method and choice of cut-point is used, these issues highlight the
necessity to carefully validate initial results using adequately powered, independent, datasets.

We were the first group to report pre-operative serum sodium as being independently prognostic for DFS amongst patients with localised ccRCC when considered as both a continuous variable (n=103) (HR=0.78, 95% CI [0.66-0.92], p=0.003) and when dichotomised to above and below the median value (139 mmol/L) (n=137) (HR=0.39, 95% CI [0.18,0.84], p=0.012). These findings have since been replicated in patients with metastatic RCC but, to our knowledge, have not been re-examined in patients with localised disease. Here, we show that serum sodium considered as a continuous variable associates with MFS at the univariate level, although not reaching significance at the p<0.05 level (p=0.064). As a dichotomised variable, using an optimal cut-point, patients with a pre-operative serum sodium >141 mmol/L had a HR for relapse of 0.09 (95% CI [0.01, 0.66], p=0.018), again suggesting that a higher serum sodium is associated with a better outcome. The mechanism underlying this association remains unclear, but has even led some to suggest that correction of relative hyponatraemia may be worth exploring as a therapeutic strategy in RCC.

As outlined in chapter 12 of this report, a number of studies have examined the prognostic utility of circulating VEGF concentrations in patients with RCC. These studies have variably employed either serum or plasma, with no consistency in reported findings. The current study is unique in that we chose to examine both serum and plasma VEGF in parallel, and, to our knowledge, is the largest to date to examine VEGF concentrations using either matrix, amongst patients with localised ccRCC. Both serum and plasma VEGF concentrations were significantly elevated amongst patients with M1 versus M0 disease, but were not different amongst stage I-III patients. However, amongst localised disease patients grouped by Leibovich score, pre-operative VEGF concentrations were elevated amongst patients with high risk versus intermediate or low risk tumours, when measured in either serum or plasma or when considering serum minus plasma concentrations. On univariate analysis, both serum and plasma VEGF associated with MFS when considered as a continuous variable, but not when dichotomised, although plasma VEGF was significant for OS (cut-point 132 pg/mL). At present, therefore, it is difficult to conclude that either fluid is superior to the other in terms of clinical relevance to RCC outcomes and requires further future analysis, which is planned.
It is recognised that both the choice and definition of time-to-event end points in clinical trials varies, making between trial comparisons imprecise.\textsuperscript{770} It is equally a potential issue when trying to compare biomarker studies. At the time of writing this report, in an effort to standardise reporting in RCC trials, a recent consensus view has been published by the DATECAN renal cancer group.\textsuperscript{770} Amongst patients with localised disease, MFS, DFS and local-regional-free recurrence were recommended as intermediate end-points. MFS was defined as ‘death from kidney cancer or appearance of metastases, whichever comes first’. In the current study, we defined MFS as time to appearance of metastases only, not including deaths, since this was the definition of MFS on which the Leibovich score was developed. Although clearly OS and CSS represent constants, the otherwise lack of consistency amongst prognostic studies in choice and definition of intermediate end-point is an important issue that biomarker reporting guidelines such as the REMARK guidelines should consider addressing.

A limitation of the current study is the relatively short median follow-up of approximately 2-years, meaning relatively few events have occurred at the time of analysis. Since most relapses occur within 18-24 months of nephrectomy, the currently reported biomarker associations may become more significant with greater length of follow-up and number of events. Future analyses are therefore planned and will be extended, for example by examining markers as continuous variables transformed via fractional polynomial methods and by looking at combinations of biomarkers and algorithms. The Leibovich score will be examined not just by risk group, but also by score (i.e. 0-11) and as individual elements i.e. T-stage, N-stage, tumour size, grade, necrosis. Furthermore, since certain elements, such as grade, are subjective and prone to inter-observer variability, and presence of necrosis open to sampling error, the value of the selected markers to the score, excluding these elements, will be examined.\textsuperscript{771, 772}

In conclusion, the multi-centre RCC biobank established within this Programme consists of a large cohort of patients with ccRCC with a typical distribution of clinicopathological characteristics and expected survival associations with known prognostic factors such as stage and grade. As such, it represents an excellent resource for validation studies of prioritised biomarkers. Despite the current relatively small number of survival events, this initial study has been able to demonstrate promising associations of the selected biomarkers.
with outcomes. Exploratory multivariable analysis suggests that, when dichotomised by optimal cut-point, pre-operative CRP may add value to the Leibovich score. The results justify further exploration in future analyses, which are planned and will be undertaken once median follow-up has been extended.
Chapter 15 - Conclusions of the Clinical Translational Workstream
The Clinical Translation Workstream (WS2) was designed to evaluate approaches to streamline and speed up the central components of the biomarker pipeline. The pipeline runs from discovery to the implementation of the appropriate biomarker testing within the healthcare system generating benefits for patients and improvements in healthcare quality and cost effectiveness. The central components include consideration of Analytic Validity and Clinical Validity. Robust evaluation of these two aspects of the pipeline are essential before clinical utility and consequent benefits for patients and healthcare services can be evaluated.

We identified modest literatures in renal cancer and renal transplantation identifying moderate numbers of candidate biomarkers, often only identified in single papers in mixed patient groups. Studies sometimes fail to distinguish the roles of candidate markers in, for example, prognosis or treatment selection. However, in both cases progress in evaluating the performance of the biomarkers and then taking them to clinical practice has been slow and few new biomarkers have been introduced in recent decades. Small study size and their heterogeneity are important factors in this. In liver disease, we were able to study the ELF test for which a substantial body of evidence for clinical utility existed. This presented us with the opportunity to take the ELF test into a formal randomised trial, the ELUCIDATE RCT described in subsequent Chapters.

The investigators believed at the outset of this workstream that they would find that many of the candidate biomarkers lacked sufficient evidence for Analytic and Clinical Validity to justify their evaluation in large prospective studies of clinical utility. They hypothesised that this would be the case because the acquisition of appropriate clinical samples, annotated with high quality clinical data, is a slow process and studies are frequently done in samples of uncertain quality, inadequate numbers and with insufficient attention to methodological considerations.

The acquisition of high quality sample banks with appropriate clinical data was deemed to be one part of the solution to speed up the biomarker evaluation pipeline. The simple hypothesis was that a standing bank of samples, carefully curated and clinically annotated in adequate numbers would provide a resource that would allow candidate biomarkers to be robustly evaluated in order to take the decision as to whether they should go through for full evaluation of clinical utility and their place in clinical practice.
The discipline of clinical biochemistry was strongly represented in the investigating team and they contributed to the definition of the appropriate sample handling and curation requirements. The synergy between research scientists and clinical bioscientists contributed to the rigour of test development and evaluation. The methodology teams from Workstream 1 and Workstream 3 advised on study design, cohort size and evaluation. The expertise of the Clinical Trials Research Unit, which was designing, delivering and analysing the ELUCIDATE RCT in Workstream 3 was used to establish a robust prospective high quality clinical data annotation process.

The Clinical Translation Workstream has delivered cohorts of patients with high quality samples and clinical annotation. The performance of some candidate biomarkers has been evaluated in the immediate term and has provided a legacy for future studies. This resides in the sample banks and clinical data which are a resource that will enable rapid validation of further biomarkers in the disease areas. However, we associate greater generic value with the outputs and learning points for the general aspects of the biomarker pipeline.

Several publications have already been generated ranging from biomarker reviews through to exploration of technical issues of specific immunoassays and clearly much of the material described in the previous chapters will also result in further publications. These will include further technical pre-analytical papers, a commentary on the various aspects of the set-up process which will be of considerable relevance when planning this type of activity going forwards, and many biomarker studies are anticipated.

We would like to emphasise the generic learning points from the work described in Chapters 10 to 14. Chapters 10 and 11 describe the preparation and delivery of the multicentre sample banks in renal diseases. They demonstrate that with rigorous attention to detail it is possible in the NHS to generate high quality sample banks and high quality clinical annotations for biomarker evaluations. The research and innovation capacity of the NHS was harnessed across multiple centres in order to create the samples, clinical data and infrastructure to rapidly evaluate candidate biomarkers in renal cancer and in patients after renal transplantation. The challenges faced were substantial and are often generalisable. Details of delays in study set up, the challenges of quality assuring samples and clinical data are well illustrated by the data in Chapters 11 and 13. This work will serve as a useful exemplar for the strategic approach we have advocated for biomarker evaluation. We have provided the
NHS, NIHR, the academic community and their partners with access to materials which allow the prompt and robust evaluation of analytic and clinical validity. The challenges in multicentre studies, the characteristics of successful centres and the energy and commitment that is necessary to deliver this approach, are clear.

In Chapter 12 the investigation team worked to review and prioritise the circulating biomarkers in renal cancer and renal transplantation that were identifiable from the literature. This chapter demonstrates that such candidate biomarkers exist in reasonable numbers but that the pace of innovation and new discovery which is leading through to biomarkers of robust analytic and clinical validity is still slow. All of the investigation team were disappointed that during the period of this study exciting new biomarkers or biomarker panels, particularly with protein biomarkers in body fluids, did not emerge. The review shows that many studies in the literature are small and inconclusive but the overview analysis clearly identified candidates that could be evaluated further. This led to the evaluation in Chapter 13 of the appropriate assays with suitable analytic validity. Chapters 13 and 14 describe the delivery of tests of good analytic validity and clinical validity for the prioritised and selected biomarkers against the sample bank and candidates for further tests of clinical utility were demonstrated.

The investigators cautiously conclude that their approach has merit and can provide an example of how this field can be streamlined. However, we have highlighted the considerable organisational and logistic challenges which must be overcome in order to effectively deliver the pipeline development.

Critical to continued improvement in the biomarker pipeline will be the multidisciplinary nature of the approaches that must be taken. For the investigation within this programme we were fortunate to have enthusiastic inputs from research scientists, clinical biochemists, methodologists, clinicians and trialists. We believe that there is little prospect of success in individual studies or in continued improvement, streamlining and speeding up of the biomarker pipeline in the absence of consistent multidisciplinary inputs of this kind.

The organisational and the logistical challenges have been highlighted. During the duration of this study we noted a steady improvement in set up times following the hard work carried out by NIHR infrastructure organisations including NIHR CRN. To continue to deliver
progress, partnerships between the NHS, universities and funders are essential. Each brings unique components of the expertise to deliver the improved biomarker pipeline which is sought by all.

Patients played a substantial part in the design, delivery and conduct of this work. Our PPI workstream and PPI commentaries were important at all stages. The engagement of patients in the methodology workstreams and of course their engagement as partners in the delivery of the cohorts and the provision of samples were essential components of the progress which has been identified in this programme workstream. We have continued to work very closely with UK industry and developed strong working relationships, most notably with Randox in a successful first phase SBRI bid.

The Clinical Translational Workstream provided the basis for the application, including many of the investigators on the NIHR Applied Programme, to become an NIHR Diagnostic Evidence Cooperative (DEC) which was successful and began work in 2013. We will describe the DEC, the learning from the Applied Programme which underpinned its development and its operations, and the extension of its scope beyond the Applied Programme to include our colleagues in musculoskeletal disease and in other aspects of oncology in Chapter 25.
Chapter 16 - Introduction to the ELUCIDATE Trial (including scientific background and explanation of rationale)
The ELUCIDATE RCT (WS3) aimed to conduct a randomised controlled trial (RCT) of an established panel of biomarkers (ELF) of potential value in chronic liver disease, to diagnose cirrhosis at an early stage when beneficial interventions to reduce dangerous complications are possible, which may lead to patient and NHS benefits. The rationale for selecting the ELF test is discussed in Chapter 1. Briefly, in the vast majority of cases, liver fibrosis is asymptomatic and cirrhosis develops insidiously with non-specific symptoms, so that opportunities for disease modification or cure are missed. Standard biochemical tests of liver function are not specific or sensitive. Liver biopsy is hazardous, inaccurate, subject to sampling error and variation in interpretation. Imaging has a major role in the detection and assessment of liver fibrosis. However all imaging modalities including ultrasound, elastography, cross-section imaging with X-rays or magnetic resonance require access to expensive technology and skilled operators.

Irrespective of the cause of chronic liver disease, progressive liver fibrosis culminates in architectural disruption of the liver by new collagen deposition termed cirrhosis. Once cirrhosis is established the most common and life threatening complications of the cirrhotic state include portal hypertension and hepatocellular cancer. Treatment of the underlying cause of chronic liver disease may prevent or delay the onset of cirrhosis. However once cirrhosis is established, whatever the cause, randomised controlled trials have demonstrated that a number of treatments (such as beta blocker therapy for the treatment of varices and surgery for low-volume hepatocellular cancer (HCC)) are effective at reducing the incidence of complications of cirrhosis. However their effectiveness depends upon cirrhosis being detected early enough to allow them to be delivered before disease is too advanced. Frequently patients present for the first time when these life-threatening complications result in avoidable morbidity, mortality and cost. In the WS3, we sought to identify a ‘pool’ of patients with progressive fibrosis, transitioning to cirrhosis, that could be treated early enough in the course of their disease to reduce the incidence of the serious complications of cirrhosis.

Evidence shows that early detection of varices and treatment with prophylactic use of beta blockers to reduce portal hypertension, or band ligation, reduces morbidity and increases survival. Respected guidelines recommend surveillance for varices because of its benefits and health economic justification. Similarly, early detection of ascites and treatment has been shown to reduce the morbidity associated with bacterial peritonitis from 17% to 2%.
The case for surveillance and early detection of HCC is more contentious with some randomised controlled trials showing evidence of benefit, while others show none. International guidelines now advocate surveillance for HCC.\textsuperscript{54-56} Retrospective analyses have identified criteria, essentially small tumours, associated with better outcomes for HCC resection and liver transplantation, but many patients are diagnosed after the growth of their tumours has ruled them out for curative resection or transplantation.\textsuperscript{57, 58}

Blood tests for fibrosis and cirrhosis are highly attractive having the potential to be automated, highly accurate and reproducible and repeatable at relatively shorter intervals than liver biopsy. Serum markers of liver fibrosis can be divided into those that are “Indirect” that measure liver biochemistry and haematological indices and those that are “Direct” measuring constituents of liver matrix and enzymes involved in fibrogenesis and fibrolysis.\textsuperscript{773, 774} Indirect measures, although useful for some clinical purposes, are subject to the influence of inflammation, drug effects and other co-morbidity. Direct markers of fibrosis have biological plausibility but theoretically may be affected by other fibrotic disorders, however this has not been a major problem in clinical evaluation.

Studies and systematic reviews have demonstrated that single direct markers are less accurate than panels of markers in the detection of liver fibrosis.\textsuperscript{59, 60} One such panel of direct markers is the Enhanced Liver Fibrosis (ELF) test, the only CE marked (EU Regulatory Approval) test for liver fibrosis, measuring constituents of liver matrix (Hyaluronic acid and Procollagen III amino-terminal peptide) and a molecule critical to the regulation of matrix re-modelling (Tissue Inhibitor of Matrix metalloproteinase 1), using sensitive automated ELISA assays designed and manufactured specifically for this purpose.\textsuperscript{61} The 3 individual biomarkers were selected as optimal from among 11 “direct” and 35 “indirect” candidates. The results of the individual assays are combined in an algorithm derived and validated in >1000 cases of liver fibrosis to generate a score that correlates with the severity of liver fibrosis on liver biopsy and subsequently fibroelastography. ELF scores have been shown to be highly predictive of clinical outcomes, including variceal bleeding, ascites, HCC and mortality. Subsequent validation studies in hepatitis C, hepatitis B, fatty liver disease, HIV-HCV co-infection, primary sclerosing cholangitis, and primary biliary cirrhosis have confirmed the performance of the test.\textsuperscript{34-37} Although performance is best in the detection of advanced fibrosis and
The ELF test was developed by Siemens Medical Solutions (formerly Bayer Healthcare) in conjunction with the University of Southampton and iQur Limited. We performed independent evaluations of the ELF test analytic validity.
Chapter 17 - Verification of ADVIA Centaur© ELF Test
analytical performance
The Siemens ELF™ test is an in vitro diagnostic assay that uses an algorithm combining quantitative measurements of serum hyaluronic acid (HA), amino-terminal propeptide of type III procollagen (PIIINP) and tissue inhibitor of metalloproteinase 1 (TIMP-1) to produce a single ELF score which reflects the degree of liver damage in patients with or at risk of cirrhosis. The ELF Test was the subject of the clinical trial described in Chapters 18, 19, 20, 21 and 22 and is available on the Siemens ADVIA Centaur automated analyser.

This Chapter describes an independent verification of the ELF Test’s analytical performance characteristics. It details two independent studies conducted within accredited NHS laboratories: an intra-laboratory study evaluating repeatability, intermediate imprecision and bias against control materials; and an inter-laboratory study evaluating reproducibility. In addition to verifying the manufacturer’s performance claims for precision and bias, recommendations are also made for further evaluation prior to routine clinical implementation. To the authors’ knowledge there have not been any published independent validation/verification studies of the analytical performance of the ELF test.

**Introduction to analytical performance evaluation, precision and bias**

In Vitro Diagnostic (IVD) tests form the basis of ~70% of clinical decision making in the NHS. The accuracy and associated uncertainty surrounding diagnostic testing consequently has a major impact on the overall quality of clinical decisions and subsequent effectiveness of this and other healthcare systems.

Numerous pre-analytical, analytical and biological factors (Figure 57) can contribute to uncertainty in diagnostic testing strategies. These uncertainties cumulate through the measurement system and may eventually affect patient outcomes. The same factors can also introduce bias into clinical trials and contribute towards lack of reproducibility in biomarker research studies.
Performance evaluation, in the form of method validation, is a legal requirement for both commercial and laboratory developed diagnostic tests. Such evaluation is recognized as a critical step in mitigating risk as part of the development of new laboratory methods.\textsuperscript{778} For commercial IVDs, such as the ELF Test, it is the manufacturers’ responsibility to validate the analytical performance of the method and provide objective evidence that the test meets the evidence requirements for the intended use, prior to seeking market approval. This includes consideration of analytical sensitivity, specificity, accuracy, repeatability and reproducibility.\textsuperscript{778}

Prior to introducing any IVD into routine clinical practice, clinical laboratories should also, in order to fulfill national accreditation requirements, independently verify the analytical performance of the test. This is to confirm that the method is performing as claimed by the manufacturer in a routine setting and is necessary to safeguard patient safety as part of clinical governance.

The performance evaluations described here include verification of the ELF Test analytical precision and bias (reflecting “trueness”), which together describe the accuracy of a measurement procedure, as illustrated in Figure 58.\textsuperscript{779}
Figure 58 Schematic illustration of the relationships between precision, bias, trueness accuracy and uncertainty. Modified from Bailey & Barwick, LGC, 2007.

Precision and Imprecision

Precision is defined as “the closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions.” It is usually expressed in terms of standard deviation (SD) and/or coefficient of variation (CV) using the equations:

\[ \sigma = \sqrt{V} \]

\[ \%CV = \left( \frac{\sigma}{\bar{x}} \right) \times 100 \]

\( \sigma \) = Standard deviation

\( \bar{x} \) = Sample mean
V = Sample variance

Precision reflects the random errors inherent in all measurement procedures and includes variability arising from multiple factors (M-Factors) including:

1. **Time**: the time between measurements
2. **Calibration**: how often the equipment is calibrated
3. **Operator**: number of staff carrying out the assay
4. **Equipment**: whether the same equipment and batches of reagents are used

Estimates of precision are strongly dependent upon the conditions in which precision is assessed. Precision is generally evaluated with respect to repeatability, intermediate precision and reproducibility. To assess repeatability, repeated measurements are made while keeping the factors above constant, so that they do not contribute to the imprecision. Reproducibility is assessed by comparing results for the same samples as measured in different laboratories, so that these and additional factors (e.g. environmental, staff training) will also contribute to variation in results. Hence, repeatability and reproducibility represent the minimum and maximum extremes of investigation conditions for precision. It is often helpful to describe precision under conditions somewhere in between repeatability and reproducibility. Such conditions are referred to as intermediate precision conditions and are described in relation to the number of M-factors that differ (M=1, 2, 3 or 4).

**Trueness and Bias**

Trueness is defined as “the closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value”. As Trueness cannot be expressed numerically, it is usually expressed in terms of measurement bias. Measurement bias is defined as “an estimate of systematic measurement error”. It is evaluated by comparing the difference between the mean of replicate measures made using a test method ($\bar{x}$) against the assigned value of a reference material or method ($x_0$), or against consensus mean result ($x_0$) for a group of laboratories (e.g. those participating in an external quality assessment (EQA) programme) using the equations:

$$\text{Bias} = \bar{x} - x_0$$
\[
\%\text{Bias} = \left( \bar{x} - x_0 / x_0 \right) \times 100
\]

**Analytical performance goals incorporating biological variation**

To validate a method one must provide objective evidence that it fulfills the evidence requirements for a specific intended use and is “fit for purpose”. However, defining these requirements remains a challenge, even after several decades of intensive efforts by members of the laboratory medicine community. In 1999 a landmark conference in Stockholm agreed a hierarchical structure for setting analytical performance goals. A recent 2014 conference in Milan revised and refined these, suggesting three approaches based on:

- The effect of analytical performance on clinical outcome (either directly or indirectly)
- Components of biological variation of the measurand, or
- "State of the art"

Analytical performance goals that incorporate biological variation can be calculated as follows:

- Analytical imprecision \((CV_A)\) should be less than 0.5*\(CV_I\)
- Analytical Bias \((B_A)\) should be less than 0.25*(\(CV_I + CV_G\))

\(CV_I\) = within individual variability \(CV_G\) = between individual variability \(CV_T\) = total measurement variability \((CV_T = \sqrt{(CV_I^2 + CV_A^2)})\)

**Single Site Evaluation of the ADVIA Centaur© ELF Test**

As part of the regulatory submission for the ADVIA Centaur© ELF Test, Siemens have undertaken an evaluation of the imprecision of the ELF Score. They performed a full validation of within-run (repeatability), between run and total intra-laboratory (intermediate imprecision) according to the CLSI protocol EP05-A2 “Evaluation of precision performance of quantitative measurement methods”. Samples were assayed in triplicate, twice daily for 20 days (n = 120 replicates per sample). The intermediate imprecision results were performed using one ADVIA Centaur XP and CP system, using one reagent lot and one calibrator lot, this data is presented in Table 94.
Table 94 Manufacturers Claimed ELF Score Within-Run (Repeatability), Between-Run and Total Intra-laboratory (Intermediate) Imprecision

<table>
<thead>
<tr>
<th>ADVIA Centaur</th>
<th>Mean ELF Score</th>
<th>Within Run SD</th>
<th>Between Run SD</th>
<th>Total Intra-Lab SD</th>
<th>Within Run CV</th>
<th>Between Run CV</th>
<th>Total Intra-Lab CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>XP</td>
<td>6.98</td>
<td>0.07</td>
<td>0.04</td>
<td>0.11</td>
<td>1.00</td>
<td>0.57</td>
<td>1.58</td>
</tr>
<tr>
<td>XP</td>
<td>7.12</td>
<td>0.04</td>
<td>0.03</td>
<td>0.08</td>
<td>0.56</td>
<td>0.42</td>
<td>1.12</td>
</tr>
<tr>
<td>XP</td>
<td>8.95</td>
<td>0.03</td>
<td>0.04</td>
<td>0.09</td>
<td>0.34</td>
<td>0.45</td>
<td>1.01</td>
</tr>
<tr>
<td>XP</td>
<td>11.05</td>
<td>0.03</td>
<td>0.04</td>
<td>0.08</td>
<td>0.27</td>
<td>0.36</td>
<td>0.72</td>
</tr>
<tr>
<td>XP</td>
<td>14.51</td>
<td>0.04</td>
<td>0.03</td>
<td>0.08</td>
<td>0.28</td>
<td>0.21</td>
<td>0.55</td>
</tr>
<tr>
<td>CP</td>
<td>7.09</td>
<td>0.04</td>
<td>0.07</td>
<td>0.11</td>
<td>0.56</td>
<td>0.99</td>
<td>1.55</td>
</tr>
<tr>
<td>CP</td>
<td>7.33</td>
<td>0.06</td>
<td>0.03</td>
<td>0.08</td>
<td>0.82</td>
<td>0.41</td>
<td>1.09</td>
</tr>
<tr>
<td>CP</td>
<td>9.08</td>
<td>0.05</td>
<td>0.04</td>
<td>0.08</td>
<td>0.55</td>
<td>0.44</td>
<td>0.88</td>
</tr>
<tr>
<td>CP</td>
<td>11.15</td>
<td>0.05</td>
<td>0.04</td>
<td>0.08</td>
<td>0.45</td>
<td>0.36</td>
<td>0.72</td>
</tr>
<tr>
<td>CP</td>
<td>14.59</td>
<td>0.06</td>
<td>0.04</td>
<td>0.09</td>
<td>0.41</td>
<td>0.27</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Study Aims - Single site evaluation

The primary aim of this study was to verify the claimed repeatability and intermediate imprecision of the ELF Score (Table 94) within a routine NHS clinical laboratory environment. Secondary objectives included verification of ELF Score bias using control materials with assigned values, and assessment of the bias, repeatability and intermediate imprecision of the individual assay components (HA, PIIINP and TIMP-1).

Study Design - Single site evaluation

The imprecision and bias studies were designed according to the ACB guideline “Measurement verification in the clinical laboratory” which is based on the guidelines of the CLSI EP15-A2 “User Verification of Performance for Precision and Trueness”. Five replicates were measured over five days using three clinically relevant concentrations (high, medium and low) of both patient pooled samples and control materials.

Study Methods - Single site evaluation

Sample collection

Serum samples were collected (REC approval 10/H1306/88), processed and stored at - 80°C according to standard methods (SOP03S Serum Plasma Urine Processing v1.0). The samples
were then transferred on dry ice to Leeds Pathology Research and Development Department, Leeds General Infirmary and stored at -80°C until required.

**Sample Preparation**

Approximately 10 anonymised serum samples (Leeds General Infirmary, REC approval 10/H1306/88) were selected from patients with systemic sclerosis and combined so as to yield serum pools at three concentrations spanning the ELF Test assay range (low, medium and high). Samples were selected so as to exclude any with physical or biochemical evidence of haemolysis, lipaemia and icterus, which could interfere in the assays. Each pool was mixed in a universal tube for approximately 5 minutes on a roller mixer at room temperature (RT). Pools were subsequently sub-aliquoted in 250 µL volumes in cryotubes and immediately frozen at -80°C.

**Measurement of Pooled Serum Samples and Reference Controls**

Study sample analysis was conducted during May and June 2015 in the Leeds Blood Sciences Department on one ADVIA Centaur XP (Serial Number 8680). The analytical conditions (e.g. reagent and calibrator lot) remained unchanged between series. The same member of laboratory staff prepared and processed the samples. This reflects the analytical conditions of the intermediate imprecision studies reported by Siemens (Table 94) (i.e. one reagent lot and one calibrator lot), in accordance with CLSI Guideline EP05-A2.788

Frozen serum samples were defrosted at room temperature, vortexed and centrifuged prior to analysis, as per local quality assurance procedures within the blood sciences laboratory. Three levels of each type of pooled serum sample and reference control material (Siemens) were analysed for each component assay and manually combined to produce the ELF Score using the Centaur XP ELF algorithm below:

\[
\text{ELF score} = 2.278 + 0.851 \ln(\text{CHA}) + 0.751 \ln(\text{CPIIINP}) + 0.394 \ln(\text{CTIMP}-1)
\]

**Data Analysis and Verification**

ELF score and component analyte measurements were determined according to the ACB Method Verification Protocol. Imprecision data analysis was obtained using nested ANOVA to determine repeatability and intermediate imprecision using the ACB “Spreadsheet A” (http://www.acb.org.uk/whatwedo/science/best_practice/MV_Terms1.aspx), which is based
upon CLSI EP 15-A2. Spreadsheet A was cross-validated with Analyse-It Software [Method Validation Edition (Leeds, UK)] and produced comparable results. A false rejection rate of 5% was used.

ELF Score and component analyte imprecision was compared with published performance claims. Test performance was also reviewed in line with FDA Bioanalytical Guidelines and the Tumor Marker Quality Requirements Guidelines of the National Academy of Clinical Biochemistry (NACB). These specify intermediate imprecision performance goals of ≤15% CV for immunoassays and ≤5% CV on high precision instruments respectively; and goals for bias of ≤15% of the nominal value, except at the Lower Limit of Quantification (LLOQ) where it should not deviate by more than 20%. Analytical performance goals for imprecision and bias were determined using estimates of total measurement variability (CV\textsubscript{T}) from the ELUCIDATE trial (Chapter 7, Table 14).

There is as yet no certified reference material or reference method for the ELF test. Analysis of bias was therefore verified by comparing the difference between the means (\bar{x}) of five replicates of three manufacturers reference quality control materials (Lot: Low-2418261, Mid-2418262, High-2418263), measured over five days, with their respective assigned values (x₀). Data was analysed according to the ACB Method Verification Protocol. However, as Spreadsheet C only enables inclusion of two replicates per day, statistical analysis was performed as above using the Analyse-it software in accord with the methods specified in CLSI EP 15-A2. Tests for equality and equivalence were both performed. Equality tests assess whether the methods are producing identical results (average bias = 0), while equivalence tests assess whether the bias is within an allowable goal specified by the manufacturer. A false rejection rate of 5% was used.

**Results - Single site evaluation**

**Assessment of imprecision**

The results shown in Tables 95 and 96 confirm that the manufacturer’s claims for ELF Test imprecision are verifiable for both repeatability and intermediate imprecision when using control materials, but not for repeatability in two of three pooled serum samples.
The Levey-Jennings plots in Figure 59 demonstrate a low level of ELF Test imprecision over time against an ELF score range spanning the clinical decision thresholds used within the ELUCIDATE trial (≥ 8.4 randomisation and ≥9.5 for management of cirrhosis). The full range of ELF Scores observed in the ELUCIDATE trial was 6.41 to 17.84 (Mean 9.304, Median 9.11).

Tables 97 and 98 show the repeatability and intermediate imprecision of the ELF component analytes HA, PIIINP and TIMP-1 for control materials and pooled serum samples respectively. The imprecision of the component analytes were worse than that of the ELF Score, with all three components exceeding the NACB ≤5% CV criteria for at least one concentration of control materials and at least two concentrations for pooled serum samples. Furthermore, TIMP-1 exceeded the manufacturers intermediate imprecision claims in two of three control materials and three of three pooled serum samples, whilst PIIINP exceeded both claims at the mid level and HA exceeded the repeatability claim at the high level in pooled serum samples.

Assessment of bias

The results of the assessment of bias are shown in Table 99 and in Figure 60. These suggest that whilst the ELF Score Bias was not equal to zero (P<0.001), it was within the acceptable bias goals assigned by the manufacturer (P<0.001). Similarly, the measured bias of 2/3 HA, 2/3 PIIINP and 3/3 TIMP-1 reference QC materials were not equal to zero (P<0.001), but all were within the acceptable bias goal (P<0.001). However, it is interesting to note that although within the manufacturer’s acceptable range, the percentage bias of TIMP-1 ranged from -21.19% to -20.26%, exceeding the FDA performance goal for bias of <20%.

Determination of analytical performance goals (APG)

The analytical performance goals based on components of biological variation of the ELF score were determined to be:

- APG for \( CV_A = 0.5 \times 4.85 = 2.4\% \)
- APG for \( B_A = 0.25 \times (4.85 + 10) = 3.7\% \)

465
(CV₁ = √[(0.47/9.3*100)^2-1.4^2]) = 4.85% (where σ total measurement variability is 0.47, CVₐ
is <1.4% and mean ELF score is 9.3); CV₉ = 0.93/9.3*100 = 10% (where σ between
individual variability is 0.93 and mean ELF score is 9.3))

<table>
<thead>
<tr>
<th>Table 95 Verification of ELF repeatability and intermediate imprecision for control materials</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean (ng/mL for HA, TIMP, PIIINP)</strong></td>
</tr>
<tr>
<td>----------------------------------------</td>
</tr>
<tr>
<td>ELF (L)</td>
</tr>
<tr>
<td>ELF (M)</td>
</tr>
<tr>
<td>ELF (H)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 96 Verification of ELF repeatability and intermediate imprecision for pooled serum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean (ng/mL for HA, TIMP, PIIINP)</strong></td>
</tr>
<tr>
<td>----------------------------------------</td>
</tr>
<tr>
<td>ELF (L)</td>
</tr>
<tr>
<td>ELF (M)</td>
</tr>
<tr>
<td>ELF (H)</td>
</tr>
</tbody>
</table>
Figure 59 Levey–Jennings charts of ELF Scores across five days: dots represent within-run means; solid line represents the between-run mean; dotted lines represent +/- 1 standard deviation. A= Low QC; B=Mid QC; C=High QC; D=Low Serum Pool; E=Mid Serum Pool; F= High Serum Pool.
Table 97 Verification of HA, PIIINP and TIMP-1 repeatability and intermediate imprecision for control materials

<table>
<thead>
<tr>
<th></th>
<th>Mean (ng/mL for HA, TIMP, PIIINP)</th>
<th>Repeatability CV%</th>
<th>Intermediate Imprecision CV%</th>
<th>Claimed repeatability CV%</th>
<th>Claimed Intermediate Imprecision CV%</th>
<th>Verification Claim (5% significance level)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA (L)</td>
<td>19.71</td>
<td>3.3%</td>
<td>4.5%</td>
<td>4.5%</td>
<td>7.5%</td>
<td>Within Claims</td>
</tr>
<tr>
<td>HA (M)</td>
<td>48.72</td>
<td>2.4%</td>
<td>5.7%</td>
<td>3.6%</td>
<td>7.7%</td>
<td>Within Claims</td>
</tr>
<tr>
<td>HA (H)</td>
<td>201.67</td>
<td>2.6%</td>
<td>4.8%</td>
<td>3.9%</td>
<td>6.6%</td>
<td>Within Claims</td>
</tr>
<tr>
<td>TIMP-1 (L)</td>
<td>89.27</td>
<td>2.8%</td>
<td>9.4%</td>
<td>2.5%</td>
<td>5.1%</td>
<td>Exceeds Intermediate claim</td>
</tr>
<tr>
<td>TIMP-1 (M)</td>
<td>272.34</td>
<td>1.8%</td>
<td>7.3%</td>
<td>1.9%</td>
<td>6.0%</td>
<td>Within claims</td>
</tr>
<tr>
<td>TIMP-1 (H)</td>
<td>499.97</td>
<td>1.8%</td>
<td>8.9%</td>
<td>1.8%</td>
<td>5.2%</td>
<td>Exceeds Intermediate claim</td>
</tr>
<tr>
<td>PIIINP (L)</td>
<td>2.3</td>
<td>3.8%</td>
<td>4.7</td>
<td>5.0%</td>
<td>6.6%</td>
<td>Within Claims</td>
</tr>
<tr>
<td>PIIINP M)</td>
<td>5.68</td>
<td>1.1%</td>
<td>4.51</td>
<td>3.3%</td>
<td>6.8%</td>
<td>Within Claims</td>
</tr>
<tr>
<td>PIIINP (H)</td>
<td>12.48</td>
<td>2.3%</td>
<td>6.0%</td>
<td>2.2%</td>
<td>4.4%</td>
<td>Within Claims</td>
</tr>
<tr>
<td></td>
<td>Mean (ng/mL for HA, TIMP, PIIINP)</td>
<td>Repeatability CV%</td>
<td>Intermediate Imprecision CV%</td>
<td>Claimed repeatability CV%</td>
<td>Claimed Intermediate Imprecision CV%</td>
<td>Verification Claim (5% significance level)</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------------------------------</td>
<td>-------------------</td>
<td>-----------------------------</td>
<td>---------------------------</td>
<td>--------------------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td><strong>HA (L)</strong></td>
<td>11.38</td>
<td>3.1</td>
<td>4.3%</td>
<td>5.2%</td>
<td>5.9%</td>
<td>Within claims</td>
</tr>
<tr>
<td><strong>HA (M)</strong></td>
<td>41.35</td>
<td>3.3</td>
<td>5.52%</td>
<td>3.6%</td>
<td>7.7%</td>
<td>Within claims</td>
</tr>
<tr>
<td><strong>HA (H)</strong></td>
<td>174.26</td>
<td>5.2%</td>
<td>7.4%</td>
<td>3.9%</td>
<td>6.6%</td>
<td>Exceeds repeatability claim</td>
</tr>
<tr>
<td><strong>TIMP-1 (L)</strong></td>
<td>141.34</td>
<td>4.6%</td>
<td>7.4%</td>
<td>1.8%</td>
<td>3.3%</td>
<td>Exceeds both claims</td>
</tr>
<tr>
<td><strong>TIMP-1 (M)</strong></td>
<td>245.82</td>
<td>2.1%</td>
<td>9.6%</td>
<td>1.9%</td>
<td>6.0%</td>
<td>Exceeds intermediate claim</td>
</tr>
<tr>
<td><strong>TIMP-1 (H)</strong></td>
<td>342.06</td>
<td>5.3%</td>
<td>9.7%</td>
<td>1.6%</td>
<td>3.1%</td>
<td>Exceeds both claims</td>
</tr>
<tr>
<td><strong>PIIINP (L)</strong></td>
<td>5.35</td>
<td>2.9%</td>
<td>6.85%</td>
<td>3.3%</td>
<td>6.8%</td>
<td>Within claims</td>
</tr>
<tr>
<td><strong>PIIINP M)</strong></td>
<td>7.78</td>
<td>4.7%</td>
<td>7.8%</td>
<td>1.9%</td>
<td>2.9%</td>
<td>Exceeds both claims</td>
</tr>
<tr>
<td><strong>PIIINP (H)</strong></td>
<td>17.88</td>
<td>2.8%</td>
<td>5.15%</td>
<td>2.2%</td>
<td>4.4%</td>
<td>Within claims</td>
</tr>
</tbody>
</table>
Table 99 Verification of ELF Score and component analyte bias using manufacturers reference QC material

<table>
<thead>
<tr>
<th>Reference Material</th>
<th>Lot</th>
<th>Target (ng/mL for HA, TIMP, PIIINP)</th>
<th>Low</th>
<th>High</th>
<th>Mean (ng/mL for HA, TIMP, PIIINP)</th>
<th>Bias (ng/mL for HA, TIMP, PIIINP)</th>
<th>Bias (%)</th>
<th>Equality Test (5% significance level)</th>
<th>Equivalence Test (5% significance level)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA</td>
<td>Low</td>
<td>20.1</td>
<td>15.07</td>
<td>25.13</td>
<td>19.71</td>
<td>-0.39</td>
<td>-1.94</td>
<td>Not equal to zero</td>
<td>Within goal</td>
</tr>
<tr>
<td></td>
<td>Mid</td>
<td>50.8</td>
<td>38.1</td>
<td>63.5</td>
<td>48.72</td>
<td>-2.08</td>
<td>-4.10</td>
<td>Not equal to zero</td>
<td>Within goal</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>200</td>
<td>150</td>
<td>250</td>
<td>201.67</td>
<td>1.67</td>
<td>0.84</td>
<td>Equal to zero</td>
<td>Within goal</td>
</tr>
<tr>
<td>PIIINP</td>
<td>Low</td>
<td>2.45</td>
<td>1.837</td>
<td>3.063</td>
<td>2.30</td>
<td>-0.15</td>
<td>-6.06</td>
<td>Not equal to zero</td>
<td>Within goal</td>
</tr>
<tr>
<td></td>
<td>Mid</td>
<td>5.96</td>
<td>4.47</td>
<td>7.45</td>
<td>5.68</td>
<td>-0.28</td>
<td>-4.68</td>
<td>Not equal to zero</td>
<td>Within goal</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>12.4</td>
<td>9.3</td>
<td>15.5</td>
<td>12.48</td>
<td>0.08</td>
<td>0.68</td>
<td>Equal to zero</td>
<td>Within goal</td>
</tr>
<tr>
<td>TIMP</td>
<td>Low</td>
<td>95.6</td>
<td>71.7</td>
<td>119.5</td>
<td>75.60</td>
<td>-20.00</td>
<td>-20.92</td>
<td>Not equal to zero</td>
<td>Within goal</td>
</tr>
<tr>
<td></td>
<td>Mid</td>
<td>296</td>
<td>222</td>
<td>370</td>
<td>236.02</td>
<td>-59.98</td>
<td>-20.26</td>
<td>Not equal to zero</td>
<td>Within goal</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>531</td>
<td>398</td>
<td>564</td>
<td>418.48</td>
<td>-112.52</td>
<td>-21.19</td>
<td>Not equal to zero</td>
<td>Within goal</td>
</tr>
<tr>
<td>ELF</td>
<td>Low</td>
<td>7.3</td>
<td>6.73</td>
<td>7.75</td>
<td>7.21</td>
<td>-0.09</td>
<td>-1.26</td>
<td>Not equal to zero</td>
<td>Within goal</td>
</tr>
<tr>
<td></td>
<td>Mid</td>
<td>9.2</td>
<td>8.63</td>
<td>9.65</td>
<td>9.10</td>
<td>-0.10</td>
<td>-1.11</td>
<td>Not equal to zero</td>
<td>Within goal</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>11.2</td>
<td>10.6</td>
<td>11.6</td>
<td>11.13</td>
<td>-0.07</td>
<td>-0.61</td>
<td>Not equal to zero</td>
<td>Within goal</td>
</tr>
</tbody>
</table>
Figure 60 Difference plot showing the bias of three manufacturers reference QC materials: for i. ELF, ii. HA, iii. PIIINP and iv. TIMP-1
Discussion - Single site evaluation

This study has verified the manufacturer’s intermediate imprecision and bias performance claims for the ELF Test, in an accredited NHS laboratory. The results suggest that the ELF Test is a precise and true assay with intermediate imprecision of <1.4% and bias of <1.26%. As the only assay factor that varied was time, the intermediate imprecision presented here is likely to underestimate the total intra-laboratory intermediate imprecision (M-factor=4).

In contrast, the manufacturer’s repeatability claims for the ELF Test were not verified in two of three pooled serum samples. This may reflect specific characteristics of the pooled clinical samples as the manufacturer’s claimed repeatability at these concentrations is highly precise and the observed repeatability of <1% is well below the specified requirements of the FDA and NACB, at ≤15% and ≤5% CV respectively. The ELF Test imprecision was also within the analytical performance goal for imprecision of 2.4% based on components of biological variation.

The imprecision of the individual components were less good and failed to meet all the manufacturers claims and NACB criteria. However, they were all within <10% CV and meet the FDA criteria, so analytical performance is within ranges generally considered to be acceptable for clinical application.

This study has also verified the ELF Test’s claims for bias using reference QC materials. The observed ELF Test %bias was low across all three QCs with <1.26% difference to the assigned value, well below the FDA goal for bias (<20%) and the analytical performance goal for bias of 3.7% based on components of biological variation.

However, the comparable performance of the component analytes was less reassuring. TIMP-1 had the highest % bias of up to -21.19%, which although within the manufacturers acceptable range, exceeded the FDA performance goal of <20%.

It is interesting to note the comparative precision and bias of the ELF Test for HA, PIIINP and TIMP-1. As might be expected, the logarithmic transformation of the biomarker concentrations within the ELF Test also transforms the variance, greatly reducing the percentage CVs. Whilst this appears to be the main reason for the perceived improvements in precision, benefit may also be derived from the “averaging” effect of the triple biomarker
panel and the preferential weighting for HA, the most precise component. It is not immediately clear from the instructions for use whether clinical laboratories should undertake quality control (QC) only for the ELF Scores, or for the ELF score and each of the component analytes. Results of the verification reported here suggest that one or more individual component analytes might fail QC, even though the ELF scores remained within acceptable limits. This might lead to rejection of an unnecessarily high number of tests. In view of the increasing number of complex decision algorithms that are being applied in laboratory medicine, this is an important issue that requires further research and guidance.

A limitation of this study was the use of pooled patient samples. Whilst this is not uncommon in precision studies, as the volume of available sample material is limited, the pooling of patient samples may dilute out any interfering substances and could also introduce interactions that would not naturally occur within an individual sample. The lack of available certified reference materials or reference methods was also a limitation of this study. Bias was therefore assessed using both manufacturer’s reference QC materials in the intralaboratory study as recommended by Khatami et al.\(^789\)

A further potential limitation was that components of biological variation (e.g. CVI and CVG) were derived from the same study and not from a prospectively designed and powered biological variation study.\(^236, 793\) However, the data presented should provide a realistic estimation of the total measurement variability and are more likely to over- than underestimate variability.

**Multi-Site Evaluation of the ADVIA Centaur\(^©\) ELF Test**

When introducing new assays, such as the ELF Test, to multiple sites within a health care system, it is essential to demonstrate that good between-laboratory agreement can be achieved across multiple laboratories. Such inter-laboratory method performance studies can helpfully contribute to the validation of analytical methods as they incorporate assessment of the additional variance encountered when comparing results between laboratories.\(^794\) Once a test has been accepted into clinical practice, continued performance surveillance is usually the responsibility of external quality assessment (EQA) providers.\(^795\)

Inter-laboratory and EQA studies usually involve distributing aliquoted samples of the same material (usually pooled human serum, plasma or urine) to laboratories participating in the
study. Participants measure the required test or tests in each of the specimens and return their results to the coordinating centre. Analysis of the submitted results enables calculation of between-laboratory agreement (and sometimes within-laboratory agreement). For heterogeneous analytes such as the ELF test components the target values are usually consensus means.

**Study Aim – Multi-site evaluation**

With the aim of assessing the feasibility of introducing the ELF test into NHS diagnostic laboratories, an inter-laboratory study was carried out to determine between-laboratory agreement (reproducibility) of all components of the ELF test.

**Study Design Multi-site evaluation**

Eight NHS diagnostic laboratories were invited to participate in the inter-laboratory study, for which availability of a Siemens ADVIA Centaur system was required. Specimens were sent to each participating laboratory together with 10 serum samples.

**Test materials and distribution:**

Suitable anonymised left-over clinical samples from patients with liver disease were identified by staff in the Blood Sciences Department at Leeds General Infirmary. Ten 5.0 mL pools were prepared by combining two to three of the clinical samples. 0.5 mLs of each pool was transferred to each of ten pre-labelled tubes (Specimens E001 through E010). Sets of ten specimens were packaged according to UK NEQAS (Edinburgh) procedures and sent, together with a personalised results sheet (Appendix 2) to each of the participating laboratories. Specimens were sent at ambient temperature in order to mimic routine clinical practice. Participants were requested to assay the specimens following the procedure recommended by Siemens for the ELF tests and to return their results sheets and assay print-outs by e-mail or fax to the UK NEQAS [Edinburgh] unit. They were asked to submit results for each individual component of the test as well as the final ELF score.

**Laboratory procedures**

The ELF test was set up on the individual analysers at the different sites with assistance from staff from Siemens Diagnostics, as is usual in clinical laboratories. Siemens also kindly
supplied all reagents and controls required. The specimens were analysed twice, in singlicate twice, in two separate runs. Manufacturers’ instructions were followed in all cases and three kit controls included in each run.

**Results - Multi-site evaluation**

Results obtained from the individual sites are shown in full in Appendix 2 and summarised in Table 100 below. Between-run agreement for all analytes and all specimens was generally very good, with CVs for each pair of runs <10% for all sites except Site 6 where a hardware issue was identified with the analyser which resulted in imprecise results for hyaluronic acid (HA, one of the components of the ELF Score) (Appendix 2). Between-method agreement for the final ELF score was excellent with between-method CVs for the ten specimens as measured at eight sites ranging from 0.4 to 1.2% (Table 100).

**Table 100 Between-laboratory agreement of the ELF score as determined for Specimens E001 through E010 in eight accredited NHS clinical laboratories**

<table>
<thead>
<tr>
<th>Specimen number</th>
<th>Low</th>
<th>Mid</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>E001</td>
<td>7.6</td>
<td>9.1</td>
<td>11.0</td>
</tr>
<tr>
<td>E002</td>
<td>7.6</td>
<td>9.2</td>
<td>11.1</td>
</tr>
<tr>
<td>E003</td>
<td>7.7</td>
<td>9.2</td>
<td>11.1</td>
</tr>
<tr>
<td>E004</td>
<td>7.6</td>
<td>9.2</td>
<td>11.0</td>
</tr>
<tr>
<td>E005</td>
<td>7.6</td>
<td>9.1</td>
<td>11.0</td>
</tr>
<tr>
<td>E006</td>
<td>7.7</td>
<td>9.3</td>
<td>11.1</td>
</tr>
<tr>
<td>E007</td>
<td>7.6</td>
<td>9.1</td>
<td>11.0</td>
</tr>
<tr>
<td>E008</td>
<td>7.6</td>
<td>9.1</td>
<td>11.0</td>
</tr>
<tr>
<td>E009</td>
<td>7.6</td>
<td>9.1</td>
<td>11.0</td>
</tr>
<tr>
<td>E010</td>
<td>7.6</td>
<td>9.1</td>
<td>11.0</td>
</tr>
</tbody>
</table>

**Discussion - Multi-site evaluation**

The multi-site study described here considered primarily verification of analytical imprecision. Results confirm that the ELF test is reproducible. The between-laboratory imprecision of <1.2% observed across eight centres is well within the analytical performance goal for imprecision of 2.4% based on components of biological variation and is also well within the specified requirements of the FDA and NACB of ≤15% and ≤5% CV respectively. However, prior to its implementation into routine clinical practice, further work would be desirable to confirm the analytical specificity (including interference and cross-reactivity), analytical sensitivity, limits of detection/quantitation and the measuring range of the tests.
Further characterisation of pre-analytical and biological factors including: within-individual variation, time and temperature from sample collection to analysis or stabilization, type of collection tube, storage temperature, duration and aliquot volume, freeze-thaw cycles and specific systematic factors (e.g. medications, fasting, alcohol and smoking) should also be conducted prior to implementation. These further studies are planned as part of an on-going NIHR Career Development Fellowship (Dr Del Galdo), which will include assessment of the clinical utility of HA, PIIINP and TIMP-1 in patients with systemic sclerosis.

**Study conclusions**

The work presented here demonstrates that the ADVIA Centaur© ELF Test assay performs as specified by the manufacturer and suggests that its transfer to routine use in NHS laboratories is feasible.

Provided the manufacturer’s instructions are followed and suitable quality control procedures implemented, the analytical performance of the ELF test should be appropriate for clinical use.

Further high quality studies of pre-analytical and biological requirements should be conducted in order to determine the total ELF test measurement uncertainty.
Chapter 18 - Design and Set up of the ELUCIDATE Trial
Study design

ELUCIDATE was a multi-centre individually RCT which aimed to determine whether the use of the ELF test in addition to standard clinical monitoring significantly alters the diagnostic timing and subsequent management of cirrhosis of the liver, compared to standard clinical monitoring alone, in order to reduce the incidence and consequences of serious complications and improve outcomes for patients and service provision. ELUCIDATE was a randomised controlled trial of testing for cirrhosis using the ELF test in patients with chronic liver disease and pre-cirrhotic moderate to severe fibrosis (as classified by clinical, laboratory, or histological evidence), due to viral liver disease, non-alcoholic fatty liver disease, alcoholic liver disease, Primary Biliary Cirrhosis (PBC), Primary Sclerosing Cholangitis (PSC), autoimmune hepatitis (AIH), haemochromatosis, or combinations of these diseases. ELF scores have been evaluated in previous studies to relate biopsy evidence to ELF results (Table 101).

Table 101 ELF scores and fibrosis staging

<table>
<thead>
<tr>
<th>Fibrosis stage (Ishak)</th>
<th>ELF Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal/Mild (F0-F2)</td>
<td>&lt;8.37</td>
</tr>
<tr>
<td>Moderate (F3)</td>
<td>8.37-8.73</td>
</tr>
<tr>
<td>Moderate/Severe (F4)</td>
<td>8.74-9.12</td>
</tr>
<tr>
<td>Severe (F5)</td>
<td>9.13-9.49</td>
</tr>
<tr>
<td>Cirrhosis (F6)</td>
<td>≥9.5</td>
</tr>
</tbody>
</table>

It can be seen that scores of under 8.4 are not associated with fibrosis on biopsy and they carry only a very small risk of serious liver complications (Figure 64, Chapter 19). Scores of over 9.5, however, are associated with cirrhosis on biopsy and a higher risk of serious complications (Figure 64).

This trial aimed to answer the following questions. Does the use of serum markers of liver fibrosis:

- permit earlier detection of liver cirrhosis in patients with Chronic Liver Disease (CLD) to allow earlier interventions?
- affect the process of care, through a) increased use of beta-blockers/band ligation of varices to prevent haemorrhage; b) increased use of endoscopy and ultrasound/AFPs
to detect HCC at a surgically curable stage; and c) effective early treatment to normalise Liver Function Tests (LFTs) in patients with Hepatitis B and Hepatitis C.

- result in patient benefit through improved survival and reduced liver-related morbidity and mortality?
- improve the cost-effectiveness of the management of end-stage liver disease?

There was also a qualitative exit study to investigate patient understandings of clinical biomarkers, experiences of testing, acceptability, perceived utility, and motivations for testing.

The design of the trial is summarised in Figure 61.

The ELUCIDATE final study protocol is provided as supplementary material.
Figure 61 - The ELUCIDATE flow chart: Flow chart showing ELUCIDATE recruitment, randomisation and follow-up procedures.
Ethical approval and research governance

Ethical approval for the RCT was given by Leeds Central Research Ethics Committee, later known as Yorkshire and Humber – Bradford Leeds Committee, (main REC) on 2\textsuperscript{nd} February 2010 (reference number 10/H1313/2). Participating sites were required to have obtained local management approvals and undertaken a site initiation meeting with the central coordinating clinical trials unit (Clinical Trials Research Unit (CTRU), University of Leeds) prior to the start of recruitment into the trial. The trial was registered with the International Standard Randomised Controlled Trial (ISRCTN) Register (ISRCTN74815110).

A summary of the changes made to the original protocol is given in Appendix 3.

Participants

The trial recruited patients with CLD from liver clinics in secondary care through the National Institute for Health Research Clinical Research Network (NIHR CRN) Comprehensive Clinical Research Networks (CCRNs).

Inclusion criteria for registration

Patients were considered eligible for registration if they met all of the following criteria:

- aged $\geq$18 years and $<$75 years
- had CLD due to any aetiology, including viral hepatitis C or B, non-alcoholic liver disease, alcoholic liver disease, primary biliary cirrhosis (PBC), primary sclerosing cholangitis (PSC), autoimmune hepatitis (AIH), haemochromatosis or combinations of these diseases, with no diagnosis of cirrhosis.
- had a life expectancy of $>$six months
- were likely to comply with the follow-up schedule
- able to provide written informed consent

Exclusion criteria for registration

Patients with any of the following criteria were not eligible for registration into the trial:
• imaging, histological or laboratory (other than ELF) diagnosis of cirrhosis /portal hypertension as evidenced by any one of the following:
  o Imaging evidence of portal hypertension (spleenomegally, varices or ascites)
  o Liver biopsy diagnostic of cirrhosis (Ishak F6 or equivalent)
  o Thrombocytopenia (platelets < 100 x 10^9/L)
  o Hypoalbuminaemia (albumin < LLN)
• acute liver injury or acute liver failure (hepatic dysfunction <6 months in duration)
• an ongoing or previous episode of hepatic decompensation (acute on chronic liver failure) including: encephalopathy, variceal bleeding, ascites, jaundice or liver synthetic dysfunction
• an established diagnosis of hepatocellular cancer or elevated alpha fetoprotein without investigation to exclude hepatocellular cancer
• Patient being treated with heparin (ELF test cannot be performed).
• Previously screened and found ineligible for the ELUCIDATE Trial

**Screening and consent procedure**
Nurses reviewed their caseload for potentially eligible participants. Subjects fulfilling the eligibility criteria above were invited to participate in the study. Wherever possible, eligible patients were sent a Patient Information Summary (PIS) to consider prior to their next clinic appointment.

At the patient’s next clinic visit, the patient was provided with the full Patient Information Leaflet (PIL) and further verbal details of the trial. Assenting patients were formally assessed for eligibility and invited to provide informed, written consent to registration, subsequent randomisation (if eligible at that time), and long-term follow-up via routine data sources of the NHS Information Centre (now known as the Health and Social Care Information Centre). Patients were permitted to have more time to consider trial participation, and if they subsequently assented, eligibility assessments and consent were undertaken at a later clinic visit. All participants were informed that they were free to withdraw at any time without reason and without it affecting the quality of their care.

**Registration and baseline ELF test**
Consenting patients were registered into ELUCIDATE via a 24 hour telephone registration system and provided a fasted serum sample for an ELF test (patients should have refrained from eating a large meal in the 2 hours prior to providing the sample). The ELF test was sent off to a central laboratory (iQUR) for analysis.

Eligibility for randomisation
Only patients with an ELF score of above a pre-defined threshold (denoting at least moderate fibrosis), and who had no clinical, histological or laboratory diagnosis of cirrhosis were eligible for randomisation. As knowledge of ELF scores may cause patients or clinicians to modify their behaviour which might influence disease progression and result in confounding, in order to ensure equipoise, results were fed back to the investigator simply as a) below the threshold and not eligible for randomisation, or b) equal to or above the threshold and eligible for randomisation.

In March 2011 after 43 patients had been randomised, the ELF threshold for randomisation was amended from 11.0 to 8.4 to incorporate patients with identified risk of progression to cirrhosis and severe complications (see Figure 64, Chapter 19).

Randomisation
Patients with a registration ELF score equal to or above the threshold (originally 11.0, amended to 8.4 after March 2011) were invited back to clinic for a randomisation visit. The randomisation visit should have occurred as soon as possible following receipt of the registration test results and preferably within 6 weeks of the registration visit, but up to 12 weeks was permissible. If more than 12 weeks had passed since registration, a repeat ELF test was taken to ensure that the patient remained eligible for the trial, and had not progressed to cirrhosis. At the randomisation visit, patients were assessed to ensure their liver disease had not progressed to clinically evident cirrhosis in the interval from their registration visit and were asked whether or not they were still happy to continue participating in the trial and willing to be randomised. Assenting patients judged to still be pre-cirrhotic were individually randomised at the end of their baseline assessments.

Randomisation was undertaken using an automated 24 hour telephone randomisation system, which was administered remotely. The randomisation service was provided by the Clinical
Trials Research Unit (CTRU) at the University of Leeds, a United Kingdom Clinical Research Collaboration (UKCRC) registered trials unit. A computer-generated minimisation program incorporating a random element was used to ensure treatment groups were well-balanced for the following characteristics:

- Centre
- Age (≥ 18 to < 40, ≥ 40 to < 65, ≥ 65 to < 75)
- Gender (Male, Female)
- Baseline ELF score: (11 to 11.49, 11.5 to 11.99, 12 to 12.49, 12.5+) or (≥ 8.4 to < 9.5, ≥ 9.5 to < 11.5, ≥ 11.5 to < 12.5, ≥ 12.5, from Protocol version 5.0 onwards)
- History of high alcohol consumption (at any time), defined as > 6 units (60 grams of alcohol)/day for 12 months or more for males and > 4 units/day for 12 months or more for females (Yes, No)
- Current alcohol consumption per day (Males: 0 units (teetotal), < 3 units (light), 3-6 units (moderate), > 6 units (high); Females: 0 (teetotal), < 2 units (light), 2-4 units (moderate), > 4 units (high))
- Type of CLD (ALD, Viral, Unknown/Other, NAFLD)

Patients were randomised to one of two treatment groups on a 1:1 ratio: standard clinical monitoring plus ELF test monitoring (intervention arm), or standard clinical monitoring alone (non-intervention arm). For patients randomised to the intervention arm, if their ELF score at registration was above the threshold for cirrhosis diagnosis, the randomisation system also notified the caller of that so that management of cirrhosis could begin.
Quality assurance of ELF test

All sites were issued with a Sample Processing Standard Operating Procedure. Samples were only used to determine eligibility for randomisation if they had been kept at room temperature for no longer than 2 days between being taken and arriving at iQUR. If shipping delays were anticipated (e.g., at a weekend), the serum sample was stored in the fridge and shipped when delivery within 2 days was possible. If a sample had been kept at room temperature for more than 2 days from the time it was taken, then a repeat sample was requested.

Patients were requested to refrain from eating a large meal in the 2 hours prior to providing the blood sample for each ELF test. For the ELF sample collected at the randomisation visit, patients were requested to have fasted (gone without food for more than 4 hours), to allow for Glucose ± HOMA-IR testing.

Treatment group allocation

Screening for cirrhosis with standard clinical monitoring
Patients allocated to the standard clinical monitoring arm were seen in clinic every 6 months and monitored as per standard practice. If the patient was deemed to be cirrhotic on clinical criteria (by examination, on the basis of laboratory tests (other than ELF) or through imaging), cirrhosis management commenced.

Screening for cirrhosis with standard clinical monitoring plus ELF test
Patients allocated to the intervention arm were seen in clinic every 6 months and monitored as per standard practice. In addition, they also had their ELF score measured every 6 months by the central laboratory. If the patient was deemed to be cirrhotic on clinical criteria (by examination, on the basis of laboratory tests (other than ELF) or through imaging), cirrhosis management commenced. If the ELF score was above a pre-defined threshold the patient was deemed to be cirrhotic. The investigator was informed that the patient was above the threshold, and the patient was recalled into clinic as soon as possible for cirrhosis management to commence.
The ELF threshold for cirrhosis was originally $\geq 12.5$, but this was changed to $\geq 9.5$ in protocol version 5.0 onwards (March 2011).

**Data collection and management**

Trial data were recorded by research staff on Case Report Forms (CRFs) and submitted to the CTRU. Sites were provided with guidance on the schedule of CRFs, data to be collected, and completion of CRFs. Data were entered into the trial database, using Infermed’s MACRO® Electronic Data Capture platform, by CTRU staff. A number of manual and in-built database cross checks were routinely performed to check for missing and inconsistent data items, which were reported back to sites for resolving at the earliest opportunity.

**Baseline registration assessment**

Baseline assessments consisting of physical examination, medical history and demographics were conducted in the month prior to registration or at the registration visit. At the randomisation visit, a blood sample was taken for the ELF test, and the patients completed the EQ-5D™ and SF-12v2™ health questionnaires.

**Baseline randomisation assessment**

At this visit, patients had another ELF sample taken and completed the EQ-5D™/SF12v2™ and Health Usage Questionnaire. Patients were not told their randomisation allocation until after they had completed the questionnaires. In addition, patients had the following assessments performed: LFTs, FBC, INR, glucose.

**Follow-up**

From the date of randomisation, patients underwent follow-up assessments every 6-months until 30 months post randomisation, unless they were diagnosed as cirrhotic. At each follow up visit all patients underwent physical examination (weight, vital signs), medical history (including details of concomitant disease and medication), blood tests (simple LFTs, platelets, albumin, and clotting) and were required to complete EQ-5D™ SF-12v2™ and
Health Usage Questionnaire. In addition, patients randomised to the follow-up arm with ELF testing also underwent blood sample collection at each follow-up visit and the sample sent to the central iQUR laboratory for ELF testing. Patients were requested to refrain from eating a large meal in the 2 hours prior to providing the sample. Where patients were diagnosed as cirrhotic (either by ELF test or clinical means) within 30 months post randomisation, patients in both treatment arms were required to attend an initial post-cirrhotic follow-up assessment at 3 months post diagnosis, then every 6 months until 30 months post randomisation.

Follow-up and management of patients diagnosed with cirrhosis

The trial protocol included recommendations for the management of varices, ascites and HCC after a diagnosis of cirrhosis, based on appraisal of national and international guidelines, but sites were permitted to follow their own established protocols provided that these were documented and adhered to for all study participants, and included as a minimum: ultrasound scanning (USS), oesophagastroduodenoscopy (OGD) and measurement of alpha-fetoprotein (AFP) levels.

All patients with a diagnosis of cirrhosis were required to have an oesophagastroduodenoscopy (OGD) as screening for varices within 3 months of their cirrhosis diagnosis, unless they had had an OGD in the previous 18 months, in which case the next OGD should have occurred within 18 months of the previous OGD unless clinically indicated sooner. If the previous OGD did not identify oesophageal varices, subsequent OGDs were repeated every 18 months. If small oesophageal varices at OGD were identified, OGDs were repeated every 6 months to look for variceal progression. For large oesophageal varices that were being treated, the timing of subsequent OGDs was dictated by local guidance.

Moderate or large oesophageal varices should have been banded as primary prophylaxis, with banding repeated weekly until the varices were obliterated. Alternatively patients could be treated with non-cardioselective beta-blockers as primary prophylaxis. As secondary prophylaxis, bleeding oesophageal varices should have been banded weekly until obliterated and in addition, patients considered for treatment with beta-blockers, unless contraindicated.
For prophylaxis of spontaneous bacterial peritonitis, all patients with ascites were treated with Norfloxacin 400mg od, or alternative antibiotic as per local protocol.

All patients diagnosed with cirrhosis were required to have AFP measured and a USS performed for HCC screening within 3 months of a diagnosis of cirrhosis, unless they had had a USS or AFP test within the previous 6 months, in which case the next USS or AFP test should have been performed within 6 months of the previous scan, unless clinically indicated sooner. The AFP test was repeated every 6 months. If the previous USS did not identify any lesions and the patient’s AFP level remained stable, subsequent USS scans were repeated every 6 months. Any space occupying lesions, equivocal USS or rising AFP in the absence of a lesion on ultrasound were followed by triple phase CT and/or MRI scans.

Suspected HCC was managed according to local, national, and international guidelines and the management documented in the patient’s CRFs.

Patients were considered for liver transplantation if they had a solitary lesion measuring less than 5 cm in diameter, or 3 lesions measuring less than 3 cm in diameter, no evidence of extrahepatic manifestations and no evidence of vascular invasion. If a patient underwent liver transplantation, further 6 monthly follow-up ceased.

All other patients were considered for therapeutic interventions as per local protocols.

Endpoints

Primary endpoint (according to protocol v7.0, 30 January 2013)

1. Time from randomisation to occurrence of first severe complication

   Severe complications are defined as:
   1. Variceal haemorrhage confirmed by one of the following:
      a) visualisation through endoscopy
      b) imaging
      c) post-mortem
2. Spontaneous bacterial peritonitis:
   Ascites confirmed by:
   a) Imaging and/or
   b) aspiration
   and
   Infection confirmed by:
   a) microscopy and/or
   b) culture

3. Hepatocellular cancer (HCC) beyond the Milan criteria.
   N.B, For the purposes of the trial, cases of HCC falling within the Milan criteria are not regarded as endpoints as they are regarded as treatable.

4. Encephalopathy - grade 3 or 4 defined using the Westhaven criteria (Table 102)

5. Liver-related mortality. Any of the following:
   a) any mention of liver disease in part one of the death certificate
   b) death due to hepatocellular cancer (HCC)
   c) death due to liver failure
   d) death due to bleeding from portal hypertension
   e) death due to hepato-renal syndrome
   f) death due to sepsis occurring as a result of chronic liver disease
   g) death due to spontaneous bacterial peritonitis
   h) death due to encephalopathy
Table 102: Westhaven criteria for semiquantitative grading of mental state

<table>
<thead>
<tr>
<th>Grade</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 1</td>
<td>Trivial lack of awareness</td>
</tr>
<tr>
<td></td>
<td>Euphoria or anxiety</td>
</tr>
<tr>
<td></td>
<td>Shortened attention span</td>
</tr>
<tr>
<td></td>
<td>Impaired performance of addition</td>
</tr>
<tr>
<td>Grade 2</td>
<td>Lethargy or apathy</td>
</tr>
<tr>
<td></td>
<td>Minimal disorientation for time or place</td>
</tr>
<tr>
<td></td>
<td>Subtle personality change</td>
</tr>
<tr>
<td></td>
<td>Inappropriate behaviour</td>
</tr>
<tr>
<td></td>
<td>Impaired performance of subtraction</td>
</tr>
<tr>
<td>Grade 3</td>
<td>Somnolence to semistupor, but responsive to verbal stimuli</td>
</tr>
<tr>
<td></td>
<td>Confusion</td>
</tr>
<tr>
<td></td>
<td>Gross disorientation</td>
</tr>
<tr>
<td>Grade 4</td>
<td>Coma (unresponsive to verbal or noxious stimuli)</td>
</tr>
</tbody>
</table>

Secondary Endpoints (according to protocol v7.0, 30 January 2013)

- Time from diagnosis of cirrhosis (by ELF or clinical means) to incidence of first severe complication
- Time from randomisation to diagnosis of cirrhosis by ELF or clinical means (to allow instigation of prophylaxis and screening)
- Process outcomes**, namely:
  - treatment with beta-blockers/band ligation (BB/BL) of varices
  - use of endoscopy and ultrasound/AFP tests
  - treatment to normalise LFTs in patients with Hepatitis B and Hepatitis C.
- Detection and timing of complications following cirrhosis, including:
  - detection of small varices
  - detection of large varices
  - incidence of treatable hepatocellular cancer (HCC)
- All causes of mortality
- Specific liver-related morbidity
• Economic evaluation of the ELF test in the early detection of cirrhosis and as such in the initiation of measures to reduce the incidence of severe complications following cirrhosis.
• Quality of Life (QoL)
• Proportion of non-randomised patients (ELF <8.4) who go on to develop cirrhosis (diagnosed by clinical means) within the follow-up period.

** The process of care outcomes were added as secondary endpoints by the way of a protocol amendment in January 2013, as part of the NIHR-approved trial extensions. The overall aim, and primary outcome, of the study is to reduce severe complications and improve survival in this patient population through the use of ELF tests. For this to happen the use of such tests would have to affect clinical practice, i.e. the process of care would have to change. This is a necessary step in order to expect an improvement in these primary outcomes, and was by no means guaranteed.

**Health-related quality of life**
Health-related quality of life was assessed at registration, randomisation and at 6-monthly intervals post randomisation for 30 months (five follow-up visits), using the Short Form questionnaire-12 items (SF-12).

**Long term follow-up**
All patients who were registered into to the trial (whether randomised or not) will be flagged with the Health and Social Care Information Centre for longer term morbidity and mortality data.

**Statistical analysis**
The analysis and reporting of this trial was undertaken in accordance with Consolidated Standards of Reporting Trials (CONSORT) guidelines. The original sample size calculation is given in detail in Appendix 4. This sample size calculation was not straightforward without the availability of the simulation approach, derived later on as part of the methodological WorkStream, and described in Chapter 19. All statistical analysis was undertaken using SAS software version 9.4, following a predefined analysis plan agreed with the Trial Steering
Committee. The primary comparative analyses between the randomised groups were conducted on an intention to treat (ITT) basis without imputation of missing data.

**Health economic methods**

The main trial was supplemented with an economic evaluation consisting of two components:

i) A within-trial economic evaluation to compare the observed costs and outcomes of the cohort of patients randomised to ELF guided detection and management with those of a cohort of patients randomised to standard care, from the perspective of the NHS and Personal Social Services.

ii) A long-term economic evaluation using the clinical trial outcome and resource utilisation data to update the parameters of the pre-existing lifetime horizon cost effectiveness model and subsequent calculation of the expected incremental cost effectiveness ratio ELF guided detection and management, using a lifetime horizon.

At the outset of the trial, the primary outcome measure for the within-trial economic evaluation was Quality Adjusted Life Years. Quality of Life weights will be calculated using the EQ-5D™ algorithm. Life years lived will be obtained from the mortality data collected at the end of the clinical trial. Costs and outcomes will be discounted at 3.5% per annum. All costs will be indexed to the trial start year (2009) using the NHS Pay and Prices Index. The primary result of the economic evaluation will be the incremental cost effectiveness ratio of ELF guided detection and management vs. standard care (ICER). This will be calculated as the difference in the mean cost of the interventions divided by the difference in the mean outcomes.

Parameter uncertainty will be examined using a non-parametric bootstrap simulation.

Results of the probabilistic sensitivity analysis will be presented as the expected incremental cost effectiveness ratio; a scatter plot on the cost effectiveness plane and as a cost effectiveness acceptability curve (CEAC). The Expected Net Benefit of ELF guided detection and management will be calculated for a range of values of lambda, including £5,000, £15,000, £20,000 and £30,000.

**Secondary within-trial economic evaluation**
The secondary within-trial economic evaluation will substitute SF-12v2™ quality of life weights for the EQ-5D™ quality of life weights used in the primary analysis. In all other regards the secondary within-trial economic evaluation will be identical to the primary within-trial economic evaluation.

**Long term economic evaluation**

The clinical trial outcome and resource utilisation data will be used. Parameter uncertainty will be examined using Monte Carlo simulation. Results of the probabilistic sensitivity analysis (PSA) will be presented as the expected incremental cost effectiveness ratio; a scatter plot on the cost effectiveness plane and as a cost effectiveness acceptability curve (CEAC). The Expected Net Benefit of ELF guided detection and management will be calculated for a range of values of lambda, including £5,000, £15,000, £20,000 and £30,000.

Delayed set up time and initially slow recruitment resulted in two successful applications for one year no cost extensions. The analysis plan required evaluation of changes in the process of care of patients as a consequence of the use of ELF monitoring. Final evaluation of the impact on QALY will follow the collection of long term outcomes including survival, from cancer registry and NHS information sources.
Chapter 19 - Value of Modelling Development, Modification and Extension of the ELUCIDATE Trial
Background

The ELUCIDATE trial was planned to run for 5 years, allowing sufficient time to randomise 1,000 patients and follow them using the active monitoring intervention with ELF based testing for 30 months. The trial began late due to delays in set-up, and by November 2012, 25 months after the first participant was recruited, only 530 patients had been randomised. At that time the Applied Programme Board conducted a site visit and raised doubts concerning the project team’s ability to complete adequate recruitment in the remaining period of grant funding and accepted the investigators’ request and recommendations for a one year extension and an additional new focus to examine process-of-care outcomes. The Board recommended early closure of the trial.

The Trial Management Group had identified the shortfall in recruitment and had worked with the 20 established sites to promote the trial, but also opened 17 additional sites bringing the total open to 37. Between the time of the site visit and the closure of recruitment which was recommended by the Board, there was an upsurge of recruitment from new and established centres and a further 348 patients were randomised to the study in 4 months (39.6% of the total recruited in 14.8% of the recruitment period) bringing the total of patients randomized and in follow-up to 878.

Rather than restricting the trial to investigate diagnostic performance and process outcomes only, this number of patients was sufficient to provide adequate power to enable completion of the trial in accordance with the original protocol, and assessment of the clinical and health service outcome endpoints identified in the original protocol that was funded as part of the Applied Programme grant application. However this required a further 12 months study extension to permit all enrolled patients to complete 30 months exposure to the monitoring and intervention in the trial (see Figures 62 and 63).

It was recognised that the delivery of the evaluation of the impact of ELF testing on the clinical and healthcare outcomes will still (as originally approved) require the collection of “routine data” from death certification, ONS, HES, cancer and transplant registry data at 5 years.
Accordingly the Project Team submitted an application for a study extension to permit the ELF monitoring intervention and collection of data on all patients for the full 30 months planned in the original protocol. This was a key requirement to allow completion of the trial and evaluation of the impact of ELF testing on the clinical and healthcare outcomes. The case was made for the extension to complete the trial based on the rapid final surge in recruitment and a modelling of the likely impact of ELF testing allowing revised estimates of study power. This was approved by NIHR.

We performed modelling incorporating data from previous studies of disease progression in similar cohorts of patients with chronic liver disease investigated for chronic liver disease with ELF and this was the basis of all previous power calculations. However, importantly at the point where the extensions were approved we were able to analyse preliminary data from the ELUCIDATE trial, allowing greater confidence in both the models themselves, and in the results and conclusions that can be drawn from them.

The model represents a significant innovation allowing us to accurately predict the development of cirrhosis (Figure 62).

*Figure 62*
*A comparison of the model prediction of the identification of patients as cirrhotic, with the observed patterns in the ELUCIDATE trial*

![Figure 62 Image](image_url)
The modelling also allowed us to demonstrate the impact on the follow up and monitoring using ELF tests that would occur with and without the one year extension (Figure 63).

**Figure 63**

*Estimated pattern of development of cirrhosis without a trial extension – showing follow-up times assuming intervention finishes at end of October 2014*

The ELUCIDATE trial, like all trials, depends for its statistical power on the number of events and the effect size. The 30 month period of active follow up within the trial was both a period of case discovery and data collection but was also, vitally, the key monitoring intervention within the trial which could produce a clinical effect. We were testing the performance of the ELF test as a means of discovering cases of cirrhosis and thereby influencing the behaviour of clinicians in delivering, as a consequence of the identification of cirrhosis by ELF criteria, the appropriate investigations and treatments necessary for patients in the diagnostic cirrhotic category.

The shortening of the period of “active follow up” therefore would have had an effect on case discovery and would also have influenced the Effect size by reducing the period of the intervention (ie ELF monitoring and treatment of cirrhosis once detected). We expected a
reduction in Effect size as well as reductions in case discovery if we were unable to complete
the period of active follow up/intervention.

We know from the existing literature that when patients are diagnosed with cirrhosis,
appropriate monitoring and treatment for the major events such as bleeding or liver cancer,
results in a reduction of such events by 50% in randomised controlled trials. However in the
ELUCIDATE trial we always carefully considered whether clinical behaviours would change
sufficiently and would result in the full effect in the cirrhotic patients. Allowing for this in
power calculations for the long term clinical end points, we always mitigated the effect size
from the 50% ceiling which is apparent in the literature, down to 40% to allow for the
possibility that clinicians would not implement the consequences of the diagnosis
comprehensively and in a timely fashion in the real world.

In estimating the impact of the extension to complete active follow up/intervention to 30
months in all patients (see Figures 62 and 63 above) we took account of reduction in case
discovery and reduction in effect size. Using our established ability to model the trial, we
calculated the loss of power as a consequence of loss of case discovery and very conservative
estimates of reduction in Effect size.

**With the extension** to complete the 30 month monitoring/intervention in all randomised
patients and retaining our previous (mitigated) effect size of 40% gave us power of 79.5%.
**Without the extension**, if we made a conservative assumption of reduced effect size to 35%,
power dropped to 63%. A more likely reduction in effect size to 30% resulted in power of
only 49%.

Such detailed modelling in the trial when preliminary data became available, provided
justification for the extension of the study to complete the monitoring intervention for all
patients enrolled in the trial based on the impact on power calculations. The additional data
generated allows more robust economic analysis of the relationships between changes in the
process of care (within the 30 month follow up data) and the longer term clinical and
healthcare outcomes (ONS and HES data).

At the end of the 30 month monitoring period we were able to conduct analyses and we
report on the impact of the ELF testing on the diagnosis of cirrhosis and on the process of
care (AFP, ultrasounds, prescription of treatments, etc). We will conduct (and have funded) the 5 year follow up analysis using ONS and HES data as planned and approved.

**Modelling methodology**

The following gives details and assumptions for the simulation model upon which the sample size estimate, and ultimately the extension request, was based.

The model relies upon the data in Figure 64 from previously published data which shows the likely occurrence of severe complications from different starting ELF ranges. It is then possible to fit a (parametric) model for each of the ELF ranges, and to use this model to predict the subsequence occurrence of severe complication events. We decided to incorporate into our model a delay before events start to occur, followed by a negative exponential event-occurrence pattern, adopting a conservative approach. So if $S_i(t)$ $(i=1,6)$ are the times to incidence of severe complications for each of a range of 6 presentation ELF scores (namely <8.39, 8.4-9.49, 9.5-10.49, 10.5-11.49, 11.5-12.49, ≥12.5), then

$$S_i(t) = \begin{cases} 1, & t \leq d_i \\ e^{-\alpha_i(t-d_i)}, & t > d_i \end{cases}$$

Where $\alpha_i$ $(i=1,6)$ are the exponential parameters, and $d_i$ $(i=1,6)$ are the delays before severe complications start to occur in each of the 6 groups.
For the purposes of the model patients were divided into those that present with an ELF value >9.5 and those where the ELF value is <9.5. In the ELF arm those patients with ELF values >9.5 are classified as cirrhotic based on this ELF result. We can establish the proportion of patients that present with ELF values >9.5, and in the ranges 9.5-10.5, 10.5-11.5, 11.5-12.5 and > 12.5, from the ELF arm, and apply these proportions, in the simulation, to both arms, since although we do not know these figures in the control arm, they would be expected to be the same.

In addition, it was necessary to consider the rate of development of cirrhosis, in both arms, for patients with starting ELF values <9.5. These rates will be different since, in the ELF arm, this rate is based on the measured ELF value, whereas in the control arm it is based on clinical grounds only. Since treatment is only initiated once cirrhosis has been diagnosed, any treatment effect was applied to patients diagnosed as cirrhotic.
We analysed the cumulative incidence curves for development of cirrhosis in the trial and assumed for simplicity, for patients who were not already defined as cirrhotic by their starting ELF value, that there was a uniform rate of development of cirrhosis, say \( \lambda \), which fitted reasonably well with the observed data, as shown in Figure 62. This would imply that the development of cirrhosis for the population not already defined as cirrhotic by their starting ELF value follows an exponential distribution, specifically

\[
P_c(t) = 1 - e^{-\lambda t}
\]

Where \( P_c(t) \) is the probability of the patient becoming cirrhotic by time \( t \), and with different rates of development of cirrhosis in the ELF and control arms, since the ELF arm rate is dependent on the ELF level crossing the 9.5 threshold, whereas it is defined by clinical judgement in the control arm.

Refinements were considered, but the simulation results do not appear to be particularly sensitive to this assumption. Note that Figure 62 shows the time to development of cirrhosis for both arms combined (either defined by ELF in the ELF arm or defined by clinical judgement in the standard of care arm). These results, broken down by treatment arm, with associated model fits, were presented, in confidence, to the DMEC. Figure 63 gave a likely projection of Figure 62 without the trial extension, depicting graphically the cohort of patients who would not receive the full trial intervention (active monitoring for 30 months) in the event that the extension had not been granted.

Further refinements of the simulation model were possible looking at different models both for the development of cirrhosis, and for subsequent occurrence of severe complications. For instance, in the latter case, we may consider models where the incidence of severe complications plateaus after a particular period of time, for example (using previous terminology):

\[
S_i(t) = \begin{cases} 
1, & t \leq d_i \\
(p_i + (1 - p_i) e^{-\alpha_i(t-d_i)}), & t > d_i
\end{cases}
\]

Where \( p_i \) is the proportion of patients in the \( i^{th} \) group that never experience a severe complication.
To complete the model, we applied an assumed treatment effect, which applies a proportional hazards improvement in time to development of severe complications for patients diagnosed as cirrhotic (in either arm). This simply changed the negative exponential parameter for the time to occurrence of severe complications in each of the different ELF ranges (9.5-10.5, 10.5-11.5, 11.5-12.5 and > 12.5). Again, using previous nomenclature, these times to occurrence of severe complications are therefore assumed to be:

\[
S_i(t) = \begin{cases} 
1, & t \leq d_i \\
 e^{-\theta \alpha_i(t - d_i)}, & t > d_i 
\end{cases}
\]

Where \( \theta \) is the assumed hazard ratio. We assumed an effect size of 40%, equivalent to a hazard ratio of 0.6, for the trial design/sample size calculations for the primary outcome. We can readily estimate possible different effect sizes using this model.

Putting together all these assumptions it is possible to simulate the trial. In the ELF arm patients are assumed to present with proportions in the different ELF ranges, and, in addition, for those with an ELF value <9.5, they are assumed to have a rate of subsequent development of cirrhosis, with such patients developing cirrhosis at different ELF values, to match what we have seen in the trial to date. They then have a probability of getting a severe complication based on the negative exponential distributions described. The situation is similar in the control arm except that treatment is only initiated somewhat later after the clinically-based diagnosis of cirrhosis. For each simulated trial the times to severe events are generated, given the actual recruitment times and a specified duration of follow-up (e.g. 5 years), and then compared between the two arms using the log-rank test. See Figure 65 as a typical example, and Figure 66 to see the theoretical, expected, curves that result from these model assumptions (in this case with an effect size of 40%).

The trial power was calculated, for different Type I error rates, by simply counting the proportion of simulations where the resulting p-value is less than the assumed type I error.

Note that it was only practicable to apply such a model when the trial was underway for a reasonable period of time, since we needed estimates of the rate of development of cirrhosis in the ELF and control arms, as well as the ELF ranges for the proportions of patients who do develop cirrhosis, at the point at which they develop cirrhosis. Using this type of simulation
modelling approach, it would have been possible to apply this model prior to the start of the trial, but these rate figures would have had to be estimated based on very little data. This should, for future trials, form part of the trial design sample size assumptions. It would, perhaps, have been possible to estimate these parameters from pilot data which looked at sequential ELF values for a cohort of patients. For future designs, it would be helpful to collect such pilot data, which would be likely to be of fundamental importance to the trial design.

None of these estimated rate values involved looking at unblinded primary event data.

Conclusions

These arguments and modelling specifically allowed us to demonstrate the opportunity to complete the ELUCIDATE trial robustly and with good power for the principal clinical and healthcare outcomes. Our approach has the potential to be generalisable to a variety of other monitoring trials, and provides a basic framework for trial design in this area, which has been, to date, lacking. For instance, the model can be applied to situations where patients are all initially below the threshold for intervention, and only pass this threshold as the study proceeds. We hope to be able to provide appropriate software so that these methods can be made widely available. This is in line with our remit in this programme grant to be using this trial as an exemplar, to aid in the design, analysis and interpretation of future monitoring trials. Note that the simulation approach described herein incorporates this variability in sequential ELF scores as they change over time through the models and parameters that relate to the rate of development of cirrhosis, since this is highly correlated with changing ELF scores.

Figure 65
Simulation of the development of severe liver complications in the trial by randomised arms
SAMPLE OUTCOME CURVES FROM A TYPICAL SIMULATION WITH FOLLOW-UP CONTINUING UNTIL DEC 2019

\[ \chi^2 = 7.78 \]

\[ P = .0053 \]
Figure 66

Simulated trial outcomes for each trial arm

THEORETICAL AVERAGE OUTCOME CURVES FROM SIMULATIONS WITH FOLLOW-UP CONTINUING UNTIL DEC 2019

% WITHOUT SEVERE COMPLICATIONS

TIME (YEARS)

ELF
CTRL
Chapter 20 - Recruitment and Delivery of the ELUCIDATE Trial
Trial Organisation

Registration and Randomisation

Eligible and consenting patients were registered and into the study via an automated 24 hour telephone registration and randomisation system. Registered patients who met the ELF threshold criteria for randomisation were randomised, preferably within 6 weeks of the registration visit.

Clinical Queries

The Co-Chief Investigators based in UCLH/Royal Free (Professor William Rosenberg and Dr Sudeep Tanwar) were the contacts for clinical queries and the review of any related and unexpected serious adverse events (RUSAEs).

Project and Trial Management

The Clinical Trials Research Unit, Leeds Institute of Clinical Trials Research at the University of Leeds was responsible for the overall project and data management of the study. The Trial Management Group (TMG), comprising the Chief Investigators, health economists, scientific advisors, trial co-ordinators and statistician, were responsible for the on-going management and promotion of the study, and for the interpretation of results.

Trial Steering Committee

The Trial Steering Committee provides overall supervision of the trial, in particular trial progress, adherence to study protocol, patient safety and consideration of new information. The committee includes an independent chair (Professor James Neuberger), an independent clinician (Dr Jonathan Fallowfield), an independent scientific adviser (Dr Christine Patch), a statistician (Dr Andrew Roddam), a health economist (Dr Simon Dixon), the Chief Investigator (Professor William Rosenberg) and members of the Trial Management Group. The patient and public involvement (PPI) representatives are Tilly Hale and Joan Bedlington.
Leeds NIHR Biomarker Bank

Patients who were recruited to the trial were asked as part of the Patient Information Sheet if they would be happy to provide an additional blood sample for the Leeds NIHR Biomarker Bank at their randomisation visit. For all patients who consented, a serum sample was obtained, processed, and stored at site within a -70°C to -80°C freezer. Upon a site reaching 50 patients or the end of trial these samples were collected for storage at the Leeds NIHR Biomarker Bank, for use in future research projects.

ELF testing and the provision of ELF and ELF/NIHR Biomarker Bank kits

Both the ELF/NIHR bioRTB kits used at randomisation and the ELF kits to be used for the tests performed at patient registration, follow up (for patients on the ELF arm), and once patients have been diagnosed as cirrhotic (patients on the standard care arm) were prepared and dispatched by the Clinical and Biomedical Proteomics Group at St James’s University Hospital, Leeds. The serum samples sent for ELF testing were processed at the iQUR laboratory based at Royal Free Hospital, London. Results from these tests were disseminated to the CTRU. The CTRU then notified sites of a patient’s eligibility and current cirrhotic status, the latter pertaining only to those patients on the ELF screening arm.

Accrual

Thirty seven centres across the UK were opened to recruitment between 23/09/2010 and 31/10/2012 (Table 103). In total, 1303 participants were registered into the trial, of whom 878 were subsequently randomised (Figure 67). 440 patients were randomised to the standard clinical monitoring arm, and 438 patients were randomised to the standard clinical monitoring plus ELF test arm. Monthly and cumulative registrations and randomisations across the Centres is shown in Table 103 and Table 104. In each case it can be seen that the additional centres added late in the trial process added very substantial numbers of patients, frequently surpassing the achievements of centres open throughout the trial. Of the 425 registered patients who did not proceed to randomisation, the main reason was because their baseline ELF score was less than the threshold required for randomisation (Table 105).

Table 103 Registrations per site per month
<table>
<thead>
<tr>
<th>ELUCIDATE REGISTRATIONS</th>
<th>Date opened</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>Accrual per centre</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Jan</td>
<td>Feb</td>
<td>Mar</td>
<td>Apr</td>
</tr>
<tr>
<td>H1 23/09/10</td>
<td></td>
<td>126</td>
<td>8</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>H2 18/05/11</td>
<td></td>
<td>104</td>
<td>14</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>H3 15/06/11</td>
<td></td>
<td>39</td>
<td>1</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>H4 24/06/11</td>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>H5 13/07/11</td>
<td></td>
<td>14</td>
<td>23</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>H6 26/07/11</td>
<td></td>
<td>4</td>
<td>6</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>H7 26/07/11</td>
<td></td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>H8 04/08/11</td>
<td></td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>H9 14/09/11</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H10 06/09/11</td>
<td></td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>H11 17/10/11</td>
<td></td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>H12 05/01/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H13 05/01/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>H14 05/01/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>H15 06/03/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H16 06/03/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H17 27/03/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>H18 03/04/12</td>
<td></td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>H19 19/04/12</td>
<td></td>
<td>2</td>
<td>13</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>H20 20/04/12</td>
<td></td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>H21 26/04/12</td>
<td></td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>H22 24/05/12</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>H23 14/05/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H24 28/05/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H25 07/06/12</td>
<td></td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>H26 03/07/12</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H27 30/07/12</td>
<td></td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>H28 21/06/12</td>
<td></td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>H29 25/07/12</td>
<td></td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H30 25/07/12</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H31 30/07/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H32 01/08/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H33 21/08/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>H34 26/09/12</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H35 10/10/12</td>
<td></td>
<td>7</td>
<td>12</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>H36 18/10/12</td>
<td></td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>H37 31/10/12</td>
<td></td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Monthly Accrual</td>
<td></td>
<td>67</td>
<td>62</td>
<td>66</td>
<td>52</td>
</tr>
<tr>
<td>Cumulative Accrual</td>
<td></td>
<td>300</td>
<td>367</td>
<td>429</td>
<td>495</td>
</tr>
<tr>
<td>GLUCIDATE RANDOMISATIONS</td>
<td>2011</td>
<td>2012</td>
<td>2013</td>
<td>Accrual per centre</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-------------------</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>2012</td>
<td>2013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1  62</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>H2  44</td>
<td>5</td>
<td>7</td>
<td>15</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>H3  19</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>H4  4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>H5  0</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>H6  1</td>
<td>2</td>
<td>10</td>
<td>12</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>H7  0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H8  1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>H9  1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H10 1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>H11 0</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>H12 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H13 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H14 0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H15 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H16 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>H17 0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>H18 0</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>H19 0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H20 0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>H21 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>H22 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H23 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H24 0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H25 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H26 0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>H27 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H28 0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H29 0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H30 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H31 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H32 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>H33 0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H34 0</td>
<td>0</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>H35 0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H36 0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H37 0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 104 Randomisations per site
|                    | 2011 |        |        |        |        |        |        |        |        |        |        |        |        |        |        | Accrual per centre |
|-------------------|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|------------------|
|                    | Jan  | Feb    | Mar    | Apr    | May    | Jun    | Jul    | Aug    | Sep    | Oct    | Nov    | Dec    | Jan    | Feb    | Mar    | Apr   |
| Monthly Accrual   | 22   | 38     | 46     | 33     | 51     | 55     | 31     | 46     | 40     | 32     | 61     | 39     | 56     | 70     | 69     | 56    |
| Cumulative Accrual| 133  | 155    | 193    | 239    | 272    | 323    | 378    | 409    | 455    | 495    | 527    | 588    | 627    | 683    | 753   | 822   | 878   | 878   |
Figure 67 Monthly and cumulative accrual

Table 105 Reasons why registered patients were not randomised

<table>
<thead>
<tr>
<th>Reason for non-randomisation of registered patients</th>
<th>Total (n=425)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient’s baseline ELF score was less than the threshold for eligibility</td>
<td>317 (74.6%)</td>
</tr>
<tr>
<td>Patient can no longer attend appointments</td>
<td>14 (3.3%)</td>
</tr>
<tr>
<td>Patient developed cirrhosis between registration and randomisation</td>
<td>2 (0.5%)</td>
</tr>
<tr>
<td>Other</td>
<td>58 (13.6%)</td>
</tr>
<tr>
<td>Missing</td>
<td>34 (8.0%)</td>
</tr>
</tbody>
</table>
Chapter 21 - Preliminary Analysis of the ELUCIDATE Trial
In this Chapter, we present the analysis of the ELUCIDATE RCT to date agreed by the DMEC. These include patient characteristics, ELF scores, compliance, time to cirrhosis-associated ELF scores and changes in the process of care.

**Baseline Characteristics**

The computerised minimisation in ELUCIDATE have worked as expected and the treatment arms are balanced with regard to all baseline characteristics (Tables 106 & 107).

### Table 106 Patient characteristics at randomisation

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (categorised)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 to 39</td>
<td>47 (10.7%)</td>
<td>51 (11.6%)</td>
<td>98 (11.2%)</td>
</tr>
<tr>
<td>40 to 64</td>
<td>306 (69.9%)</td>
<td>304 (69.1%)</td>
<td>610 (69.5%)</td>
</tr>
<tr>
<td>65 to 75</td>
<td>85 (19.4%)</td>
<td>85 (19.3%)</td>
<td>170 (19.4%)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>54.0 (11.64)</td>
<td>53.8 (11.19)</td>
<td>53.9 (11.41)</td>
</tr>
<tr>
<td>Median (Range)</td>
<td>55.0 (23.0, 74.0)</td>
<td>54.0 (22.0, 74.0)</td>
<td>55.0 (22.0, 74.0)</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>246 (56.2%)</td>
<td>246 (55.9%)</td>
<td>492 (56.0%)</td>
</tr>
<tr>
<td>Female</td>
<td>192 (43.8%)</td>
<td>194 (44.1%)</td>
<td>386 (44.0%)</td>
</tr>
<tr>
<td><strong>ELF value at registration (categorised)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.4 to 9.49</td>
<td>240 (54.8%)</td>
<td>242 (55.0%)</td>
<td>482 (54.9%)</td>
</tr>
<tr>
<td>9.5 to 11.49</td>
<td>174 (39.7%)</td>
<td>172 (39.1%)</td>
<td>346 (39.4%)</td>
</tr>
<tr>
<td>11.5 to 12.49</td>
<td>14 (3.2%)</td>
<td>15 (3.4%)</td>
<td>29 (3.3%)</td>
</tr>
<tr>
<td>12.5 or higher</td>
<td>10 (2.3%)</td>
<td>11 (2.5%)</td>
<td>21 (2.4%)</td>
</tr>
<tr>
<td><strong>Registration ELF value</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>9.6 (1.07)</td>
<td>9.7 (1.03)</td>
<td>9.6 (1.05)</td>
</tr>
<tr>
<td></td>
<td>ELF (n=438)</td>
<td>Standard (n=440)</td>
<td>Total (n=878)</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------</td>
<td>-----------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Median (Range)</td>
<td>9.4 (8.4, 17.4)</td>
<td>9.4 (8.4, 14.2)</td>
<td>9.4 (8.4, 17.4)</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**History of alcohol consumption**

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>101 (23.1%)</td>
<td>107 (24.3%)</td>
<td>208 (23.7%)</td>
</tr>
<tr>
<td>No</td>
<td>337 (76.9%)</td>
<td>333 (75.7%)</td>
<td>670 (76.3%)</td>
</tr>
</tbody>
</table>

**Current alcohol consumption**

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teetotal</td>
<td>230 (52.5%)</td>
<td>231 (52.5%)</td>
<td>461 (52.5%)</td>
</tr>
<tr>
<td>Light</td>
<td>153 (34.9%)</td>
<td>151 (34.3%)</td>
<td>304 (34.6%)</td>
</tr>
<tr>
<td>Moderate</td>
<td>41 (9.4%)</td>
<td>41 (9.3%)</td>
<td>82 (9.3%)</td>
</tr>
<tr>
<td>Heavy</td>
<td>14 (3.2%)</td>
<td>17 (3.9%)</td>
<td>31 (3.5%)</td>
</tr>
</tbody>
</table>

**Primary diagnosis**

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcoholic liver disease</td>
<td>35 (8.0%)</td>
<td>33 (7.5%)</td>
<td>68 (7.7%)</td>
</tr>
<tr>
<td>Viral liver disease</td>
<td>176 (40.2%)</td>
<td>177 (40.2%)</td>
<td>353 (40.2%)</td>
</tr>
<tr>
<td>Other/unknown</td>
<td>114 (26.0%)</td>
<td>117 (26.6%)</td>
<td>231 (26.3%)</td>
</tr>
<tr>
<td>Non-alcoholic fatty liver disease</td>
<td>113 (25.8%)</td>
<td>113 (25.7%)</td>
<td>226 (25.7%)</td>
</tr>
</tbody>
</table>

**Time from registration to randomisation (in days)**

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>43.5 (32.42)</td>
<td>42.2 (34.03)</td>
<td>42.9 (33.23)</td>
</tr>
<tr>
<td>Median (Range)</td>
<td>35.0 (7.0, 349.0)</td>
<td>35.0 (6.0, 328.0)</td>
<td>35.0 (6.0, 349.0)</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*3 patients were randomised before the 22/03/2011 were done so on the old baseline ELF score categories*

*13 participants were already registered with ELF results that meant they were ineligible before the protocol amendment. After the approval of v5.0 of the protocol, these participants were re-approached and randomised into the ELUCIDATE trial.*
Table 107 Number of patients diagnosed with Viral Liver disease at randomisation (overall)

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient diagnosed with Viral Liver disease at randomisation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>176 (40.2%)</td>
<td>177 (40.2%)</td>
<td>353 (40.2%)</td>
</tr>
<tr>
<td>No</td>
<td>262 (59.8%)</td>
<td>263 (59.8%)</td>
<td>525 (59.8%)</td>
</tr>
</tbody>
</table>

As expected, the distribution of patients with and without viral disease varied across the sites (Table 108). Sites in areas where the patient population consists of a high proportion of patients with viral liver disease specialised in treating this group of patients, notably Royal Free Hospital, University College London, Kings College Hospital, Royal Blackburn Hospital, Victoria Hospital Blackpool and Torbay General Hospital.

Table 108 Number of patients diagnosed with Viral Liver disease at randomisation (by site)

<table>
<thead>
<tr>
<th>Site</th>
<th>Viral Liver disease</th>
<th>No Viral Liver disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (%)</td>
<td>N (%)</td>
<td>N (%)</td>
</tr>
<tr>
<td>Royal Free</td>
<td>81 (59.1%)</td>
<td>56 (40.9%)</td>
<td>137 (100%)</td>
</tr>
<tr>
<td>UCLH</td>
<td>85 (89.5%)</td>
<td>10 (10.5%)</td>
<td>95 (100%)</td>
</tr>
<tr>
<td>Royal Liverpool</td>
<td>5 (10.2%)</td>
<td>44 (89.8%)</td>
<td>49 (100%)</td>
</tr>
<tr>
<td>Royal Devon and Exeter Hospitals</td>
<td>14 (29.8%)</td>
<td>33 (70.2%)</td>
<td>47 (100%)</td>
</tr>
<tr>
<td>Royal Blackburn Hospital</td>
<td>43 (100.0%)</td>
<td>0 (0.0%)</td>
<td>43 (100%)</td>
</tr>
<tr>
<td>St James University Hospital</td>
<td>11 (26.8%)</td>
<td>30 (73.2%)</td>
<td>41 (100%)</td>
</tr>
<tr>
<td>Bradford Royal Infirmary</td>
<td>2 (4.9%)</td>
<td>39 (95.1%)</td>
<td>41 (100%)</td>
</tr>
<tr>
<td>Singleton Hospital</td>
<td>22 (57.9%)</td>
<td>16 (42.1%)</td>
<td>38 (100%)</td>
</tr>
<tr>
<td>Royal Bournemouth Hospital</td>
<td>6 (18.2%)</td>
<td>27 (81.8%)</td>
<td>33 (100%)</td>
</tr>
<tr>
<td>Queen Alexandra Hospital</td>
<td>4 (13.3%)</td>
<td>26 (86.7%)</td>
<td>30 (100%)</td>
</tr>
<tr>
<td>Basingstoke and North Hampshire Hospital</td>
<td>0 (0.0%)</td>
<td>29 (100.0%)</td>
<td>29 (100%)</td>
</tr>
<tr>
<td>Southampton General Hospital</td>
<td>2 (7.7%)</td>
<td>24 (92.3%)</td>
<td>26 (100%)</td>
</tr>
<tr>
<td>Royal London Hospital</td>
<td>11 (42.3%)</td>
<td>15 (57.7%)</td>
<td>26 (100%)</td>
</tr>
<tr>
<td>James Cook University Hospital</td>
<td>0 (0.0%)</td>
<td>21 (100.0%)</td>
<td>21 (100%)</td>
</tr>
<tr>
<td>Nottingham Queens Medical Centre</td>
<td>7 (33.3%)</td>
<td>14 (66.7%)</td>
<td>21 (100%)</td>
</tr>
<tr>
<td>University Hospital Lewisham</td>
<td>7 (33.3%)</td>
<td>14 (66.7%)</td>
<td>21 (100%)</td>
</tr>
<tr>
<td></td>
<td>Viral Liver disease</td>
<td>No Viral Liver disease</td>
<td>Total</td>
</tr>
<tr>
<td>------------------------------</td>
<td>---------------------</td>
<td>------------------------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>N (%)</td>
<td>N (%)</td>
<td>N (%)</td>
</tr>
<tr>
<td>Kings College Hospital</td>
<td>20 (100.0%)</td>
<td>0 (0.0%)</td>
<td>20 (100%)</td>
</tr>
<tr>
<td>Torbay District General Hospital</td>
<td>10 (58.8%)</td>
<td>7 (41.2%)</td>
<td>17 (100%)</td>
</tr>
<tr>
<td>Derriford Hospital</td>
<td>0 (0.0%)</td>
<td>16 (100.0%)</td>
<td>16 (100%)</td>
</tr>
<tr>
<td>Royal Hallamshire Hospital</td>
<td>4 (26.7%)</td>
<td>11 (73.3%)</td>
<td>15 (100%)</td>
</tr>
<tr>
<td>Princess Alexandra</td>
<td>0 (0.0%)</td>
<td>15 (100.0%)</td>
<td>15 (100%)</td>
</tr>
<tr>
<td>Royal Albert Edward Infirmary</td>
<td>0 (0.0%)</td>
<td>14 (100.0%)</td>
<td>14 (100%)</td>
</tr>
<tr>
<td>Chelsea and Westminster</td>
<td>2 (15.4%)</td>
<td>11 (84.6%)</td>
<td>13 (100%)</td>
</tr>
<tr>
<td>Totals for other sites* which randomised &lt;10 patients/site</td>
<td>17 (24.3%)</td>
<td>53 (75.7%)</td>
<td>70 (100%)</td>
</tr>
</tbody>
</table>

*Hull Royal Infirmary, Freeman Hospital, Royal Sussex County Hospital, University Hospital Durham, Kingston Hospital, Victoria Hospital Blackpool, Royal Preston Hospital, Rotherham District General Hospital, Bronglais General Hospital, Royal Hampshire County Hospital, University Hospital Bristol, Queen Elizabeth Hospital (Birmingham), Royal Lancaster Infirmary, Warrington Hospital

**Protocol Violators**

Protocol violators in ELUCIDATE include patients who scored lower than 8.4 in any post-randomisation ELF test and patients who were randomised later than 12 weeks after registration.

**ELF test below 8.4 at follow-up**

55 (25.5%) out of 216 patients with at least one reported ELF test after randomisation in the ELF arm and 1 (7.1%) out of 14 in the standard arm had an ELF score lower than 8.4 (Table 109). The proportion of patients that have ELF scores lower than this threshold value was largest (28.6%, 28/98) in patients with viral liver disease as primary diagnosis (Table 110). Moreover, within this group of patients, the proportion of patients with ELF scores below 8.4 was largest in those patients treated for hepatitis B or hepatitis C: 40.0% (8/20) and 42.3% (11/26) as compared to 17.3% (9/52) in patients that have not received any treatment since randomisation (Table 111). Response to treatment might therefore play a role. This is also possibly the case for patients with non-alcoholic fatty liver disease: here, weight loss might be the reason for a decrease in the ELF score. However, we have not collected data on weight loss to explore this hypothesis.
Table 109 Patients with ELF scores <8.4 at follow-up

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=216)</th>
<th>Standard (n=14)</th>
<th>Total (n=230)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELF score below 8.4 at follow-up</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>55 (25.5%)</td>
<td>1 (7.1%)</td>
<td>56 (24.3%)</td>
</tr>
<tr>
<td>No</td>
<td>160 (74.1%)</td>
<td>13 (92.9%)</td>
<td>173 (75.2%)</td>
</tr>
<tr>
<td>Missing</td>
<td>1 (0.5%)</td>
<td>0 (0.0%)</td>
<td>1 (0.4%)</td>
</tr>
</tbody>
</table>

Denominator: all patients with at least one reported ELF test after randomisation

Table 110 Protocol violators (too low ELF score at follow-up) by primary diagnosis

<table>
<thead>
<tr>
<th></th>
<th>Alcoholic liver disease (n=16)</th>
<th>Viral liver disease (n=98)</th>
<th>Other/unknown (n=62)</th>
<th>Non-alcoholic fatty liver disease (n=54)</th>
<th>Total (n=230)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELF score below 8.4 at follow-up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1 (6.3%)</td>
<td>28 (28.6%)</td>
<td>13 (21.0%)</td>
<td>14 (25.9%)</td>
<td>56 (24.3%)</td>
</tr>
<tr>
<td>No</td>
<td>15 (93.8%)</td>
<td>70 (71.4%)</td>
<td>48 (77.4%)</td>
<td>40 (74.1%)</td>
<td>173 (75.2%)</td>
</tr>
<tr>
<td>Missing</td>
<td>0 (0.0%)</td>
<td>0 (0.0%)</td>
<td>1 (1.6%)</td>
<td>0 (0.0%)</td>
<td>1 (0.4%)</td>
</tr>
</tbody>
</table>

Table 111 Protocol violators (too low ELF score at follow-up) with primary diagnosis viral liver disease by hepatitis treatment status (since randomisation)

<table>
<thead>
<tr>
<th></th>
<th>Hep. B treatment (n=20)</th>
<th>Hep. C treatment (n=26)</th>
<th>No Hep. treatment (n=52)</th>
<th>Total (n=98)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELF score below 8.4 at follow-up</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>8 (40.0%)</td>
<td>11 (42.3%)</td>
<td>9 (17.3%)</td>
<td>28 (28.6%)</td>
</tr>
<tr>
<td>No</td>
<td>12 (60.0%)</td>
<td>15 (57.7%)</td>
<td>43 (82.7%)</td>
<td>70 (71.4%)</td>
</tr>
</tbody>
</table>

More than 12 weeks between registration and randomisation
44 (10.0%) out of 438 randomised patients in the ELF arm and 26 (5.9%) out of 440 randomised patients in the Standard arm were not randomised within 12 weeks of registration (Table 112). For some of these patients, the ELF test was repeated before randomisation: only 15 (3.4%) out of 438 randomised patients in the ELF arm and 8 (1.8%) out of 440 randomised patients in the standard arm were not randomised within 12 weeks of the (repeated) registration ELF test.

Table 112 Patients who were randomised more than 12 weeks after registration, respectively the date of the registration ELF test

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomisation took place within 12 weeks of registration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>394 (90.0%)</td>
<td>414 (94.1%)</td>
<td>808 (92.0%)</td>
</tr>
<tr>
<td>No</td>
<td>44 (10.0%)</td>
<td>26 (5.9%)</td>
<td>70 (8.0%)</td>
</tr>
<tr>
<td>Randomisation took place within 12 weeks of registration ELF test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>421 (96.1%)</td>
<td>432 (98.2%)</td>
<td>853 (97.2%)</td>
</tr>
<tr>
<td>No</td>
<td>15 (3.4%)</td>
<td>8 (1.8%)</td>
<td>23 (2.6%)</td>
</tr>
<tr>
<td>Missing</td>
<td>2 (0.5%)</td>
<td>0 (0.0%)</td>
<td>2 (0.2%)</td>
</tr>
</tbody>
</table>
Visit Compliance

Pre- and post-cirrhotic follow-up visit compliance in ELUCIDATE is generally poor in both ELF and standard arm.

Compliance with pre-cirrhotic follow-up visits

According to the protocol, patients in both the ELF and the standard of care arm should be seen every six months. If compliance is defined as being compliant with all visits, allowing visits to take place between five and seven months after the previous visit, 67 (29.8%) out of 225 patients with expected pre-cirrhotic follow-up visits in the ELF arm and 81 (19.7%) out of 411 in the Standard arm are compliant (Table 113). The denominator in the ELF arm is substantially smaller than in the standard arm because many patients in the ELF arm were diagnosed with cirrhosis at randomisation and thus do not have any pre-cirrhotic follow-up visits.

| Table 113 Overall pre-cirrhotic follow-up compliance for patients who were expected to have at least one visit post-randomisation |
|---|---|---|
| | ELF (n=225) | Standard (n=411) | Total (n=636) |
| Compliant | 67 (29.8%) | 81 (19.7%) | 148 (23.3%) |
| Non-compliant | 158 (70.2%) | 330 (80.3%) | 488 (76.7%) |

Patients are defined as compliant if they have been compliant for all visits so far.

Timing of pre-cirrhotic follow-up visits

A total of 597 pre-cirrhotic follow-up visits took place in the ELF arm with on average 7.4 (SD 3.42) months between visits. In the standard arm, there were 1424 reported pre-cirrhotic follow-up visits with on average 7.0 (SD 2.89) months between visits (Table 114 and Figure 68).
Table 114 Time between pre-cirrhotic follow-up visits

<table>
<thead>
<tr>
<th>Duration between follow up visits (months)</th>
<th>ELF (n=597)</th>
<th>Standard (n=1424)</th>
<th>Total (n=2021)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>7.4 (3.42)</td>
<td>7.0 (2.89)</td>
<td>7.1 (3.06)</td>
</tr>
<tr>
<td>Median (Range)</td>
<td>6.2 (1.8, 33.2)</td>
<td>6.2 (1.0, 31.0)</td>
<td>6.2 (1.0, 33.2)</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 68 Time between pre-cirrhotic follow-up visits
Compliance with post-cirrhotic follow-up visits

According to the study protocol, all patients diagnosed as cirrhotic either by ELF or clinical means should attend an initial post-cirrhotic follow up visit, at three months after the diagnosis of cirrhosis. All subsequent follow-up visits for the purposes of data collection should take place every six months. If compliance is defined as being compliant with all visits, 43 (16.2%) out of 266 patients with expected post-cirrhotic follow-up visits in the ELF arm and 4 (21.1%) out of 19 in the Standard arm are compliant (Table 115). The denominator in the standard arm is substantially smaller than in the ELF arm because only a small number of patients in the standard arm were diagnosed with cirrhosis.

Table 115 Overall compliance for patients who were expected to have at least one visit post diagnosis of cirrhosis (by arm)

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=266)</th>
<th>Standard (n=19)</th>
<th>Total (n=285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compliant</td>
<td>43 (16.2%)</td>
<td>4 (21.1%)</td>
<td>47 (16.5%)</td>
</tr>
<tr>
<td>Non-compliant</td>
<td>223 (83.8%)</td>
<td>15 (78.9%)</td>
<td>238 (83.5%)</td>
</tr>
</tbody>
</table>

Patients are defined as compliant if they have been compliant for all visits so far.

Timing of post-cirrhotic follow-up visits

Looking only at post-cirrhotic visits that were expected to take place 6 months after the previous visit, the time between visits was on average 6.5 (SD 2.80) months in the ELF arm (n=695). In the standard arm (n=35), the time between visits was on average 7.2 (SD 4.71) (Table 116 and Figure 69).

Table 116 Time between post-cirrhotic follow-up visits (for visits that are expected to take place six months after the previous visit)

<table>
<thead>
<tr>
<th>Duration between follow up visits (months)</th>
<th>ELF (n=695)</th>
<th>Standard (n=35)</th>
<th>Total (n=729)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>6.5 (2.80)</td>
<td>7.2 (4.71)</td>
<td>6.5 (2.92)</td>
</tr>
<tr>
<td>Median (Range)</td>
<td>6.1 (0.7, 23.4)</td>
<td>5.9 (2.3, 25.2)</td>
<td>6.1 (0.7, 25.2)</td>
</tr>
<tr>
<td>(Q1, Q3)</td>
<td>(5.2, 7.2)</td>
<td>(4.4, 7.4)</td>
<td>(5.2, 7.2)</td>
</tr>
</tbody>
</table>
Figure 69 Time between post-cirrhotic follow-up visits (for visits that are expected to take place six months after the previous visit)

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=694)</th>
<th>Standard (n=35)</th>
<th>Total (n=729)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**ELF Test Compliance**

Patients in the ELF arm should have an ELF test at every pre-cirrhosis follow up visit. Patients in the standard arm should have one ELF test after diagnosis of cirrhosis.

**ELF Test compliance in the ELF arm**

Compliance to ELF tests in the ELF arm is 72%, with compliance being defined as having an ELF test at every recorded pre-cirrhosis follow-up visit (Table 117). Note that the denominator only includes ELF arm patients that were expected to have at least one pre-cirrhotic follow-up visit and for instance does not include those patients that were diagnosed with cirrhosis at randomisation.
Most ELF tests were not done because of administrative errors (31) or lack of staff (27) and most cases of non-compliance occurred at a small number of sites.

**Table 117 ELF arm - Overall ELF test compliance**

<table>
<thead>
<tr>
<th></th>
<th>Total (n=225)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compliant</td>
<td>162 (72.0%)</td>
</tr>
<tr>
<td>Non-compliant</td>
<td>63 (28.0%)</td>
</tr>
</tbody>
</table>

**ELF Test compliance in the standard of care arm**

According to protocol, patients in the standard arm should have one ELF after the confirmed clinical diagnosis of cirrhosis. 12 out of 20 patients (60%) in the standard arm that were expected to be tested were compliant with ELF test, for one patient there was not enough information to assess compliance (Table 118).

**Table 118 Standard arm - ELF test compliance**

<table>
<thead>
<tr>
<th></th>
<th>Total (n=20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compliant</td>
<td>12 (60.0%)</td>
</tr>
<tr>
<td>Non-compliant</td>
<td>7 (35.0%)</td>
</tr>
<tr>
<td>Insufficient information</td>
<td>1 (5.0%)</td>
</tr>
</tbody>
</table>
**Compliance with cirrhosis management**

Compliance with timing of Alpha-Fetoprotein (AFP) tests, scans (USS, CT or MRI) and Endoscopies (OGD) was assessed for patients who were diagnosed with cirrhosis.

**Alpha-Fetoprotein (AFP) tests**

**AFP compliance**

According to the protocol, the timing of initial AFP tests following a diagnosis of cirrhosis depends on whether there was an AFP test done in the 6 months prior to the diagnosis: If the patient had an AFP test within 6 months prior to being diagnosed as cirrhotic, the next AFP measurement should be performed within 6 months of the previous test unless clinically indicated sooner. If the patient did not have an AFP test within 6 months prior to being diagnosed as cirrhotic, they should undergo an AFP test within 3 months of diagnosis of cirrhosis. After the initial AFP test, it should be repeated every 6 months from the previous test. AFP test compliance was low in both ELF and standard arm with 12.6% and 11.8%, respectively (Table 119). Compliance is defined as being compliant with all visits so far and one month tolerance is applied.

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=231)</th>
<th>Standard (n=17)</th>
<th>Total (n=248)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compliant</td>
<td>29 (12.6%)</td>
<td>2 (11.8%)</td>
<td>31 (12.5%)</td>
</tr>
<tr>
<td>Non-compliant</td>
<td>202 (87.4%)</td>
<td>15 (88.2%)</td>
<td>217 (87.5%)</td>
</tr>
</tbody>
</table>

**AFP timings**

For all AFP measurements that were due to take place 6 months after the previous one, median time between measurements was 6.0 (Range: 1.8, 26.8) in the ELF arm and 6.2 (3.0, 25.3) in the standard arm (Table 120 and Figure 70).
Table 120 Overall timings (for all AFP measurements due to take place 6 months after the previous one)

<table>
<thead>
<tr>
<th>Time (months)</th>
<th>ELF (n=465)</th>
<th>Standard (n=22)</th>
<th>Total (n=487)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>6.7 (3.56)</td>
<td>7.9 (5.06)</td>
<td>6.8 (3.64)</td>
</tr>
<tr>
<td>Median (Range)</td>
<td>6.0 (1.8, 26.8)</td>
<td>6.2 (3.0, 25.3)</td>
<td>6.0 (1.8, 26.8)</td>
</tr>
<tr>
<td>(Q1, Q3)</td>
<td>(4.4, 7.4)</td>
<td>(5.1, 8.5)</td>
<td>(4.4, 7.6)</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 70 Overall AFP timings

Scans (USS, CT or MRI)

If a cirrhotic patient underwent a scan within 6 months prior to being diagnosed as cirrhotic, the next scan should be performed within 6 months of the previous scan unless clinically indicated sooner. If the patient did not have a scan within 6 months prior to being diagnosed
as cirrhotic, they should undergo a scan as screening for HCC within 3 months of diagnosis of cirrhosis. Subsequent scans should be repeated every 6 months from the timing of the previous scan.

**Scan compliance**

Compliance with scans was low in both ELF and standard arm with 29.5% and 14.3%, respectively (Table 121).

<table>
<thead>
<tr>
<th>Table 121 Overall scan compliance for patients who were expected to have at least one measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Compliant</td>
</tr>
<tr>
<td>Non-compliant</td>
</tr>
</tbody>
</table>

**Scan timings**

For all scans that were due to take place 6 months after the previous one, median time between measurements was 6.8 (Range: 0.0, 30.0) in the ELF arm and 7.5 (1.3, 11.7) in the standard arm (Table 122 and Figure 71).

<table>
<thead>
<tr>
<th>Table 122 Overall scan timings (for all scans due to take place 6 months after the previous one)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Time (months)</td>
</tr>
<tr>
<td>Mean (SD)</td>
</tr>
<tr>
<td>Median (Range)</td>
</tr>
<tr>
<td>(Q1, Q3)</td>
</tr>
<tr>
<td>Missing</td>
</tr>
</tbody>
</table>
Endoscopies (OGD)

According to the protocol, all patients with diagnosis of cirrhosis should have an oesophagastroduodenoscopy (OGD) as screening for varices within 3 months of diagnosis, unless they have had an OGD in the last 18 months prior to the diagnosis of cirrhosis.

Compliance with timings of endoscopies was fair in both ELF and standard arm with 54.9% and 50.0%, respectively (Table 123).
OGD compliance

Table 123 Overall compliance for patients who are expected to have at least one Endoscopy

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=153)</th>
<th>Standard (n=12)</th>
<th>Total (n=165)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compliant</td>
<td>84 (54.9%)</td>
<td>6 (50.0%)</td>
<td>90 (54.5%)</td>
</tr>
<tr>
<td>Non-compliant</td>
<td>69 (45.1%)</td>
<td>6 (50.0%)</td>
<td>75 (45.5%)</td>
</tr>
</tbody>
</table>

Withdrawals

32 patients in the ELF arm and 24 patients in the standard arm have withdrawn from different aspects of the trial. 21 patients in the ELF arm withdrew consent for further ELF testing. Of these 21 patients 10 did not want to attend or were unable to attend (with 2 specifically mentioning distance), 5 moved location, 2 did not want the ELF test – one due to difficulties with bleeding the patient and one where the patient mentioned that ELF indicated cirrhosis but a fibroscan indicated that they were not cirrhotic so they didn’t want more ELF tests, and 4 gave no reason for withdrawing consent. It can be seen that only 2 patients specifically mentioned issues with the ELF test itself. 51 patients (28 in the ELF arm and 23 in the standard arm) were no longer willing to be followed up as per the protocol schedule. 14 of these patients (9 ELF arm, 5 standard arm) were willing for further data to be collected at their standard visits, if available. 27 patients (16 ELF arm, 11 standard arm) were willing to have their long-term data collected via a patient registry.
Disease Progression to Cirrhosis

281 (64.2%) out of 438 randomised patients in the ELF arm and 20 (4.5%) out of 440 in the Standard arm were diagnosed with cirrhosis (Table 124). 197 (70.1%) ELF arm patients were diagnosed at randomisation but none of the patients in the standard arm (Tables 124, 135). Almost all ELF arm patients were diagnosed with an ELF test (99.6%) while the most frequent method of diagnosis in the standard arm were Fibroscans (45.0%), Tables 124 – 129.

Table 124 Diagnosis of Cirrhosis overall (by arm)

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis of cirrhosis during the trial</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>281 (64.2%)</td>
<td>20 (4.5%)</td>
<td>301 (34.3%)</td>
</tr>
<tr>
<td>No</td>
<td>157 (35.8%)</td>
<td>420 (95.5%)</td>
<td>577 (65.7%)</td>
</tr>
</tbody>
</table>

Table 125 Timing of diagnosis i.e. at/after randomisation (by arm)

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=281)</th>
<th>Standard (n=20)</th>
<th>Total (n=301)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis of cirrhosis after randomisation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>84 (29.9%)</td>
<td>20 (100.0%)</td>
<td>104 (34.6%)</td>
</tr>
<tr>
<td>No</td>
<td>197 (70.1%)</td>
<td>0 (0.0%)</td>
<td>197 (65.4%)</td>
</tr>
</tbody>
</table>
### Table 126 Method of diagnosis of Cirrhosis (overall)

<table>
<thead>
<tr>
<th>Method</th>
<th>ELF (n=281)</th>
<th>Standard (n=20)</th>
<th>Total (n=301)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver biopsy</td>
<td>1 (0.4%)</td>
<td>2 (10.0%)</td>
<td>3 (1.0%)</td>
</tr>
<tr>
<td>Ultrasound</td>
<td>2 (0.7%)</td>
<td>7 (35.0%)</td>
<td>9 (3.0%)</td>
</tr>
<tr>
<td>Liver CT</td>
<td>0 (0.0%)</td>
<td>3 (15.0%)</td>
<td>3 (1.0%)</td>
</tr>
<tr>
<td>MRI</td>
<td>1 (0.4%)</td>
<td>0 (0.0%)</td>
<td>1 (0.3%)</td>
</tr>
<tr>
<td>Gastroscopy</td>
<td>1 (0.4%)</td>
<td>7 (35.0%)</td>
<td>8 (2.7%)</td>
</tr>
<tr>
<td>Fibroscan</td>
<td>2 (0.7%)</td>
<td>9 (45.0%)</td>
<td>11 (3.7%)</td>
</tr>
<tr>
<td>Clinical judgement</td>
<td>2 (0.7%)</td>
<td>6 (30.0%)</td>
<td>8 (2.7%)</td>
</tr>
<tr>
<td>ELF test</td>
<td>280 (99.6%)</td>
<td>0 (0.0%)</td>
<td>280 (93.0%)</td>
</tr>
</tbody>
</table>

Note: Methods of diagnosis are not mutually exclusive

### Table 127 Method of diagnosis of Cirrhosis (for patients who were not diagnosed at randomisation)

<table>
<thead>
<tr>
<th>Method</th>
<th>ELF (n=84)</th>
<th>Standard (n=20)</th>
<th>Total (n=104)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver biopsy</td>
<td>1 (1.2%)</td>
<td>2 (10.0%)</td>
<td>3 (2.9%)</td>
</tr>
<tr>
<td>Ultrasound</td>
<td>1 (1.2%)</td>
<td>7 (35.0%)</td>
<td>8 (7.7%)</td>
</tr>
<tr>
<td>Liver CT</td>
<td>0 (0.0%)</td>
<td>3 (15.0%)</td>
<td>3 (2.9%)</td>
</tr>
<tr>
<td>Gastroscopy</td>
<td>0 (0.0%)</td>
<td>7 (35.0%)</td>
<td>7 (6.7%)</td>
</tr>
<tr>
<td>Fibroscan</td>
<td>1 (1.2%)</td>
<td>9 (45.0%)</td>
<td>10 (9.6%)</td>
</tr>
<tr>
<td>Clinical judgement</td>
<td>2 (2.4%)</td>
<td>6 (30.0%)</td>
<td>8 (7.7%)</td>
</tr>
<tr>
<td>ELF test</td>
<td>83 (98.8%)</td>
<td>0 (0.0%)</td>
<td>83 (79.8%)</td>
</tr>
</tbody>
</table>

Note: Methods of diagnosis are not mutually exclusive
Table 128 First method of diagnosis of Cirrhosis (overall)

<table>
<thead>
<tr>
<th>Method</th>
<th>ELF (n=281)</th>
<th>Standard (n=20)</th>
<th>Total (n=301)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver biopsy</td>
<td>1 (0.4%)</td>
<td>2 (10.0%)</td>
<td>3 (1.0%)</td>
</tr>
<tr>
<td>Ultrasound</td>
<td>0 (0.0%)</td>
<td>6 (30.0%)</td>
<td>6 (2.0%)</td>
</tr>
<tr>
<td>Liver CT</td>
<td>0 (0.0%)</td>
<td>2 (10.0%)</td>
<td>2 (0.7%)</td>
</tr>
<tr>
<td>Gastroscopy</td>
<td>0 (0.0%)</td>
<td>3 (15.0%)</td>
<td>3 (1.0%)</td>
</tr>
<tr>
<td>Fibroscan</td>
<td>1 (0.4%)</td>
<td>6 (30.0%)</td>
<td>7 (2.3%)</td>
</tr>
<tr>
<td>Clinical judgement</td>
<td>1 (0.4%)</td>
<td>1 (5.0%)</td>
<td>2 (0.7%)</td>
</tr>
<tr>
<td>ELF test</td>
<td>278 (98.9%)</td>
<td>0 (0.0%)</td>
<td>278 (92.4%)</td>
</tr>
</tbody>
</table>

Table 129 First method of diagnosis of Cirrhosis (for patients who were not diagnosed at randomisation)

<table>
<thead>
<tr>
<th>Method</th>
<th>ELF (n=84)</th>
<th>Standard (n=20)</th>
<th>Total (n=104)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver biopsy</td>
<td>1 (1.2%)</td>
<td>2 (10.0%)</td>
<td>3 (2.9%)</td>
</tr>
<tr>
<td>Ultrasound</td>
<td>0 (0.0%)</td>
<td>6 (30.0%)</td>
<td>6 (5.8%)</td>
</tr>
<tr>
<td>Liver CT</td>
<td>0 (0.0%)</td>
<td>2 (10.0%)</td>
<td>2 (1.9%)</td>
</tr>
<tr>
<td>Gastroscopy</td>
<td>0 (0.0%)</td>
<td>3 (15.0%)</td>
<td>3 (2.9%)</td>
</tr>
<tr>
<td>Fibroscan</td>
<td>1 (1.2%)</td>
<td>6 (30.0%)</td>
<td>7 (6.7%)</td>
</tr>
<tr>
<td>Clinical judgement</td>
<td>1 (1.2%)</td>
<td>1 (5.0%)</td>
<td>2 (1.9%)</td>
</tr>
<tr>
<td>ELF test</td>
<td>81 (96.4%)</td>
<td>0 (0.0%)</td>
<td>81 (77.9%)</td>
</tr>
</tbody>
</table>
Process of Care Outcomes

The frequency of biopsy, AFP, imaging and endoscopy and visits is summarised by treatment arm for both non-cirrhotic and cirrhotic patients. The difference in the proportion of patients undergoing/receiving each process endpoint between the treatment groups is compared using logistic regression, adjusting for the stratification factors. Where the process outcome relates to numbers of tests being performed (e.g. ultrasound, AFP) the two arms are compared using the Mann-Whitney non-parametric test.

Frequency of Biopsies

12.1% of patients randomised into the ELF arm have had at least one biopsy post randomisation as opposed to 9.1% of the patients in the standard arm (Table 130). The odds ratio is 1.405, with higher odds for patients in the ELF arm to have a biopsy than patients in the standard arm, but the 95% confidence interval around this odds ratio (0.884-2.233) includes the critical value of 1 (equal odds) and the difference between ELF and standard arm is therefore not statistically significant.

Table 130 Number of randomised patients having at least one biopsy (pre or post diagnosis of cirrhosis)

<table>
<thead>
<tr>
<th>Has the patient had at least one Biopsy during the trial</th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
<th>Lower Confidence Limit for Odds Ratio</th>
<th>Upper Confidence Limit for Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>53 (12.1%)</td>
<td>40 (9.1%)</td>
<td>93 (10.6%)</td>
<td>1.405</td>
<td>0.884</td>
</tr>
<tr>
<td>No</td>
<td>385 (87.9%)</td>
<td>400 (90.9%)</td>
<td>785 (89.4%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Odds ratio is adjusted for stratification factors
Frequency of Alpha-Fetoprotein (AFP) tests

Table 131 shows the frequency of AFP measurements post randomisation by arm. AFP measurements were more frequent in patients in the ELF arm, which can also be seen in the histogram in Figure 72 where the mass of the distribution of the number of AFP tests per randomised patient in the ELF arm is located to the right of the distribution in the standard arm. The median number of AFP measurements in the ELF arm is 2 (Range: 0-7) as compared to 1 (Range: 0-5) in the standard arm and the difference between the two distributions is statistically significant with a p-value smaller than 0.001 (Table 132). AFP measurements are a process of care in cirrhotic patients and this difference suggests that patients in the ELF arm are not only more likely to be diagnosed with cirrhosis (see section on Disease Progression), they are also more likely to receive the processes of care to monitor cirrhotic patients for major complications.

The timing of post-diagnosis AFP tests follows the same pattern in the ELF arm as in the standard arm (Figure 73). Moreover, before diagnosis of cirrhosis, the average number of AFP tests per year is very similar in both arms with 0.68 in the ELF arm and 0.56 in the standard arm (Table 133). After diagnosis of cirrhosis, this number is 0.90 in the ELF arm and 0.95 in the standard arm. This is further evidence for the causal link between diagnosis of cirrhosis and higher frequency of AFP tests, independent of treatment allocation. Whilst we have seen in the section on Disease Progression that participants in the ELF arm are more likely to be diagnosed with cirrhosis, we can see here that this diagnosis will also lead to a higher frequency of AFP in the ELF arm.
Table 131 Number of AFP tests per randomised patient by arm

<table>
<thead>
<tr>
<th>Variable</th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of AFP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>89 (20.3%)</td>
<td>130 (29.5%)</td>
<td>219 (24.9%)</td>
</tr>
<tr>
<td>1</td>
<td>66 (15.1%)</td>
<td>92 (20.9%)</td>
<td>158 (18.0%)</td>
</tr>
<tr>
<td>2</td>
<td>68 (15.5%)</td>
<td>80 (18.2%)</td>
<td>148 (16.9%)</td>
</tr>
<tr>
<td>3</td>
<td>89 (20.3%)</td>
<td>67 (15.2%)</td>
<td>156 (17.8%)</td>
</tr>
<tr>
<td>4</td>
<td>67 (15.3%)</td>
<td>50 (11.4%)</td>
<td>117 (13.3%)</td>
</tr>
<tr>
<td>5</td>
<td>37 (8.4%)</td>
<td>21 (4.8%)</td>
<td>58 (6.6%)</td>
</tr>
<tr>
<td>6</td>
<td>16 (3.7%)</td>
<td>0 (0.0%)</td>
<td>16 (1.8%)</td>
</tr>
<tr>
<td>7</td>
<td>6 (1.4%)</td>
<td>0 (0.0%)</td>
<td>6 (0.7%)</td>
</tr>
</tbody>
</table>

Figure 72 Number of AFP measurements per randomised patient by arm
Table 132 Number of AFP tests per randomised patients by arm

<table>
<thead>
<tr>
<th>Variable</th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
<th>P-value, Wilcoxon Test (Two-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of AFP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>2.4 (1.82)</td>
<td>1.7 (1.53)</td>
<td>2.1 (1.72)</td>
<td></td>
</tr>
<tr>
<td>Median (Range)</td>
<td>2.0 (0.0, 7.0)</td>
<td>1.0 (0.0, 5.0)</td>
<td>2.0 (0.0, 7.0)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Figure 73 Number of AFP measurements post-diagnosis per 6 month period

Number of AFP tests post-diagnosis per 6 month period after diagnosis
(Number of patients diagnosed with cirrhosis: 281 in the ELF arm; 20 in the Standard arm)

<table>
<thead>
<tr>
<th>Frequency</th>
<th>1 to 6</th>
<th>7 to 12</th>
<th>13 to 18</th>
<th>19 to 24</th>
<th>25 to 30</th>
<th>31 to 36</th>
<th>37 to 42</th>
</tr>
</thead>
<tbody>
<tr>
<td>months since diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Randomisation result</td>
<td>ELF</td>
<td>Standard</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% in ELF arm</td>
<td>93.5</td>
<td>93.7</td>
<td>95.4</td>
<td>95.1</td>
<td>97.3</td>
<td>97.1</td>
<td></td>
</tr>
</tbody>
</table>
Table 133 Frequency of AFP before and after diagnosis of cirrhosis

<table>
<thead>
<tr>
<th>Randomisation result</th>
<th>Total number of AFP</th>
<th>Total years of follow-up time</th>
<th>Average number of AFP per year of follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>any time</td>
<td>before diagnosis</td>
<td>after diagnosis</td>
</tr>
<tr>
<td>ELF</td>
<td>1060</td>
<td>383</td>
<td>677</td>
</tr>
<tr>
<td>Standard</td>
<td>758</td>
<td>721</td>
<td>37</td>
</tr>
</tbody>
</table>

**Frequency of Ultrasound Scans (USS)**

Table 134 shows the frequency ultrasound scans (USS) post randomisation by arm. USS were performed more frequently in patients in the ELF arm, which can also be seen in the histogram in Figure 74 where the mass of the distribution of the number of USS per randomised patient in the ELF arm is located to the right of the distribution in the standard arm. The median number of USS in the ELF arm is 1 (Range: 0-6) as compared to 0 (Range: 0-5) in the standard arm and the difference between the two distributions is statistically significant with a p-value smaller than 0.001 (Table 135).

USS are a process of care in cirrhotic patients and this difference demonstrates that patients in the ELF arm are not only more likely to be diagnosed with cirrhosis (see section on Disease Progression), they are also more likely to receive the processes of care of cirrhotic patients. This is a really important finding as only regular and frequent USS can pick up hepatocellular cancer early.

The timing of post-diagnosis USS follows the same pattern in the ELF arm as in the standard arm (Figure 75). Moreover, before diagnosis of cirrhosis, the average number of USS per year is very similar in both arms with 0.29 in the ELF arm and 0.26 in the standard arm (Table 136). After diagnosis of cirrhosis, this number is 0.66 in the ELF arm and 0.74 in the standard arm. This is further evidence for the causal link between diagnosis of cirrhosis and higher frequency of USS, independent of treatment allocation. Whilst we have seen in the section on Disease Progression that participants in the ELF arm are more likely to be diagnosed with cirrhosis, we can see here that this diagnosis will also lead to a higher frequency of USS in both trial arms. Table 136 shows an excess of 306 scans in the trial ELF arm.
Table 134 Number of Ultrasound Scans (USS) per randomised patient by arm

<table>
<thead>
<tr>
<th>Variable</th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of USS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>151 (34.5%)</td>
<td>233 (53.0%)</td>
<td>384 (43.7%)</td>
</tr>
<tr>
<td>1</td>
<td>98 (22.4%)</td>
<td>111 (25.2%)</td>
<td>209 (23.8%)</td>
</tr>
<tr>
<td>2</td>
<td>77 (17.6%)</td>
<td>55 (12.5%)</td>
<td>132 (15.0%)</td>
</tr>
<tr>
<td>3</td>
<td>53 (12.1%)</td>
<td>30 (6.8%)</td>
<td>83 (9.5%)</td>
</tr>
<tr>
<td>4</td>
<td>40 (9.1%)</td>
<td>8 (1.8%)</td>
<td>48 (5.5%)</td>
</tr>
<tr>
<td>5</td>
<td>16 (3.7%)</td>
<td>3 (0.7%)</td>
<td>19 (2.2%)</td>
</tr>
<tr>
<td>6</td>
<td>3 (0.7%)</td>
<td>0 (0.0%)</td>
<td>3 (0.3%)</td>
</tr>
</tbody>
</table>

Figure 74 Number of USS per randomised patients by arm
Table 135 Number of Ultrasound Scans (USS) per randomised patients by arm – Wilcoxon Test

<table>
<thead>
<tr>
<th>Variable</th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
<th>P-value, Wilcoxon Test (Two-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of USS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>1.5 (1.52)</td>
<td>0.8 (1.08)</td>
<td>1.2 (1.36)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Median (Range)</td>
<td>1.0 (0.0, 6.0)</td>
<td>0.0 (0.0, 5.0)</td>
<td>1.0 (0.0, 6.0)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Figure 75 Number of USS post diagnosis per 6 month period

Number of USS post-diagnosis per 6 month period after diagnosis
(Number of patients diagnosed with cirrhosis: 281 in the ELF arm; 20 in the Standard arm)

% in ELF arm: 94.7, 93.5, 93.8, 94.6, 98.0, 92.9
Table 136 Frequency of USS before and after diagnosis of cirrhosis

<table>
<thead>
<tr>
<th>Randomisation result</th>
<th>Total number of USS</th>
<th>Total years of follow-up time</th>
<th>Average number of USS per year of follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Any time before diagnosis</td>
<td>after diagnosis</td>
<td>Any time before diagnosis</td>
</tr>
<tr>
<td>ELF</td>
<td>664</td>
<td>163</td>
<td>501</td>
</tr>
<tr>
<td>Standard</td>
<td>358</td>
<td>329</td>
<td>29</td>
</tr>
</tbody>
</table>

For 5 scans (4 in ELF and 1 in the standard arm) scan dates are missing. These scans are excluded from this table.

**Frequency of Oesophago-Gastro-Duodenoscopies (OGD)**

34.9% of patients randomised into the ELF arm have had at least one endoscopy (OGD) post randomisation as compared to 2.7% of the patients in the standard arm. The (adjusted) odds ratio is 83.9, with higher odds for patients in the ELF arm to have an endoscopy than patients in the standard arm. The 95% confidence interval around this odds ratio is 36.9-192.4, therefore the difference between ELF and standard arm is clearly significant. This is likely to be because a larger proportion of patients are diagnosed with cirrhosis in the ELF arm than in the standard arm and these patients have endoscopies as part of their cirrhosis management, as expected.

When looking only at the subset of patients that were diagnosed with cirrhosis, 54.4% of the ELF arm patients have had at least one OGD post-diagnosis as compared to 60% in the standard arm. The (adjusted) odds ratio is still in favour of the ELF arm (1.3) but the 95% confidence interval (0.4-4.9) shows that this difference in odds is not significant anymore. This means the odds of having an OGD post diagnosis of cirrhosis are equivalent, suggesting that patients received the same processes of care with regard to OGDs in both arms (Tables 137 and 138). However, as a result, overall in the trial to date there were a total of 165 additional OGDs performed in the ELF arm – with only 14 OGDs performed in the standard arm compared to 179 in the ELF arm. This is probably the single largest difference in process of care tests between the arms (Table 139).
### Table 137 Number of randomised patients having at least one OGD since randomisation

<table>
<thead>
<tr>
<th>Variable</th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
<th>Odds Ratio</th>
<th>Lower Confidence Limit for Odds Ratio</th>
<th>Upper Confidence Limit for Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Has the patient had at least one OGD?</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>153 (34.9%)</td>
<td>12 (2.7%)</td>
<td>165 (18.8%)</td>
<td>83.894</td>
<td>36.573</td>
<td>192.443</td>
</tr>
<tr>
<td>No</td>
<td>285 (65.1%)</td>
<td>428 (97.3%)</td>
<td>713 (81.2%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Odds ratio is adjusted for stratification factors

### Table 138 Number of patients having at least one OGD post diagnosis of Cirrhosis

<table>
<thead>
<tr>
<th>Variable</th>
<th>ELF (n=281)</th>
<th>Standard (n=20)</th>
<th>Total (n=301)</th>
<th>Odds Ratio</th>
<th>Lower Confidence Limit for Odds Ratio</th>
<th>Upper Confidence Limit for Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Has the patient had at least one OGD since diagnosis of cirrhosis?</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>153 (54.4%)</td>
<td>12 (60.0%)</td>
<td>165 (54.8%)</td>
<td>1.304</td>
<td>0.350</td>
<td>4.861</td>
</tr>
<tr>
<td>No</td>
<td>128 (45.6%)</td>
<td>8 (40.0%)</td>
<td>136 (45.2%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Odds ratio is adjusted for stratification factors

### Table 139 Number of OGDs by randomisation result

<table>
<thead>
<tr>
<th>Randomisation result</th>
<th>ELF</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (n=193)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELF</td>
<td>179 (92.7%)</td>
<td>14 (7.3%)</td>
</tr>
</tbody>
</table>
Frequency of Beta-Blocker/Band litigation treatment

9 (2.1%) patients in the ELF arm 7 (1.6%) in the standard arm were diagnosed with varices (Table 140). 5 (1.1%) patients in the ELF arm and 3 (0.7%) were treated with beta blockers or band ligation (Table 141). The adjusted odds ratio is 1.3 (95% CI 0.1-13.8) and shows that there is no significant difference in the likelihood to be treated between ELF and standard arm patients.

Table 140 Number of patients who have developed varices (by arm)

<table>
<thead>
<tr>
<th>Has the patient developed Varices?</th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>9 (2.1%)</td>
<td>7 (1.6%)</td>
<td>16 (1.8%)</td>
</tr>
<tr>
<td>No</td>
<td>429 (97.9%)</td>
<td>433 (98.4%)</td>
<td>862 (98.2%)</td>
</tr>
</tbody>
</table>

Table 141 Number of randomised patients being treated with beta blockers or band ligation

<table>
<thead>
<tr>
<th>Variable</th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
<th>Lower Confidence Limit for Odds Ratio</th>
<th>Upper Confidence Limit for Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has the patient been treated with beta blockers or band ligation?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>5 (1.1%)</td>
<td>3 (0.7%)</td>
<td>8 (0.9%)</td>
<td>1.343</td>
<td>0.131 13.754</td>
</tr>
<tr>
<td>No</td>
<td>433 (98.9%)</td>
<td>437 (99.3%)</td>
<td>870 (99.1%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Odds ratio adjusted for stratification factors

Frequency of treatment to normalise LFTs

57 (13%) patients in the ELF arm and 48 (10.9%) in the standard arm were receiving treatment to normalise liver function tests (LFTs, Table 142). The adjusted odds ratio is 1.5
(95% CI 0.9-2.7) and suggests that there is no statistically significant difference in the odds of receiving treatment to normalise LFTs between the two arms. However, again the numbers are small, and a few more ELF arm patients are receiving treatment to normalise LFTs, which is in the right direction even though this is not significant.

Table 142 Number of randomised patients receiving treatment to normalise LFTs

<table>
<thead>
<tr>
<th>Variable</th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
<th>Odds Ratio</th>
<th>Lower Confidence Limit for Odds Ratio</th>
<th>Upper Confidence Limit for Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has the patient received treatment to normalise LFTs?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>57 (13.0%)</td>
<td>48 (10.9%)</td>
<td>105 (12.0%)</td>
<td>1.510</td>
<td>0.850</td>
<td>2.681</td>
</tr>
<tr>
<td>No</td>
<td>381 (87.0%)</td>
<td>392 (89.1%)</td>
<td>773 (88.0%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Odds ratio is adjusted for stratification factors

Related Unexpected Serious Adverse Events (RUSAEs)

There were no RUSAEs reported in the ELUCIDATE trial.

Health Economic consequences

The follow-up period for the ELUCIDATE trial has been extended to 5 years beyond the end of the NIHR Programme Grant, meaning that we are unable, at this stage, to access data on resources used and quality of life by arm. This prevented the estimates of QALY and assessment of the cost-effectiveness of the ELF test in the early detection of cirrhosis.

Considering that one of the aims of the trial is to assess how the use of ELF affects the process of care, the aim of the health economics at this stage was restricted to a descriptive analysis of the costs associated with the process of care outcomes. These are namely, increased use of endoscopy, biopsy, ultrasound, alpha-fetoprotein (AFP) tests to detect
hepatocellular cancer (HCC) at a surgically curable stage; and increased use of beta-blockers/band ligation of varices to prevent haemorrhage/HCC.
Descriptive analysis

Cost associated with process of care outcomes

For costing out the process of care we have assigned a unit cost to mean values (as reported in the statistical analysis) for the various procedures. Unit costs and their sources are presented in Box 6. Costs for endoscopy OGD and ultrasound scans (USS) were taken from the NHS Reference Costs 2013-2014. Where costs were not available from national databases they were obtained from the literature. This was the case for alpha-fetoprotein (AFP) test and for liver biopsy which utilised unit costs a HTA report on antiviral therapy for mild chronic hepatitis C.\textsuperscript{796} Finally, drug costs were taken from the British National Formulary. Costs were adjusted using 2015 prices and were discounted at 3.5%. Costs in each arm of the trial for biopsies, USS, AFP, OGD and beta blockade are given in Tables 143–147.

Statistical results showed that the average number of biopsies performed is very similar in the two arms of the trial, 0.1 (SD 0.3). The cost is just over £13,000 in both arms.

Statistical results also showed that diagnosis of cirrhosis (defined in ELUCIDATE as an ELF Score $\geq 9.5$) leads to higher frequency of USS, AFP tests and endoscopy OGD in both trial arms – as it should be, given that these constitute process of care in cirrhotic patients. USS, AFP tests and endoscopy OGD though are more frequent in patients in the ELF arm because they are more likely to be diagnosed with cirrhosis. This translates into a higher average cost of USS for the ELF arm, specifically £28,935 against £15,503 in the standard arm.

The same applies to AFP tests. AFP measurements are process of care in cirrhotic patients, hence why these tests were administered more frequently in the ELF arm. On average, £7,670 was spent on AFP tests in the ELF arm, whilst £5,458 is the average cost of AFP tests in the standard arm.

Statistical results showed that there is a significant difference in the mean values of numbers of OGD performed, 0.4 (SD 0.6) in the ELF arm and 0.03 (SD 0.2) in the standard arm. This translates into a remarkable difference in cost: £66,653 spent on endoscopies in the ELF arm against £5,022 spent in the standard arm.
It was not possible to distinguish between use of beta blockers or band ligation (the data recorded only if they had been treated with either) so we assumed that first choice of treatment was the drug therapy. This is supported by the UK guidelines on the management of variceal haemorrhage in cirrhotic patients, which advise towards pharmacological treatment with Propranolol as first line (40 mg twice daily). Carvedilol (12.5 mg once daily) or Nadolol (40 mg once daily) are suggested as alternatives to Propranolol. Once initiated, the treatment continues indefinitely. In the ELF arm, five patients were treated for varices; whilst in the standard arm three patients were treated. We assumed that treatment started in the last year of follow-up and we calculated the cost for one year. The least expensive course treatment is carvedilol which would have treated the patients in the ELF arm for £74 and patients in the standard arm for £44. The most expensive course is Propranolol with £170 for the ELF arm patients and £102 for the standard arm.

Severe complications and deaths

Severe complications are the primary outcome and will be analysed using registry (ONS/HES) data in the long-term follow-up analysis in 2021. In this report, the number of severe complications and the number of deaths are therefore not broken up by arm.

16 (1.8%) of all randomised patients had at least one severe complication. When looking at the first identified severe complication in these 16 patients, hepatocellular cancer was the most common with 10 cases (62.5%). Two patients had variceal haemorrhage and 2 encephalopathy. Two patients with hepatocellular cancer and 1 patient with encephalopathy subsequently died of liver-related causes.

17 of 878 randomised patients (1.9%) died during follow-up. 5 of these deaths (29.4%) were liver related (1 variceal haemorrhage, 2 HCC, 1 liver failure, 1 sepsis), for one patient this information is missing. Hepatocellular cancer was the most common liver related cause with two cases (40.0%).

Discussion

Although there was relatively low compliance to trial procedures, this was not unexpected in this group of patients. It should also be noted that fairly strict definitions of compliance were reported, with a tight window, usually of one month, in which to have the test performed, and
the ‘window’ graphs show that the majority of tests were still done, if often considerably later than would be ideal. Furthermore, in terms of looking at a series of tests, if one test at one particular time point was missed, that test would have been likely to have been carried out at the next visit, reducing the impact of the missed test.

Again, the analysis of the process of care outcomes show substantial differences in the care delivered between the ELF arm and the control arm, despite the apparent lack of compliance. Furthermore, when considering tests performed after the diagnosis of cirrhosis, which is defined by ELF threshold value in the ELF arm, and therefore includes vastly more patients in the ELF arm, differences in test such as Alpha-Fetoprotein (AFPs), Ultrasounds, and Oesophago-Gastro-Duodenoscopies (OGDs) are very large (18 times as many AFPs, 17 times as many Ultrasounds, 13 times as many OGDs), so it seems clear that a very different treatment package has been delivered between trial arms despite the low compliance under the definition used.

The main effect of low compliance, in terms of the design and outcome of the trial, will be its potential impact on the effect size for the primary endpoint of occurrence of severe complications. In order for the intervention to be successful, it has to be successfully delivered, otherwise it cannot be expected to result in a difference in outcome. It is difficult to judge whether the magnitude of the low compliance would be sufficient to cause a substantial reduction in the effect size. However, the prior comments about the stringency of compliance reporting, combined with the large differences found in the use of tests such as AFP, ultrasound and OGD, are crucial, and any reduction in effect size due to low compliance, as defined, would be expected to be relatively minor. We will be able to analyse effect size in relation to adherence to protocol guidelines in the final analysis, at which time it should be possible to quantify any such effect. This analysis will be included in the Statistical Analysis Plan for analysis of the long-term follow-up for the primary endpoint.

The unsurprisingly higher cost associated with the ELF arm (£116,629 against £39,345) is due to the additional care, mainly in the form of additional diagnostic investigations, provided as a consequence of more patients with ELF score ≥9.5, hence diagnosed with cirrhosis. This additional care is directed at preventing complications (in the case of beta-blockers therapy) or detecting complications early and at a curable stage. This should translate in improved survival rates and improved quality of life at the end of the extended
follow-up. The cost-effectiveness of the ELF test in the early detection of cirrhosis will ultimately depend on the impact that the additional process of care in the ELF arm has on survival rates, quality of care and health care resource use.

The descriptive analysis has a number of limitations and assumptions. Information was not available on when the diagnostic investigations occurred during the 30 month follow-up, so we assumed that, for example, all USS happened in the last year of follow-up and we discounted costs on the third year. With regard to treatment of varices, patients can be treated surgically or with pharmacological treatment; we assumed that patients were all treated with beta-blockers, based on UK guidelines which advise on pharmacological treatment as first line. Finally, information on when treatment was initiated was not available, hence we assumed treatment starts in the last year of follow up and we discounted cost on the third year. Also, we did not know which particular drug was used; thus we included the cost of both the most and least costly treatment courses.

These data will allow us to complete the planned evaluation of the cost-per-QALY generated by the ELF monitoring strategy when the outcome data become available. In the mean time we will explore further modelling approaches to relate the cost of the strategy against potential improvements in outcomes that may be estimated from the changes in process of care.

<table>
<thead>
<tr>
<th>Table 143 Liver Biopsy</th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>0.1 (0.3)</td>
<td>0.1 (0.3)</td>
<td>0.1 (0.3)</td>
</tr>
<tr>
<td>Average Cost</td>
<td>£ 14,635</td>
<td>£ 14,702</td>
<td>£ 29,337</td>
</tr>
<tr>
<td>Discounted cost</td>
<td>£ 13,200</td>
<td>£ 13,260</td>
<td>£ 26,460</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 144 Ultrasound scans</th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>1.5 (1.52)</td>
<td>0.8 (1.08)</td>
<td>1.2 (1.36)</td>
</tr>
<tr>
<td>Average cost</td>
<td>£ 32,081</td>
<td>£ 17,188</td>
<td>£ 51,447</td>
</tr>
<tr>
<td>Discounted cost</td>
<td>£ 28,935</td>
<td>£ 15,502</td>
<td>£ 46,402</td>
</tr>
</tbody>
</table>
### Table 145 AFP test

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>2.4 (1.82)</td>
<td>1.7 (1.53)</td>
<td>2.1 (1.72)</td>
</tr>
<tr>
<td>Average cost</td>
<td>£ 8,504</td>
<td>£ 6,051</td>
<td>£ 14,916</td>
</tr>
<tr>
<td>Discounted cost</td>
<td>£ 7,670</td>
<td>£ 5,458</td>
<td>£ 13,454</td>
</tr>
</tbody>
</table>

### Table 146 Endoscopy OGD

<table>
<thead>
<tr>
<th></th>
<th>ELF (n=438)</th>
<th>Standard (n=440)</th>
<th>Total (n=878)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>0.4 (0.6)</td>
<td>0.03 (0.2)</td>
<td>0.2 (0.5)</td>
</tr>
<tr>
<td>Cost</td>
<td>£ 73,899</td>
<td>£ 5,567</td>
<td>£ 74,068</td>
</tr>
<tr>
<td>Discounted cost</td>
<td>£ 66,653</td>
<td>£ 5,022</td>
<td>£ 66,805</td>
</tr>
</tbody>
</table>

### Table 147 Treatment with beta-blockers

<table>
<thead>
<tr>
<th>Treatment</th>
<th>1 Year cost for standard arm (3 patients)</th>
<th>1 Year cost for ELF arm (5 patients)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propranolol</td>
<td>£102</td>
<td>£170</td>
</tr>
<tr>
<td>Carvedilol</td>
<td>£44</td>
<td>£74</td>
</tr>
<tr>
<td>Nadolol</td>
<td>£88</td>
<td>£147</td>
</tr>
</tbody>
</table>
### Box 6 - Unit costs and their sources

<table>
<thead>
<tr>
<th>Activity</th>
<th>Unit Cost</th>
<th>2015 prices</th>
<th>Comments</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Liver Biopsy</strong></td>
<td>£ 249.00</td>
<td>£ 334.14</td>
<td>Average across three hospitals</td>
<td>Wright et al. 2006 (HTA Report).</td>
</tr>
<tr>
<td><strong>Alpha-Fetoprotein test</strong></td>
<td>£ 6.03</td>
<td>£ 8.09</td>
<td>Average across three hospitals</td>
<td>Wright et al. 2006 (HTA Report).</td>
</tr>
<tr>
<td><strong>Ultrasound Scan</strong></td>
<td>£ 47.00</td>
<td>£ 48.83</td>
<td>-</td>
<td>National schedules Reference cost (2013-14) - Diagnostic imaging</td>
</tr>
<tr>
<td><strong>Endoscopy</strong></td>
<td>£ 406.00</td>
<td>£ 421.80</td>
<td>-</td>
<td>National schedules Reference cost (2013-14) - Day case</td>
</tr>
<tr>
<td><strong>Beta-Blockers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propranolol</td>
<td>£ 1.45</td>
<td></td>
<td>40 mg, 28 tab pack Dose: 40 mg twice daily</td>
<td>BNF Nov 2015</td>
</tr>
<tr>
<td>Carvedilol</td>
<td>£1.26</td>
<td></td>
<td>12.5 mg, 28 tab pack Dose: 12.5 mg once daily</td>
<td>BNF Nov 2015</td>
</tr>
<tr>
<td>Nadolol</td>
<td>£5.00</td>
<td></td>
<td>80 mg, 28 tab pack Dose: 40 mg once daily</td>
<td>BNF Nov 2015</td>
</tr>
</tbody>
</table>
Chapter 22 - Workstream 3 – Next Steps and Preliminary Conclusions
Summary and Discussion of the Results of ELUCIDATE to date

At the time of compiling this report the data gathered during the conduct of ELUCIDATE have only been divided by randomization arm for the analysis of indicators of cirrhosis and changes in the process of care.

A number of observations can be made concerning two broad categories of data; those relating to the clinical aspects of the study and those relating to the conduct of the study.

Clinical Aspects

Baseline Demographics
Randomisation was effective in that there are no significant differences in the baseline characteristics between the two arms. Median age was 55yrs and there was a slight overrepresentation of men at 56%.

ELF at registration
The ELF values at registration were median 9.4. Nearly 55% of those registered had ELF scores in the range 8.4-9.49 indicating that they were pre-cirrhotic, and at low risk of liver related complications of CLD (Figure 64, Chapter 19). Only 5.7% (50 participants) exceeding 11.49 and thus at higher risk of liver related complications within the next 5 years. This distribution of liver fibrosis amongst the participants means that only a relatively small proportion are likely to develop liver related outcomes during the course of the trial. Longer term follow up will be required to capture these events and to determine if earlier detection of cirrhosis and management alters the course of disease and the incidence of morbidity and mortality.

The change from a randomisation threshold to 8.4 in March 2011 did not introduce bias but did reduce the risk of serious liver related complications in the trial overall (see Figure 64).

Alcohol Consumption
The levels of alcohol consumption in the cohort were relatively low compared to what was anticipated for patients with CLD. Only 9.3% reported moderately heavy levels of current
consumption and only 3.5% reported current heavy alcohol consumption. Again this lower than anticipated alcohol consumption is likely to reduce the incidence of liver related events in the cohort during the trial and during subsequent follow-up. Patients may have under reported alcohol use in both arms of the trial.

Aetiology of Chronic Liver Disease

The distribution of aetiologies of CLD amongst the cohort differs from that seen in the majority of liver clinics with alcoholic liver disease underrepresented at 7.7% and viral hepatitis overrepresented at 40.2%. Non-alcoholic fatty liver disease accounted for 25% of the cohort and is represented at a level similar to most liver clinics. The course of ELUCIDATE paralleled the introduction of highly effective therapies for HCV infection and more widespread treatment of HBV. As a consequence it is likely that patients recruited with these conditions will have been treated during the course of the trial and are likely to experience improvement in their liver fibrosis. While this is likely to reduce the anticipated incidence of liver related events it will have created the opportunity to analyse the ability of ELF and other biochemical tests to monitor improvements in fibrosis consequent on control or eradication of hepatitis virus infection.

It is unfortunate that so few patients were recruited with alcoholic liver disease as their primary aetiology as these patients are at highest risk of liver related events and are likely to be the group in which early detection of cirrhosis might be most beneficial. However patients with alcoholic liver disease are often difficult to recruit and retain in clinical trials.

The significant representation of NAFLD may be valuable. Fatty liver disease is increasingly recognised as an important cause of CLD and one in which the currently available therapeutic interventions (diet and exercise) have limited effect. However new specific and general antifibrotic therapies are undergoing trials in NAFLD so it will be valuable to know what strategies are most effective in detecting cirrhosis in this condition.

The proportion of patients with viral hepatitis amongst those recruited varied considerably between sites from 0 to 100%. This is likely to reflect the specialist interest of the local principle investigator, the representation of these patients in the local clinic population and the willingness of the investigators and patients to enroll these patients in the study.
**Progression to Cirrhosis**

Inevitably the proportion of patients progressing to “cirrhosis” as defined by ELF testing was greater in the ELF arm than in the Standard arm (64.2% compared to 4.5%). However the definition of cirrhosis for each arm differed at this stage of the analysis, being defined by ELF score > 9.5 in the ELF arm and on clinical criteria in the Standard arm. Patients with ELF scores of > 9.5 have a significant risk of serious complications (Figure 64) and the investigators and Steering Group judged that to justify the initiation of measures to reduce the risk of serious complications.

Seventy percent of the patients registered were diagnosed as cirrhotic at randomization at the time of their first trial test in the ELF arm. None of the patients in the Standard arm were diagnosed as cirrhotic when they attended their randomization visit. This difference supports the hypothesis that ELF will detect cirrhosis in patients who would not be recognized as cirrhotic using clinical criteria. However this may merely reflect a difference in case definition or at best “lead time” bias. Longer term follow up will determine what proportion of patients in each arm progressed to clinically relevant outcomes associated with cirrhosis.

Only 9 patients in the ELF arm (0.4%) were diagnosed as cirrhotic using criteria other than ELF. The commonest method for diagnosing cirrhosis in the Standard arm were elastography (3.7%), ultrasound (3%), gastroscopy (2.7%) and clinical judgement (2.7%).

Following randomization 84 patients in the ELF arm and 20 patients in the Standard arm were subsequently diagnosed as cirrhotic during the course of the trial. 3 patients in the ELF arm were diagnosed using clinical measures (1 each of biopsy, elastography and clinical judgment). Ultrasound and elastography were the commonest means of diagnosing cirrhosis in the Standard arm patients with 6 attributable to each method.

**Study Conduct**

**Recruitment**

Recruitment to ELUCIDATE was slower than anticipated and start up time was longer than anticipated in many Centres. However, 46 NHS Trusts participate in ELUCDATE, with wide
geographical distribution across the country. Many hospitals had little previous experience of recruitment into liver RCTs. The extension of the trial to additional Centres ultimately allowed the recruitment of almost 90% of the target but the delays have limited our ability to yet analyse long term outcomes.

Compliance

Pre-cirrhotic
Compliance with follow-up visit attendance was generally poor in both arms. This reflects clinical experience in liver clinics where “Did Not Attend” rates vary between 5 and 25%. The ELUCIDATE protocol required 6 monthly attendance at clinics for pre-cirrhotic patients. It was apparent during the feasibility planning for the study that many clinics booked less frequent appointments for patients with CLD. It was decided that rather than change the protocol, this variance in clinic practice would be captured in the conduct of the study. Thus the low rates of compliance with the protocol were anticipated. Only 23.3% of patients complied with all planned visits and the rate of compliance was higher in the ELF arm at 29.8% compared to 19.7% in the standard care arm.

Post-cirrhotic
Compliance with the protocol was even worse for patients following the diagnosis of cirrhosis. Again this is likely to be due to local clinical practice deviating from national and international guidelines as well as the ELUCIDATE protocol. Compliance with clinic visits was worse in the ELF arm at 16.2% than in the Standard arm at 21.1%. This may reflect greater concern for patients diagnosed as cirrhotic in the Standard arm but the number of patients complaint with the protocol in the Standard arm is very small (n=4).

Reasons for non Compliance
The most frequently reported reason for non-compliance with clinic visits in the study were administrative errors or lack of staff. Interestingly the majority of these deviations occurred at UCLH and the Royal Free where large numbers of patients were recruited and King’s College Hospital.

ELUCIDATE provided no separate funding for research nurses to conduct the study. It was anticipated that NIHR supported staff would recruit and consent patients to the study and this
was certainly the case at the majority of sites. However the relatively light research tasks involved in the conduct of ELUCIDATE became “onerous” at those centres recruiting large numbers of patients when processing large numbers of blood samples and case report forms became time consuming. These observations have implications for future NIHR studies that seek to rely on Portfolio adoption and access to NIHR faculty to conduct “non-onerous” tasks critical to the completion of the study. Direct costing of the research component of observational studies should be considered but this will have significant impact on the cost of studies such as ELUCIDATE.

Smaller centres with fewer competing studies performed particularly well and this must be recognized as a success of the NIHR goal to have more NHS Trusts engaged in clinical research that is relevant to the NHS.

**Compliance with Cirrhosis Management**

Compliance with AFP testing was low in both arms at 12.6% and 11.8% in the ELF and Standard arms respectively.

**Process of Care Outcomes**

**Frequency of biopsies**

There was no significant difference in the number or proportion of patients undergoing liver biopsy in each arm of the study. This suggests that clinicians did not regard an ELF based diagnosis of cirrhosis as an indication for a liver biopsy.

**AFP tests**

A diagnosis of cirrhosis should be followed by monitoring for the development of hepatocellular cancer. Guidelines recommend the use of AFP measurement and ultrasound scanning every 6 months and this practice was incorporated into the ELUCIDATE protocol. Measurement of AFP was more frequent in ELF arm than in the Standard arm confirming that the investigators adhered to the protocol and conducted appropriate tests following the diagnosis of cirrhosis by ELF. Once cirrhosis was diagnosed the adherence to protocol was same in both arms.
This suggests that if ELF correctly defines cirrhosis then clinicians will test for AFP as a screening test for HCC. This suggests that cancers may be diagnosed more frequently and earlier through the use of biochemical testing for cirrhosis and raises the possibility that more HCC’s could be cured and that the outcomes in HCC could be improved. It remains to be determined whether this strategy will improve survival from HCC.

**Ultrasound Scans**
More ultrasound scans were performed in the ELF arm than in the Standard care confirming that the investigators adhered to the protocol and conducted appropriate tests following the diagnosis of cirrhosis by ELF. Once cirrhosis was diagnosed the adherence to protocol was same in both arms.

This suggests that if ELF correctly defines cirrhosis then clinicians will test for screen for HCC appropriately. This suggests that cancers may be diagnosed more frequently and earlier through the use of biochemical testing for cirrhosis and raises the possibility that more HCC’s could be cured and that the outcomes in HCC could be improved. It remains to be determined whether this strategy will improve survival from HCC.

**OGD**
The overall frequency of OGD was 34.9% in the ELF arm compared to 2.7% in the Standard arm. Following the diagnosis of cirrhosis the frequency of OGD was similar in both arms with 54.4% of ELF arm cirrhotics and 60% of Standard arm cirrhotics undergoing endoscopy. This is a critical investigation for diagnosing treatable varices and so is an important step in reducing the incidence of life threatening complications of cirrhosis. As more patients underwent OGD for a cirrhotic indication in the ELF arm (and randomized controlled trial evidence has shown that treatment of varices reduces morbidity and mortality in CLD), the greater use of OGD in the ELF arm may translate into an overall benefit in terms of a reduction in morbidity and mortality in the ELF arm over time.

**Diagnosis of varices**
The number of patients deemed to have developed varices was greater in the ELF arm than in the Standard arm (9 or 2.1% compared to 7 or 1.6%) but this difference was not statistically significant.
Of the patients diagnosed as cirrhotic who had an OGD in order to detect varices 9/153 (5.9%) in the ELF arm and 7/12 (58.3%) in the Standard Care arm were found to have varices. At this stage of analysis it is not possible to determine the diagnostic accuracy of either strategy but the numbers of cases detected was similar in each arm suggesting that either strategy is equally effective in detecting varices.

Six of the 20 cases of cirrhosis in the Standard Care arm were detected using Fibroscan. As an alternative method of non-invasive testing for fibrosis with similar effectiveness to ELF, the use of Fibroscan in the Standard arm will have enriched for cirrhosis in the control arm, reducing the measured effectiveness of the intervention (ELF). Access to Fibroscan is limited in the UK and so once unblinding has occurred it will be possible to discern the impact of Fibroscan on the detection of cirrhosis and treatment in the control or Standard Care arm.

**Beta-Blocker prescription or Band Ligation**

The number of patients initiated on treatment for portal hypertension and oesophageal varices was greater in the ELF arm compared to the Standard arm (5 compared to 3 or 1.1% vs 0.9%). Of the cases of diagnosed varices 5/9 in the ELF arm and 3/7 in the Standard Care arm began treatment. Neither difference was statistically significant. This suggests that the instigation of life saving treatment was no greater in the ELF arm and suggests that the study is unlikely to provide evidence of benefit from early diagnosis of cirrhosis. However the numbers reaching this endpoint to date are small and, at this stage the impact of the use of Fibroscan in the Standard Care arm cannot be determined.

**Treatment**

Similar numbers of patients underwent treatment to normalize their liver function during the course of ELUCIDATE with 13% of ELF patients and 10.9% of Standard care patients receiving treatment. This suggests that eradication or control of viral hepatitis is unlikely to account for the lower than anticipated incidence of complications of cirrhosis during the trial study period.

**The Trial Outcomes**
The ELUCIDATE trial was designed to test the hypothesis that earlier diagnosis of cirrhosis in patients with chronic liver disease through the use of serum testing would permit earlier and more widespread targeted screening for oesophageal varices and hepatocellular cancer linked to treatment and that this would result in reductions in morbidity and mortality. When compared to Standard Care, more patients in the ELF arm were diagnosed with cirrhosis and then underwent screening for oesophageal varices and hepatocellular cancer. The excess of cases of cirrhosis in the ELF arm was anticipated as the criterion for diagnosing cirrhosis in the ELF arm was a universally applied biochemical threshold for ELF score > 9.5 while diagnosis of cirrhosis in the Standard Care arm relied on clinical recognition. Whether this early diagnosis proves to be of long term clinical benefit remains to be determined. It is inevitable that earlier diagnosis will incur greater costs attributable to investigation and preventative treatment of complications. However these costs may be exceeded by the costs of managing decompensated liver disease presenting de novo in patients who have been diagnosed with cirrhosis later.

The rate of screening amongst cirrhotic cases detected in each arm was similar, suggesting that the investigators adhered to the trial protocol and complied with guidelines for the management of cirrhosis equally for cirrhotic patients in each arm in accordance with the hypothesis underpinning the trial.

The number of cases of oesophageal varices diagnosed and treated were similar in each arm and thus the proportion of cirrhotic patients in each arm with varices and started on treatment were far higher in the Standard Care arm. This suggests that Standard Care may be more efficient and equally as effective as the use of ELF testing to detect and treat varices although at the present stage we do not know if cases of varices were missed in the Standard Care arm. Six of the 20 cases of cirrhosis in the Standard Care arm were diagnosed following the use of Fibroscan, a non-invasive test for cirrhosis that performs as effectively as ELF. Because the randomization blinding has not been broken we do not know at this stage what proportion of the patients with varices in the Standard arm were diagnosed using Fibroscan but it is clear that a third of the cirrhotic cases in the Standard arm were detected using this non-invasive test.

Fibroscan is a technology that emerged into use in the NHS during the course of the ELUCIDATE trial. It was not CE marked when the trial was initiated and access to the
technology remains limited within NHS Trusts. The authors of the trial did not include Fibroscan as part of the intervention arm of the trial due to the lack of regulatory approval at the start of the trial, limited access to the technology and the lack of consensus on thresholds for the diagnosis of cirrhosis. However during the course of the study Fibroscan has entered more widespread, but not universal use. While the use of an alternative novel non-invasive test for fibrosis may be regarded as “contaminating” the control arm of the trial, the Chief Investigators took a pragmatic decision to not exclude the use of Fibroscan but to record its use by investigators in both arms of the trial. It may be that the use of Fibroscan in the Standard Care arm will have eliminated the intended difference between the arms and hence the power of the study to detect any benefit of early detection of cirrhosis. Once the trial is un-blinded it may be possible to determine the proportion of cases of cirrhosis diagnosed and managed in the Standard care arm using Fibroscan as an alternative to ELF as a screening NIT.

Another explanation for the high pick-up rate for varices amongst the cirrhotic patients in the Standard arm is a “Hawthorne Effect.” It is likely that investigators were more vigilant for the onset of cirrhosis in patients in the Standard Care arm than for patients outside the trial. If one of the benefits of conducting ELUCIDATE in a large number of NHS Trusts is diffusion of best practice then this should be seen as a benefit of conducting research in clinical services. This possibility could be explored through qualitative research conducted with the principle investigators.

If the use of ELF to detect cirrhosis is beneficial then those patients most likely to benefit will be those who are asymptomatic and have few clinical signs of cirrhosis. These patients are unlikely to be diagnosed through Standard Care and are likely to be over-represented amongst the patients presenting for the first time with life threatening variceal haemorrhage and inoperable HCC. It remains to be seen if complications will be detected and managed more effectively in this group of patients in the ELF arm compared to the Standard Care arm but this should be discernable with longer follow-up of the two arms once randomization is un-blinded.

**The Trial Process**
The trial process was ultimately successful in approaching the final target but with important delays. The randomization process was effective in that there were no significant differences between the patients recruited to the two arms of the trial. However the recruitment process was slower than anticipated due to a number of problems with the regulatory processes, enforced changes in the protocol and problems at the participating sites. Much has been learnt through the conduct of the study and some of the issues with regulatory processes have been addressed by NIHR and will be addressed by the Health Research Authority. It is apparent that some of the smaller centres made proportionally greater contributions to recruitment than some of the larger centres. It was clear that site specific visits by the Co-Principal Investigator made a significant difference to recruitment.

The spectrum of aetiologies of chronic liver diseases amongst the participants was similar to that seen in the outpatient departments of specialist centres but this differs from the case-mix presenting to A&E departments and admitted to hospital wards. Thus compared to hospital in-patients the cohort under represented alcoholic liver disease and over represented viral hepatitis and NAFLD. However as the purpose of the study was to investigate the impact of detecting cirrhotic cases amongst hospital clinic attenders this is a representative sample.

The spectrum of liver fibrosis amongst trial participants reflected the trial design and focussed on pre-cirrhotic patients. While this targeted the cohort of patients most likely to benefit from early diagnosis of cirrhosis it meant that longer follow-up was required than allowed for in the period of the Programme Grant. Thus a later follow-up will be required. As this will not incorporate continued monitoring for progression of fibrosis this will diminish any potential benefit from the intervention, limiting the main benefit to those patients reaching cirrhosis during the on-trial monitoring period of 30 months.

The patients most likely to benefit from monitoring for cirrhosis are those with progressive liver disease who develop cirrhosis during the study period. The modest levels of drinking (again possibly due to a Hawthorne effect) and the introduction of effective therapies for HCV and HBV during the conduct of the study are likely to reduce the proportion of patients with progressive fibrosis.
Compliance with trial visits was moderate and provides a good reflection of clinical practice. These data will permit more accurate estimates of feasibility and power when designing future monitoring and interventional studies in CLD.

Longer term follow-up and analysis of the process of care, clinical events and outcomes by randomisation arm as well as investigation of the impact of non-invasive screening for cirrhosis in the control (Standard Care arm). Will have to await un-blinding of the trial and the passage of time. However frameworks for the analysis of the data can now be developed to permit a full evaluation of the trial in due course.

**Long term follow up studies**

When the DMEC determine that the study can be fully unblinded we will conduct an analysis of outcomes in each arm as planned. Specifically we will determine the numbers and proportions of patients in each arm who:

- Were diagnosed with cirrhosis
- Underwent OGD
- Had AFP measurement
- Ultrasound scans
- Were diagnosed with HCC
- The size and number of HCC at the time of diagnosis
- Underwent treatment for HCC
- Were diagnosed with Oesophageal varices
- Started therapy for oesophageal varices, including Band ligation Beta blockers
- Presented with haematemesis due to portal hypertension
- Presented with ascites
- Presented with spontaneous bacterial peritonitis
- Presented with encephalopathy
- Presented with any other complication of cirrhosis
- Underwent liver transplantation
- Died from any cause
Died due to a liver related cause

In the ELF arm, cases diagnosed with cirrhosis will be analysed in order to determine the true and false positive rates for ELF by assigning a diagnosis of cirrhosis using clinical parameters including:

Non-invasive blood tests for cirrhosis including:
- Platelet count
- AST/ALT ratio
- APRI
- FIB4

Fibroelastography including Fibroscan and ARFI

Imaging including
- USS
- CT
- MRI

Clinical judgement of the clinician where recorded.

In cases diagnosed as having cirrhosis in the Standard Care arm, we will attempt to determine the impact of the use of Fibroscan in the “control” arm by comparing the number and rate of diagnosis of cirrhosis in participants at those centres using Fibroscan with that at centres that had no access to Fibroscan.

The accuracy of ELF and Standard Care in diagnosing clinically important cirrhosis will be determined by comparing the number and proportion of cases developing clinical signs of decompensated cirrhosis, hepatocellular cancer, transplantation or death in each arm at the time of censoring. These rates will be compared in those cases undergoing treatment for portal hypertension as well as the total numbers of cases per arm.

We will evaluate the impact of treatment for underlying chronic liver diseases by analysing the frequency and rates of treatments of disease by aetiology in each arm. Specifically we will investigate changes in ELF score in response to therapy in the ELF arm. We will determine any differences in treatment rates by disease status and by arm.
We will investigate the correlation between ELF score at randomisation and the incidence of complications of cirrhosis in all participants.

**Feasibility of the long term data collection: Sustaining data collection from the end of the active period of follow up to the long term outcomes**

There are few prospective inception cohort studies of patients with chronic liver disease investigated for long-term complications of CLD. We have successfully monitored cohorts of patients with liver disease using ONS and HES data in previous studies. In a former study we followed those patients initially recruited for the original ELF study of biomarkers of liver fibrosis from which the ELF test was derived. Using routine data sources including death certification and detailed interrogation of clinical case notes. For those patients lost to follow up by the hospital services we contacted their General Practitioners. Using these methods we were able to obtain data on over 95% of the patients recruited into the study. While examining the case notes of the participants was feasible it proved to be labour intensive, necessitating one or two investigators spending one to two days extracting data at each participating site.

In a separate study of middle-aged women participating in screening for ovarian cancer (UKCTOCS) who had provided prior consent for follow-up using routine data sources we investigated their incidence of hospital admissions and deaths from liver-related conditions using routine data sources and their NHS numbers. In this study of 110,000 women we established a set of ICD-10 codes that could be used to define the incidence of liver related outcomes and showed that NHS numbers could be used to access these data through routinely gathered, centrally stored hospital episode statistics.\(^798\)

The cohort of patients with chronic liver disease recruited for the ELUCIDATE trial represent the range of chronic liver diseases commonly encountered in NHS clinical practice and were recruited at a similar stage of disease severity. From the outset we proposed to follow the cohort for their long-term morbidity and mortality in order to obtain a clearer picture of the course and consequences of CLD in the NHS.
It became apparent when the threshold for entry into ELUCIDATE was reduced from 11.0 to 8.4 in March 2011 that a minority of patients recruited would develop liver related morbidity or mortality during the course of the funded period of the study. Recognising this, the Trial Steering Group agreed to plan for a separately funded follow-up of the ELUCIDATE cohort at 5 and 10 years. We consulted with NHS Information Centre (now Health & Social Care Information Centre) to discuss data collection on the trial cohort at two time points, including exploration of process, costs, and anticipated outputs. We were informed that this was a provided service and accessed by many health and research teams, and that we should ensure consent for such a follow-up was obtained from the patients at the time of recruitment to the trial and that 6 months prior to the censoring date a request should be submitted to the HSCIC for mortality and morbidity data on the cohort of recruited patients.

The proposed methodology is to use routine data sources including death certification, cancer registries and Hospital Episode Statistics to gather clinical outcome data on the whole cohort. The NHS numbers of all participants who have provided informed consent for long-term follow-up will be provided to the NHS Health and Social Care Centre (http://www.hscic.gov.uk) to obtain details of their clinical encounters with NHS secondary care providers. In addition to HES data, cancer registries and death certification registries will be searched for morbidity and mortality using the participants’ NHS numbers.

Management

The long-term follow up for clinical outcomes will be conducted at 5 years after the completion of the ELUCIDATE study. A working group led by the Chief Investigator, Prof William Rosenberg, comprising Prof Peter Selby, Prof Walter Gregory and Dr Julie Parkes will meet on an annual basis to ensure a sustained effort to deliver the additional data. The full protocol will be developed and renewed annually, and will take into account the relevant processes for application to HSCIC at the time of censoring.

Planned Analyses

Using NHS numbers, patients’ records will be surveyed for clinical outcomes associated with chronic liver disease including the following ICD-10 codes: K70, K73, K74, K76, I850, I859,
Z944 and C220. Codes K70, K73, K74 and K76 relate to chronic liver disease or cirrhosis. I850, I859, Z944 and C220 code for events associated with decompensation of chronic liver disease.

Cancer registries will be interrogated for incidence of hepatocellular cancer and cholangiocarcinoma. The UK Liver Transplant register will be interrogated for liver transplantation, the indication and outcome of transplantation.
**Anticipated Outcomes from the Long-term Follow-up:**

The long-term follow up of the ELUCIDATE cohort is of considerable importance to the full evaluation of the trial’s impact and will permit us to:

- determine the relationship between change in ELF and incidence of liver related events. These analyses will provide the definitive assessment of the clinical effectiveness of ELUCIDATE and provide the data upon which full health economic analysis of the cost effectiveness of use of biochemical strategies for the early diagnosis and management of cirrhosis.
- determine the prognostic performance of the ELF test in predicting liver disease morbidity and liver and all-cause mortality;
- determine the incidence of clinical events in patients diagnosed as having cirrhosis based on the ELF test and;
- to compare the incidence of liver related events in patients diagnosed with cirrhosis based on ELF and based on standard clinical criteria.

**Limitations**

The ELUCIDATE trial is subject to several limitations. The slow start and slow initial recruitment (despite the “late surge” in recruitment) mean that prolonged follow up is now needed for us to be able to report the primary, health related, endpoints. Our focus was always on an “end to end” trial (from diagnostic test to patient outcomes and service outcomes), so this is disappointing. However, the lessons learned and the “process of care” analysis will be very valuable as an exemplar trial. We run the risk that new technologies will supervene during the follow up period. Compliance with the trial process was satisfactory but compliance with the interventions required by the test outcomes was lower than expected which may also reduce the impact on the primary outcomes.

ELUCIDATE is certainly of value as an example of an exemplar trial in a initially important area of healthcare. It should help those who are strategically planning the evaluation of new biomarkers to judge the feasibility of timely delivery of end-to-end trials and judge the pace of alternative strategies which we discuss in Chapter 24.
Chapter 23 - Patient and public perspectives
The multiple possible roles for biomarkers in patient management are likely to influence patient experience in many different ways. Monitoring of disease progression has potential to cause anxiety, but also to stimulate possible lifestyle change by bringing home the reality of the liver damage to the patient, much earlier. Little research has been carried out into the psychosocial aspects of biomarkers, and the possible benefits and harms of their use need to be examined explicitly in future studies.

This chapter is in two parts. The first part presents the results of the research exit interviews with patient participants in the RCT reported in the ELUCIDATE RCT Workstream (WS3). The second part pulls together this information with the contributions made by PPI representatives consulted about the methodology work conducted in the Methodology Workstream (WS1) and reported in Chapter 9.

**Part one: Patients Experiences in the ELUCIDATE trial: A Qualitative study about patient experiences of taking part in a trial to test biological fluid biomarkers for Liver Disease**

This qualitative sub study of the ELUCIDATE trial aimed to explore the experiences and perspectives of patients who were enrolled on the ELUCIDATE trial, in order to provide additional insight about using the Enhance Liver Fibrosis (ELF) test as part of patient care.

The starting point for the work is that patients may have very different experiences of having their disease monitored by ELF or similar tests. For some it may be reassuring to know that they are been tested regularly and allow them a feeling of control through knowing what is going on with their health. Others might believe that the symptoms are not indicative of cancer or may be anxious about their condition potentially getting worse even though this might never happen if patients are anxious they might consequently utilise a number of coping strategies to cope with the associated distress. Anxious patients might be hyper vigilant in regards to potential symptoms, engage in information seeking, and/or develop avoidant attitudes and behaviours.
Some patients might not adhere to monitoring because they fear its iatrogenic effects while older patients might believe that they are less likely to develop neoplasm because of their more advanced age, or overall perceived lower susceptibility to cancer. Systemic factors also importantly determine surveillance experiences. Health information given to the patients, the role of healthcare professionals and previous experiences with cancer importantly determine the acceptance and understanding of monitoring practices.

Patients’ understanding of clinical biomarkers and experiences of testing, their acceptability to the patients, their perceived utility and patient experiences and motivations for testing are therefore important factors of translation of biomarkers into clinical practice. By exploring patients’ experiences of being monitored by ELF test, this study will assess how acceptable the testing is to patients and so enable better support to be provided to liver patients undergoing monitoring in the future.

Method

Thirteen patients took part in an in-depth semi-structured interview that was guided by a semi-structured topic guide.

Sites

It was planned to recruit participants from three sites, chosen to represent different sizes of institution and different types of catchment area: Leeds Teaching Hospitals NHS Trust, Bradford Teaching Hospitals Foundation NHS Trust and University College Hospitals NHS Trust. However, due to delays in obtaining permissions, it was only possible to conduct the research in two sites, Leeds and Bradford.

Sample

Patients were sampled purposively with the aim of recruiting 8 participants from the standard care arm and 8 from the intervention arm of the ELUCIDATE trial and other factors such as age and gender were considered. They had to be able to comply with the requirements of the study protocol and be able to provide written, informed consent. Patients were excluded if they were unable to comply with the requirements of the protocol or could not provide informed consent.
Recruitment
Eligible patients were first approached by a research nurse who was part of the hepatology team. They provided each patient with a patient information pack containing an information sheet, opt out form, a demographic form and a freepost envelope. The research nurse asked for verbal consent to pass the patient’s contact details to the researchers from the University of Leeds who would conduct the interviews. All patients were given seven days from the day they were given the patient information pack to consider taking part. If after considering the information they did not want to participate they could choose to opt out by completing the opt out form and returning it in a freepost envelope; after this no further contact with the research team was made. The use of an opt out approach for recruitment was chosen as this is something that most patients find acceptable (particularly in this population of trial participants) and can minimise response bias.  

Interviews
After seven days the researchers from the University of Leeds contacted the patient to confirm they were still interested and to schedule an interview. Patients could either agree to the interview or request more time to consider the information and contact could be made at a later date. The interviews were 30-45 minutes in length and took part over the telephone or in a location that was convenient for the patient. Several studies have evidenced that there are minimal differences in the results of semi-structured interviews that are conducted over the telephone or face to face. They were audio-recorded with the patient’s consent.

The interviews were guided by a semi-structured topic guide that included questions about patient’s understanding and experience of taking part in the ELUCIDATE trial and of the ELF test. We asked them about the ELF test and the trial and whether they would recommend this type of study to patients in future.

Analysis
The interviews were professionally transcribed verbatim and managed with help of NVivo. The data was analysed inductively, with no prior hypotheses, using thematic analysis and was undertaken by two researchers independently coding for emerging themes and comparing themes and codes. The analysis was further refined by using a constant comparison and
contrastive approach, and looking for negative cases in order to examine for similarities and differences within and between the patients in different centres, and within and between trial arms.

**Results**

Thirteen patients were contacted by the research nurses at the two sites. Of these, one patient opted out of the study, one was too ill to take part and two could not be contacted at existing addresses. Ten patients were available for interview and nine interviews were completed, three with participants from the usual care arm and six from the intervention arm.

**Patient Characteristics**

All of the patients were of a similar ethnic origin, their age was mean 63.4 years with a range of (56-75 years). Six were female and three were male. The conditions which participants had were:

- Haemachromatosis – 1
- Fatty Liver – 1
- Non-alcoholic fatty Liver -1
- Methotrexate related -1
- Hepatitis C - 2
- Autoimmune hepatitis – 1
- Primary biliary cirrhosis - 1
- Not stated – 1

There was a range of educational abilities

- Secondary school - 3
- College/ diploma - 3
- University – 3

Four participants were retired, three were in professional roles, one was unable to work due to ill health and one chose not to state their employment status (Table 148).
**Table 148: ELUCIDATE Patient characteristics**

<table>
<thead>
<tr>
<th>Patient Number</th>
<th>Site</th>
<th>Gender</th>
<th>Liver condition</th>
<th>ELF arm</th>
<th>Result</th>
<th>Education</th>
<th>Occupation</th>
<th>Ethnic Origin</th>
<th>Religion</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELU01</td>
<td>1</td>
<td>M</td>
<td>Fatty Liver</td>
<td>Usual care</td>
<td>No</td>
<td>University degree</td>
<td>Professional</td>
<td>White British</td>
<td>Christian</td>
</tr>
<tr>
<td>ELU02</td>
<td>1</td>
<td>M</td>
<td>Haemachromatosis</td>
<td>Usual care</td>
<td>No</td>
<td>University degree</td>
<td>Retired</td>
<td>White British</td>
<td>Christian</td>
</tr>
<tr>
<td>ELU03</td>
<td>1</td>
<td>F</td>
<td>Methotrexate related</td>
<td>Usual care</td>
<td>No</td>
<td>College/ diploma</td>
<td>Retired</td>
<td>White British</td>
<td>Christian</td>
</tr>
<tr>
<td>ELU04</td>
<td>1</td>
<td>M</td>
<td>Not stated</td>
<td>ELF</td>
<td>No</td>
<td>Secondary School</td>
<td>Not stated</td>
<td>White British</td>
<td>Christian</td>
</tr>
<tr>
<td>ELU05</td>
<td>1</td>
<td>F</td>
<td>Non-alcoholic fatty Liver</td>
<td>ELF</td>
<td>No</td>
<td>University degree</td>
<td>Professional</td>
<td>White British</td>
<td>None</td>
</tr>
<tr>
<td>ELU06</td>
<td>2</td>
<td>F</td>
<td>Hep C</td>
<td>ELF</td>
<td>No</td>
<td>Secondary School</td>
<td>Professional</td>
<td>White British</td>
<td>Prefer not to disclose</td>
</tr>
<tr>
<td>ELU07</td>
<td>2</td>
<td>F</td>
<td>Autoimmune hepatitis</td>
<td>ELF</td>
<td>ELF indicated cirrhosis Followed cirrhosis pathway</td>
<td>College/ Diploma</td>
<td>Retired</td>
<td>White</td>
<td>Christian</td>
</tr>
<tr>
<td>ELU08</td>
<td>2</td>
<td>F</td>
<td>Hep C genotype 1</td>
<td>ELF</td>
<td>ELF indicated cirrhosis which was later found not to be the case</td>
<td>College diploma</td>
<td>Retired</td>
<td>White British</td>
<td>Buddhist</td>
</tr>
<tr>
<td>ELU09</td>
<td>2</td>
<td>F</td>
<td>Primary Billiary Cirrhosis</td>
<td>ELF</td>
<td>No</td>
<td>Secondary school</td>
<td>Unable to work due to ill health</td>
<td>White British</td>
<td>None</td>
</tr>
</tbody>
</table>
Key themes

The themes that were found tended to be related to the questions on the topic guide:

- Patients experience and understanding of the ELUCIDATE trial
- Patients experience and understanding of the Enhance Liver Fibrosis (ELF) test
- Support and information offered to patients

Generally the patients interviewed here had a good experience and found the ELF test acceptable compared to current available alternatives. This was reflected in the interviews here, an alternative view was presented by a patient who had been told she was at risk of being cirrhotic. The use of language when communicating risk of cirrhosis to patients requires attention.

Patients’ experience and understanding of the ELUCIDATE trial

Patients’ understanding of the ELUCIDATE study was generally clear, they appreciated the fact that the ELF test would indicate changes from fibrosis to cirrhosis

“my understanding is that, as biopsies are quite invasive, I found it quite invasive, and the trials are looking at different markers in the blood to see whether they could find the beginning of the cirrhosis so that you know, ……people could start the treatment maybe sooner” (ELU03)

Although some people struggled to remember what the trial entailed

“Yeah it was a bit long ago I can’t really remember.” (ELU06)

This probably reflected the length of time between taking part in the study and the qualitative exit interview. Most patients that took part in the trial appeared to do so for altruistic reasons, and partly to benefit themselves through being monitored and assessed more regularly.

“as far as I’m concerned if you can help people by being a guinea pig it’s a common sense, and it might help me as well as other people you know”. (ELU02)

“it gave me a little bit of confidence that I was involved, and people were looking out for what was going on with my liver so I was quite happy to do that.” (ELU03)

“a closer eye would be kept on that when I had the specialist scans looking at the physical nature of the liver” (ELU05)

Although one participant did refer to the fact that their treatment was expensive and so they wanted to give something back for that reason

“I just felt I had a very expensive treatment and it was the least I could do” (ELU06).
The ELUCIDATE trial compared patients in two trial arms, one where people were monitored with the ELF test and the other was routine expert care. Some patients were clear about the fact that there were two arms in the trial:

“My understanding was some people had the ELF test and some people didn’t I think they were looking to compare the two, and the outcomes” (ELU01)

Whereas others were less clear:

“No, I was just told they were doing this trial, you know, obviously to see, to try and get some way of finding out if people were going to develop any other problems later on” (ELU07)

Patients struggled to recall the difference between the arms due to length of time between interview and being on the trial. It was also because most perceived no negative impact and were used to having blood tests as part of routine care. However, patients’ expectations of the study were quite high as the opportunity to have a less invasive test was appealing:

“so obviously so yeah anything rather than going through a biopsy from my point of view, just having blood samples taken would be a lot better, so I said fine I’d go for that.” (ELU01)

“I found the biopsy quite distressing, and apparently I’ve got really tough muscles! And it took a lot of getting into my muscle, it was awful, so I was quite happy with the bloods and weights and things” (ELU03)

After considering the information most patients thought that taking part would be less of a problem than routine care:

“I didn’t think there would be any detrimental effects, not, any.” (ELU06)

Patients weighed up the potential risks and benefits and judged the ELUCIDATE trial to be low risk and low effort as demonstrated by:

“I think it was a positive thing really it certainly didn’t make me worry about it but it certainly gave me a bit more depth of understanding of what was going on I think really.” (ELU01)

The patients interviewed in this study were satisfied with the fact that they would be monitored more regularly:

“No, in fact, I mean to some degree it helped me because it had to be done every three months, and X used to make sure I got an appointment. Some of my appointments, I would have waited for far longer.” (ELU07)

The main negative issues were reported in terms of having difficulty parking at the research unit and the cost of travelling in their own vehicles:

“The only negative experience is travelling from X to [name of hospital site] and back you know and parking there.” (ELU02)
Patients’ experience and understanding of the Enhanced Liver Fibrosis (ELF) test.

Patients were not told the results of the ELF test unless there was a change in their condition. “No I did ask and he said normally, unless there was a major problem, where you needed treatment then they wouldn’t get back to you on your blood results” (ELU03)

In most cases this meant that no additional information was provided to the patient. This was an acceptable situation for seven patients that were interviewed. With any diagnostic and monitoring test some patients will show a false positive or false negative result. We interviewed a further patient who said “I was told that I was cirrhotic… But I wasn’t.” And that worried me considerably, being left, you know, just being told I was… cirrhotic, and just being told that, with no follow up, if you see what I mean. So I had to do my own follow up with my own doctor.” (ELU08)

The professional team were able to use other information to confirm that the patient was not cirrhotic. However it had caused the patient to be concerned and even though they were reassured other tests indicated that they were not cirrhotic. The nurse told me, “Oh well, you know, it’s unlikely because, you know, your blood tests are fine,” but when you’re told something like that, it’s very shocking” (ELU08)

This participant would have liked to have spoken to a doctor or been offered a further scan. Some discussion of why the professionals were certain she did not have cirrhosis could have helped. “So, I’m not really sure, I think I would have liked to have talked to a doctor and then they would have said, “Look, this really is rubbish,” you know, “You’re absolutely fine” (ELU08)

In this case the patient sought additional support from a patient forum. Being familiar with the internet and potential support options seemed to be a benefit for this participant but this might not readily be accessed by those with a lower level of health literacy. “Well yes, I mean, I was on a forum as well and we shared a lot of information on the forum about treatments that we were having at various hospitals through the country, so it was quite interesting, really” (ELU08)

Seven of the nine participants interviewed did not have changes in test results so were unable to comment on the question concerned with changes made in relation to test results. One of the two patients who had received above threshold ELF test results reported that they were now following the cirrhosis pathway, but the other said
“Well, I’m not sure, [I’d make a change] especially as the test got me wrong”. (ELU08)

Again the way in which information about monitoring tests is communicated to patients should be paid attention, this includes the description of the test and what happens when a result is not correct. The issue of the test indicating a diagnosis was also raised by this patient who said

“Yeah I’d been seeing Dr X and he explained...[they could] possibly look at some alternatives to give me a good diagnosis, one of the things would have been a liver biopsy, which......I didn’t particularly fancy, and then he explained that they were then looking at this trial, and would be looking at alternative ways of helping make diagnoses and things and asked me if I’d be interested.” (BRA01)

Support and information received as part of the ELUCIDATE trial

Overall patients had positive experiences whilst being enrolled on the ELUCIDATE trial and felt that the study was explained well

“No no not at all it was explained very well in the first place...it was very much, full detail was given and you know explanation all the way through as to what was happening and the whys and the wherefores so yeah it was good, yeah.” (ELU01)

Patients reported having a high level of trust in the professional team caring for them and doing the research with most of the study tasks and paperwork were completed alongside routine care.

“That was usually what happened, yes, occasionally it didn’t work but usually it did, and we, we linked the two together so, I went for the appointment and did the study at the same time.” (ELU01)

Completing questionnaires and associated forms was found to be acceptable

“Well, it was mainly just filling out a form, so yes, it was absolutely fine.” (ELU08)

Although one participant had noted that the order of the questions had been changed and commented

“I personally found it quite tedious, filling them in every time but I know you needed the information.” (ELU05)

Patients did not remember being signposted to sources of information about research per se. Overall those that took part in the study felt it was important “that someone is the guinea pig”
and felt that they “wanted to give something back” and did expect any further reward for example:

“They did also offer to pay my expenses which I said no, I wasn’t bothered about that”. (ELU03)

Patients talked about the fact that the research nurses went above and beyond to help and support them. Most patients that took part in these interviews would recommend taking part in this type of research study. They wanted to be able to give something back as part of a reciprocal process. In general they fitted a profile of those that usually agree to take part in research studies including those having a high level of trust in the professional team and wanting to give something back.

Patients’ perspective of being part in the ELUCIDATE study: discussion points

Patients may have had very different experiences in the ELUCIDATE trial depending on a number of factors. The specific arm of the trial a patient followed and whether they were told they were at elevated risk of cirrhosis would have influenced their experience. Additional influencing factors may be dependent upon a patient’s personal circumstances such as the condition they had, whether their employer would accommodate regular hospital appointments and the costs of taking part including absence from work, travel to hospital sites and parking. Withdrawals from the trial may indicate that ongoing monitoring is not acceptable or convenient for some patients.

The trial information provided to participants as part of the ELUCIDATE study was perceived to be sufficient with few gaps noted, although some direction to professional support or peer support should be considered in future trials. The issue of information needing to be tailored to individuals makes this difficult to get right in every case. The research nurses at both these sites provided support to patients that was found to be personal and effective.

The process of having a blood test in addition to existing tests and in place of more invasive tests was found to be acceptable for the patients. This was reflected in the experiences of patients who were not indicated to be at increased risk of cirrhosis by the ELF test. One patient we spoke to, had a different experience having been told she was cirrhotic based on a test result which was resolved in the context of other information. Participants who were
interviewed had a high level of trust in professionals and were comfortable in NHS environments. There were no reports here of people who had problems with medical tests or that had previous negative experiences though the sample was limited.

There was only one example of a patient with what was perceived as an inaccurate result so it would be good to know how well this reflected other “false positive” experiences in the trial. This patient experienced a level of shock and worry that should be further considered in terms of how this information is presented to the patient and the type of support that could be offered. (This is supported by work from PPI).

When implementing this test in clinical practice some attention needs to be paid to how the risk of cirrhosis is communicated to patients. A monitoring test is not necessarily the basis of a diagnosis however the language used for diagnostic and monitoring tests is often used interchangeably and can be confusing to both professionals and patients.

**Limitations**

The sample here is small and it’s possible that other sites may have shown different experiences. Most participants reported positive experiences and there may have been some bias in the selection and availability of patients for interview: none of those interviewed had undergone additional invasive tests in light of their ELF test result (e.g. endoscopy) and nor were those who dropped out of the study represented. This means that the generally benign nature of participants’ experiences may not fully reflect any potential for harm, as relevant accounts may have been missed. Additional information could have been collected if there had been time to develop the insight from this qualitative study into a questionnaire for a larger sample of study participants and to confirm these experiences.

**Conclusion**

Since the alternative to a blood test is a more invasive biopsy, the ELF test proves to be an acceptable method of assessment. However, caution needs to be applied and some scrutiny given over to how test results are communicated to patients as part of a considered implementation strategy for any biomarker proven to be effective in a trial. Patients have the highest expectations that ethical methods are applied at all stages. Lessons from Prostate
Specific Antigen (PSA) testing demonstrate a need to fully understand all the factors that affect test interpretation prior to implementation in clinical practice.

**Part two: drawing threads together**

Part one above reports on a piece of qualitative research, specifically addressing participation in the ELUCIDATE trial. As will be recalled, more general issues about biomarker evaluation research were discussed with patient and public representatives in a consultation exercise reported in an earlier section (Chapter 9).

While fully recognising that these are two different types of activity, each generating very salient points of their own, a number of common themes can also be identified.

The patient interviews clearly indicate that adding ELF testing to an ongoing monitoring regime, as was done in the intervention arm of the ELUCIDATE trial, was acceptable to participants and fitted well with routine care in this group of at-risk patients. Reassurance of continuing low scores (below cut point) could be obtained by these means.

Individuals who were test/diagnosis positive were obviously in a different position. The term used in the trial protocol (and hence used by staff when communicating with patients) for people testing positive in the intervention arm was “diagnosis”. A test score above the specified cut point triggered the initiation of a cirrhosis management protocol, which included further tests. The interviewee who had experienced this care pathway was told by staff that the subsequent tests had not supported this “diagnosis”, i.e., that the ELF result had been a false positive, with upsetting as well as confusing consequences.

No quantitative argument is being made here. Clearly, interview numbers were small, and many factors will have influenced who was available and willing to be interviewed. However, almost two thirds of trial participants in the intervention arm were found to be test positive at some stage, and the interviewee’s experience does chime with concerns about false alarms expressed by members during the WS1 PPI consultation.

Clinicians’ desire not to miss preventable disease is very powerful, but the possibility of causing harm is real and must be considered. National screening programmes, for example,
have been criticised for ignoring negative effects and a more balanced presentation of the pros and cons is now advocated.

The pathway from the introduction of a new monitoring strategy through to patient benefit is clearly complicated, but the health economic approach (Chapter 8) lays out the fundamental trade off very clearly: does the gain in “utility” (Quality of Life multiplied by cost) achieved by successfully treating more cases outweigh the loss in utility incurred by unnecessarily treating more non-cases? Patients and family members can readily recognise the trade off being made here, and can relate it to their own circumstances and experience. Members of the PPI consultation group could also see that the numbers (e.g., test scores) on which decisions were based reflected professionals’ judgements and preferences as well as scientific “facts”.

Thinking first about judgements and preferences, members of the PPI consultation group could readily appreciate that a traditional clinical approach tends to focus on the benefits of successfully treating those patients who can benefit, and to have relatively little to say about “the price” – in every sense of the word – being paid for that. PPI members appreciated the application of a “precautionary principle” to detecting the recurrence or progression of disease – assuming of course that early information is preventative – but they also emphasised the possibilities for harm. It could be additionally pointed out here (echoing a point previously made in Chapter 5) that the case for paying more attention to negative consequences when designing monitoring strategies is even stronger than for screening because monitoring offers multiple opportunities for harm.

The established health economic argument (Chapter 8) here is that an inefficient use of resources will displace more efficient use of those same resources. This argument is more customarily thought of in terms of more and less effective treatments, but it also applies to more and less efficient use of follow-on testing resources. An additional point can also be made: if the eligibility criteria for enhanced monitoring are made too broad, then people at high risk of developing (or already having) the condition of interest will be on the same waiting lists for follow-on tests as people at much lower risk. PPI representatives familiar with an overburdened NHS, its resource constraints and waiting lists, had no difficulty appreciating this point and wanted a more transparent and considered approach to be adopted. Simply widening the definition of who is eligible for enhanced care could be seen as seldom
the best approach, even if the enhanced regime was acceptable and relatively benign from an individual patient’s perspective.

As for “scientific facts”, the existence of sizeable evidence gaps in the biomarker pipeline came as quite a surprise to members of the PPI consultation group. They were aware that very substantial sums continue to be spent worldwide on biomarker research and assumed that all components of the ACCE pipeline would already be of the highest methodological quality. When they realised that the quality of evidence sufficient to justify the use of a given test strategy in the NHS was not in fact uniformly high, they were rather shocked.

A number of other points with implications for patients are worth spelling out here, even though they were not discussed in detail with patients themselves. In the intervention arm of the trial, test positive patients were managed as though they had been diagnosed with cirrhosis. The first management change was to implement more intensive monitoring, some of which entailed quite unpleasant procedures, eg endoscopies. One of the interviewed patients commented (see above) that being monitored with a blood test like ELF was an attractive alternative to having a biopsy, but the alternatives are actually more complicated than that, as the trial results (Chapter 22) show. Many more endoscopies were conducted in the intervention arm, but there was no reduction in the biopsy rate. Inevitably, such management is also more costly.

Patient benefit cannot, of course, result from extra monitoring per se, but only from a timely response to information generated by the more frequent/intensive testing schedule. Process outcomes provide an early signal that changes to management are being triggered by test results, but it is important to maintain a distinction between process outcomes that reflect extra monitoring as distinct from process outcomes that reflect the initiation of treatment.

The clinical effectiveness of treatment at the particular stage of disease identified by the initial test (and any confirmatory tests) must next be evaluated, and here a robust randomised design is essential. Correctly identified early cases (as well of course as early identified false positives) may not be better off if early treatment does not lead to better outcomes than late treatment – and all will have spent more time under investigation with associated anxiety.
To repeat the crucial question asked earlier: does the gain in utility (QL times cost) achieved by successfully treating more cases outweigh the loss in utility incurred by unnecessarily treating more non-cases? A single RCT can answer this question in relation to a particular combination of measure, cut point, monitoring schedule, management protocols and patient population, but experience suggests that the choice of many of these study attributes is not evidence based. A research pipeline which operates in this way is inevitably inefficient, as it places no burden on investigators to justify many of their choices, still less to make some effort to optimise the combination selected. The implications for trial participants as well as for patients more generally are substantial.

Given the sums of money spent on biomarker research to date, and the lack of demonstrable patient benefit, some kind of change is clearly desirable. Patients and the public purse need all ACCE ingredients to be of the highest quality, not just those currently favoured by the research pipeline as it currently operates. This chapter concludes with an overview of the case for doing things better.

**Guidelines to improve research quality**

Guidelines for researchers about how best to conduct and report different kinds of biomarker studies have been available for some time. The REMARK guidelines, for example, offer reporting recommendations for tumor marker prognostic studies, and STARD does a similar job for diagnostic accuracy studies. More recently, the MONITOR group have proposed a four phase model for biomarker monitoring trials, and a position statement drawing together all the main good practice guidelines has been issued by the European Group on Tumor Markers (EGTM). 296, 807

In a welcome convergence, the four stages identified in the EGTM position statement, aimed at an oncology readership, are similar to the stages identified by the MONITOR group for trialists, and by the test evaluation working group of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM). 308 In their first three stages, all bear a close resemblance to the ACCE framework that was drawn upon for the work reported in Chapter 9, but the last stage in the tumour markers statement refers to regulatory approval, the last stage in the trial design proposal refers to audit and economic impact (including quality of
life, assessed using established methods), and the last stage in the clinical chemistry
document refers to “the impact of testing on the patient, the organisation and society.” In
2015, the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC)
took the emphasis on patients one step further in a report entitled “Current evidence and
future perspectives on the effective practice of patient-centered laboratory medicine”. The
report argued that laboratory medicine specialists needed to work with multidisciplinary
groups seeking to, “optimize clinical outcomes and patient experiences in an efficient and
cost-effective way.”

**Gaps in the evidence**

From a standard methodological perspective, the research pipeline on non-invasive liver
markers can be seen to have gaps in it, but in more heavily researched areas, the same is also
ture. Some gaps remain in territory which is well charted in other respects. In respect of a
basic analytical validity question, for example, a study showing that a prompt repeat of PSA
testing could reduce the number of unnecessary biopsies in men being screened for prostate
cancer was only published in 2015. Further down the pipeline, Bessen et al (2015) modelled the cost utility of different mammographic follow-up schedules, and showed they
could be tailored according to risk of recurrence. Assumptions about the effects on patients
of receiving reassuring test results may also not stand up to scrutiny. Rolfe and Burton (2013)
found little evidence of psychological benefit for patients in relation to diagnostic tests which
had essentially been ordered for reassurance purposes, although there may of course be
benefits for the patients’ doctors.

There are also areas of relatively uncharted research territory. Further targeted
methodological work could help to fill important evidence gaps, for example on the inter-
dependency of patient mix, cut points and schedules in maximising overall patient benefit.
The implications for both modelling and care of variation between patients in e.g. progression
rates, have been little studied in a monitoring context (as compared to the screening context),
and nor have the implications for research and service provision of omitting diagnostic
testing from a monitoring care pathway. Traditionally, diagnosis is seen as informing
treatment options, depending on the cause identified, and monitoring as informing the timing
of treatment, but these boundaries are blurring and no longer cover all possibilities. One
important piece of work arising out of the present study will be the drawing together of the models described in Section 1, and another will be further study of patient preferences for monitoring schedules with different properties and different implications for preventative behaviour and lifestyle change.

The work we presented to our PPI consultation group raises an important issue about the role of patient and public involvement in future ‘methodological’ studies. At the outset of this programme of work we were hesitant about trying to communicate ‘methodological’ knowledge to lay members and some colleagues took the view that it could not meaningfully be done. However, our meeting showed that the public have a more sophisticated palate, and a greater interest in this topic than researchers have traditionally given them credit for. Participants demonstrated a genuine interest in these issues and asked pertinent questions, but we had only scheduled to meet with the group once. With hindsight, a series of four to five meetings with the group to facilitate a more wide reaching discussion about the implications of the work and possible ways forward would have been more useful, both to them, and us.

Choosing wisely
In 2012 the American Board of Internal Medicine (ABIM) Foundation launched a campaign called “Choosing Wisely”, which has the goal “of advancing a national dialogue on avoiding wasteful or unnecessary medical tests, treatments and procedures”. In 2015, the Academy of Medical Royal Colleges publicized their initiative to bring the campaign to the UK. The emphasis in both campaigns is on avoiding unnecessary treatments, but unnecessary tests and assessments are also addressed. The first action point of the UK campaign reads, “Doctors should provide patients with resources that increase their understanding about potential harms of interventions and help them accept that doing nothing can often be the best approach.” The second says, “Patients should be encouraged to ask questions such as, “Do I really need this test or procedure? What are the risks? Are there simpler safer options? What happens if I do nothing?” High quality research evidence is clearly needed if clinicians are to be able to answer questions of this kind.
Personalised medicine, innovation and the funding of research

At the end of Chapter 9, the point was made that the acknowledged role for patient choice in treatment decisions also needs to be thought through in relation to testing decisions. For example, the choice of cut point for ELF – the case management threshold – was lowered in the Elucidate trial from 12.5 to 9.5, primarily on the basis of judgements about which risk categories were likely to carry weight with clinicians. By increasing the numbers of test positives in this way, fewer people with incipient cirrhosis will have been missed, but inevitably this will have been accompanied by an increase in the numbers of people switched to an intensive monitoring regime unnecessarily. It seems likely that many patients would arrive at a different trade-off here, not least because of the need to take probabilities of benefits and harms into account as well as their potential magnitude.

Bossuyt and Tajik (2015) draw a similar conclusion in their paper about the evaluation of biomarkers used to guide treatment decisions – a not dissimilar function in principle from the monitoring markers under scrutiny in this report.814 After describing studies in which cancer patients were asked about the size of treatment benefit they would need to expect for adjuvant chemotherapy to be worthwhile, these authors concluded, “for some, the required gain may be fairly large, while for others extending survival is extremely important and their threshold for accepting treatment is close to zero.” Bossuyt and Tajik went on to say, “This is definitely an area for personalised medicine: not in the abundant use of next generation sequencing, but in the recognition that personal values and trade-offs differ.” Although direct research evidence is not available, the patient perspectives described in this report make it likely that very similar conclusions could be drawn about the use of monitoring tests for disease progression.

As Bossuyt and Tajik note, in respect of test evaluation metrics, “The classical clinical performance measures, such as clinical sensitivity and specificity, can only be used in rare circumstances.” It can be argued however that research funding is much more likely to be available for studies of the sensitivity and specificity of innovative new markers than it is to fill gaps in our understanding of the real world performance of existing ones, and funding for trials – even quite speculative ones – is much more widely available than funding for well designed, descriptive longitudinal studies, or studies of patient perspectives on the quality and
role of tests used in their care. Although the need to fill gaps in the evidence base for using biomarkers in patient care is scientifically and professionally convincing, as long as the current incentive structure for researchers remains, little is likely to change and the potential for patient benefit is unlikely to be realised.
Chapter 24 - Programme Conclusions and the Framework for Biomarker Evaluation
General conclusions

In Chapters 9, 15 and 23 we have summarised and discussed the principal conclusions from each of the workstreams in Methodology (WS1), Clinical Translation (WS2) and the ELUCIDATE RCT (WS3).

Taken together they represent a substantial overview of the state-of-the-art and the challenges in the introduction of known biomarkers into clinical practice and healthcare systems. We have shared the results and the experience of this investigation with patients at the application, study design, study delivery, analysis and evaluation stages. In each workstream we have confirmed the importance of a rigorous and systematic approach to biomarker evaluation and made recommendations to help future work in this area. The incomplete and inadequate nature of some of the methodological approaches taken in this field was clearly identified in WS1; similar limitations of approaches in the laboratory and in sample preparation and study design in many cases in WS2. The challenges, both conceptual, logistic and organisational of delivering “end-to-end” clinical trials like ELUCIDATE to determine the place of new biomarkers in clinical practice were shown in WS3.

The investigators conclude that there are a number of important, generic and recurring themes in what we have learned in these studies:

- Multidisciplinary research teams are essential to establish individual and portfolios of biomarker evaluation projects. All of the parts of the pipeline must be critically evaluated to a high standard using standardised approaches if the field is to move forward adequately.

- The organisational and logistic challenges have to be addressed by well trained teams with adequate resources to operate to the standards that are required.

- Innovation in study design and research methodology is essential. “End-to-end” RCTs will remain the gold standard for evaluating the place of a new biomarker or panel of biomarkers in clinical practice. Such study designs must be carefully and comprehensively grounded in appropriate prior knowledge of the performance of the test and the evaluation of the test in appropriate large current clinical populations.
Statistical methods for the design of the trials and the calculation of power have been developed in the course of this programme. To be useful at the beginning of an RCT, they require access to high quality population based cohorts of well characterised patients with the disease under study, whose data reflect the current biological basis of disease (such as causative factors) and the current treatment environment.

- New study designs and logistic solutions have to be deployed to streamline the biomarker pipeline. The Methodology workstream highlights these issues for monitoring studies.

- For Clinical Translation (WS2), studies of analytical and clinical validity are crucial parts of the evaluation of tests but are beset by the challenge of having large numbers of good quality samples with annotated excellent clinical data which a representative of the key clinical populations, current and have adequate follow up to answer long term questions. They have to be carefully curated but managed, used, sustained and understood by specialised multidisciplinary research and innovation teams. WS2 demonstrates this very clearly. Broadly, it took us 5 years to design and deliver the samples and clinical datasets for renal disease; it took us 5 months to review all of the literature on available biomarkers in RCC, find and validate the assays and report the findings in this Report. Future strategies will require access to pre-existing sets of data and samples which we will curate and sustain and use as a resource for academic and commercial collaborators and through the model of the NIHR DEC. This should mean more “5 month” study turnaround times. Even this will be at risk if our specialised multidisciplinary team, nationwide in its clinical contacts, lost its energy or focus. Establishing and sustaining this approach in other topics will be challenging. Nevertheless, we have shown that with the appropriate infrastructure and planning, the analytic and clinical validity of tests may be evaluated promptly for the purposes for which they are intended – allowing robust initial evaluation of tests which can then be evaluated for their impact on healthcare outcomes using RCTs or other strategies.

- The ELUCIDATE RCT illustrates what can be achieved through the commitment and engagement of investigators in the NHS. It was slow in set up and in initial
recruitment. More Centres were opened; existing Centres often responded and recruitment flourished to bring in a large proportion of patients in the last months, even under threat of closure. To some extent the lessons are therefore conventional – work harder and faster; open more centres; drive the process by motivating the centres – and large RCTs get delivered.

- We think, however, that the learning points from ELUCIDATE are less conventional and the response need to be more radical. The “end-to-end” nature of this RCT, from initial use of a biomarker through the consequent change in clinical behaviour to yield the desired changes in the process of care, through to long term impacts on survival and other major clinical events – although desirable, is going to be rarely deliverable without radical changes in how we all work. ELUCIDATE suffered from “conventional” RCT challenges of scale, recruitment and set up and the responses are listed above. It also suffered from changes which occurred in the disease population (more alcohol related liver disease), changes in the diagnostic options (Fibroscan in more centres) and changes in the therapeutic options (antiviral drugs for hepatitis) which happened between the studies of cohorts which resulted in its validation and during the conduct of the trial. They will continue during the planned long follow up using NHS Informatics in ONS and HES data. So the value of the trial to the NHS is restricted by its scale, duration and, of course, cost.

- Our solution for ELUCIDATE was rapid recruitment in the latter phases; innovative modelling to support design and power calculations; analysis of process of care endpoints which reflect the impact of the biomarker monitoring on clinical behaviour and practice, and the longer follow up using health informatics, we believe, will often be appropriate. This approach needs to be prospectively developed for future trials. Modern health informatics can provide large, current, relevant clinical datasets to complete the study design and early economic modelling quickly and in the currently appropriate clinical populations which reasonably stable epidemiological, diagnostic and therapeutic environments. These will have to come from patient electronic records through to carefully developed technical, confidentiality and governance routes that are in the process of being put in place. The RCTs may then focus on changing clinical behaviour and the process of care. Long follow up will come from
Health Informatics (as with ELUCIDATE) but will use EPR based sources rather than the derived databases of ONS/HES origins.

- Patient engagement has been constructive in the course of the programme and should remain a central feature of this field, as with all fields of biomedical and health research. We drew extensively on patient input not only in evaluating the methodological findings of Workstream 1 but also in the design, delivery and evaluation of the preliminary results of Workstream 3. Patients were involved in the clinical translational workstream in study design and advised at all stages of the delivery of the cohorts and have commented on the analysis and reporting of the three WS.

There is increasing concern about the volume of biomedical research undertaken that is irreproducible, with systematic reviews estimating that between 50-89% of pre-clinical research contains one or more errors, flaws, inadequacies, or omissions that prevent the replication of results. With growing financial pressures on researchers in the UK and elsewhere, research funders and publishers have a responsibility to ensure that the time and public money spent on research are spent wisely. This programme has highlighted the need for appropriate validation and verification of biomarker assays and diagnostic tests prior to conducting research studies (see also Chapter 13). The level of validation/verification should be appropriate to the stage of test development and its intended clinical use. If a test is used to inform clinical decision making within a trial, or provides an end point or outcome measure within an interventional trial, it is essential that such tests and/or biomarker assays are validated fully in a clinical laboratory as early as possible during its development. Where a test or assay is to be used purely in an observational “research use only” context then less rigorous assay validation, as described in Chapter 13, may be sufficient. However, appropriate assay validation should be an essential requirement for funding and publication of biomarker studies.

The programme has identified new methodological approaches, new biomarkers which justify evaluation of clinical utility in kidney disease and shown that the deployment of the ELF test will alter clinical practice in ways which are likely to be associated with improved
outcomes. It is disappointing that we do not have the final evidence of the impact of the ELF test on the important outcomes of liver disease including cancer, haemorrhage and survival.

The experience with the Applied Biomarker Programme has contributed to two broad aspects of a Framework for the introduction of biomarkers into healthcare systems.

1) A framework for introducing biomarkers into clinical practice through the NIHR Diagnostic Evidence Cooperative (DEC).

2) A framework for the design and conduct of clinical trials for biomarker evaluation based on innovative study design and modern health informatics, and early engagement with health economics discussed above.

The samples and clinical data accumulated in the programme will be an important asset for future studies both for the programme investigators and collaborating centres who have delivered the samples and the data and guided the research and for new collaborations for academia and industry through the NIHR DEC and other routes.
The NIHR Diagnostic Evidence Cooperative at Leeds

The In Vitro Diagnostic (IVD) industry is the 2nd largest UK medical technology sector by employment and 5th by turnover, so is of fundamental importance to the UK economy and a huge area for growth with 17% increase between 2010-2011. The Leeds NIHR DEC was designed based substantially on the experience in this programme and take four complementary strategic approaches to enhancing the evidence base for IVDs. We will:

![Diagram of diagnostic evaluation pipeline and development areas]

**Figure 76 Leeds NIHR DEC diagnostic evaluation pipeline and development areas**
a) Deploy and refine methods in IVD study designs, health economics and health informatics to improve and speed up the way IVDs can be evaluated for NHS use drawing on our experience in this Programme.

b) Sustain and strengthen our working networks of cooperating NHS sites to deliver studies effectively with their patients and samples. Drawing on the experience in WS2, we can sustain the capability and capacity to more rapidly evaluate new biomarkers in renal and liver diseases. This approach has been extended to musculoskeletal diseases and oncology/haematology.

c) Invite, select and prioritise specific IVD candidates in our clinical areas, from our own work partners and interested parties and help them develop and deliver appropriate studies and evidence.

d) Create a strong stakeholder engagement group to work with our teams, patients and our academic and commercial partners, to shape our strategies, research programmes and projects and identify new opportunities together.

Investigators of this Applied Programme are playing a major role in the national NIHR DEC developments. Jon Deeks leads on methodology for the four DECs (Imperial, Oxford, Newcastle and Leeds). He also leads on methodology for the Newcastle DEC. Peter Selby, Mike Messenger and Steph Roberts organise and lead the NIHR DEC Leeds and Applied Programme investigators (William Rosenberg, Cathie Sturgeon, Andrew Lewington, Naveen Vasudev, Claire Hulme, Carys Lippiatt, Chris McCabe, Roz Banks, Doug Altman, Walter Gregory) have leading roles.

IVD evaluation methodology research

Our NIHR DEC Leeds is promoting a coherent philosophy based on continuous decision modelling through all phases of development. A central model for each project will draw on real-time NHS data-mining to define the clinical pathway and key clinical decision points and their relationship with clinical outcomes and costs. Based on recent successful examples using these methods, time to adoption can be accelerated and research design efficiency can be promoted. Finding solutions to the challenges of evaluating diagnostics will be an
important part of the DEC mission. Such challenges include the rapidly evolving nature of multiple competing technologies, real-world characterisation of diagnostic test properties and their impact on clinical and economic outcomes and optimisation of case definition thresholds. Members and collaborators of the NIHR DEC Leeds have been at the centre of recent methodology innovations aimed at addressing some of these challenges. The expertise is divided into three themes:

i. Study Design and Conduct. The gold standard for demonstrating clinical utility is the randomised controlled trial. In response to the need for faster time to market, coupled with the complexities of evaluating diagnostic tests, new trial methods are required. Drawing on experience from our NIHR Biomarkers programme and the ELUCIDATE trial we have developed concepts for RCTs to evaluate changes in care processes and modelling strategies to test long term patient and NHS centred outcomes, after our experience with recruiting almost 900 patients to such a trial. The DEC strategy will build on this to streamline randomised designs, making more use of early field trials, surrogates, modelling, the use of routine linked data and integrated economic data collection.

ii. Health Economics. The gold standard for demonstrating cost-effectiveness is a model-based economic evaluation. Modelling will commence at the very start of IVD development under guidance from the DEC during selection and prioritisation. By introducing a model early, it is possible to characterise the potential impact of an IVD on the clinical pathway, clinical decision points and expected clinical end economic outcomes. The optimal case definition threshold for tests can be proposed for cost-effectiveness in addition to clinical validity alone. Models will be maintained as IVD development progresses, populated by meta-analysis of evidence generated both within and external to the DEC. Probabilistic modelling will be used to characterise areas of uncertainty in the evolving evidence for value between each phase of development; thus enabling iterative research design efficiency. Expected cost-effectiveness can be established as well as commercial headroom for the manufacturer. The trade-off between investment in a large RCT versus alternative cheaper or quicker study designs can be described through the modelling process including the use of Bayesian decision modelling and value of information analysis.
iii. Health Informatics. Central to improving the efficiency of IVD development will be early testing in real NHS settings. For this to be achievable, outcomes monitoring through pre-established clinical data collection is necessary. In order to demonstrate the potential impact of a new IVD, the standard care pathway with real-world outcomes needs first to be defined. A framework has been developed by Leeds researchers, which allows the pathway (including event probabilities and patient characteristics) to be defined directly from observed clinical events held within clinical databases (The Leeds Patient Pathway Manager) using data-mining techniques. This allows a central decision model to be populated at patient-level with observed outcomes. Linkage with NHS finance and resource usage databases such as HES, PLICS and primary care databases allows a full economic model to be constructed.

Applied examples and a set of modelling tools have been developed in Leeds where standard-care and IVD-specific clinical data feeds directly into a decision-analytic model to produce estimates of longer-term health impact and cost-effectiveness.

**Qualify Analytical Validity**

Independent verification of the technical performance of an IVD can be conducted by the DEC IVD Validation Group, including assessments of analytical sensitivity, specificity, precision, parallelism, recovery, selectivity, limit of quantitation (LOQ) and vulnerability to interferences.

**Qualify Clinical Validity**

The IVD clinical performance (e.g., sensitivity, specificity, predictive values) will be reviewed and can be verified by the IVD Validation Group using samples from our established networks of Co-Operating NHS sites and biobanks, established as part of the NIHR Biomedical Research Unit, ELUCIDATE trial and NIHR Biomarker Applied Programme. The DEC will identify funding (commercial consortia and/or public) to replenish biospecimen resources and support multi-centre biobanking programmes alongside registries,
providing low-cost high value prospective samples and data for rapid IVD evidence
generation in priority clinical areas.

Evaluation of Clinical Utility and Cost Effectiveness

Decisions about the research design priorities will be decided by the DEC Methodology
Group, following an update of the pathway model developed in stage 1, to include: data
gathered in stages 2 and 3; pathway outcomes from linked NHS datasets; and re-estimation of
uncertainties and value of information. Where further evidence is required the Methodology
Group will make recommendations on research designs. Where necessary the DEC will then
utilise the expertise of the Leeds CTRU for running trials of IVDs, incorporating the
enhanced methodologies developed in DEC research programmes.

The Interactions between Workstreams

We have given clear examples where the interaction between workstreams has been
synergistic. The methodology work provided a literature based and evidence based
framework for the development of Workstream 2 and Workstream 3. For instance the
simulation modelling made possible a radical re-design of the ELUCIDATE RCT and
provided a robust approach to power calculations which underpin the completion of the trial
and its extensions. The methodology workstream also emphasised the critical importance of
the determination of Analytic and Clinical Validity of tests in a meaningful way which can
underpin evaluations of clinical utility and cost effectiveness. This resulted in the re-
evaluation of the ELF test as a laboratory test reported in Chapter 17 and the rigorous re-
appraisals of the assays used throughout Workstream 2 including those prioritised as a
consequence of the literature reviews. The experience of developing the RCT informed the
development of the cohorts for the Workstream 2 renal cancer and renal transplantation work.
Our ability to work with centres, promote trial set up, open new centres and motivate our
collaborators, allowed delivery of the recruitment to target which they substantially
succeeded in doing. However, it is important that we acknowledge again that the delay in
delivering the ELUCIDATE trial meant that large bodies of information broken down by trial
arm were not available for our methodology research.
Incorporating large clinical cohorts and randomised controlled trials into an Applied Programme as the attraction of the synergies of the kind described in the above paragraph. However, delays in setting up in recruiting into those cohorts do have potential to undermine the delivery of the programme as a whole and limit the synergy between the workstreams. There is no absolute ideal model, but we would urge caution in the integration of large RCTs into integrated programmes of work where other workstreams are dependent on the timely completion of those trials.

**Future Methodology Research**

To validate a method one must provide objective evidence that it fulfills the evidence requirements for a specific intended use and is “fit for purpose”. However, defining the requirements remains a challenge, even after several decades of intensive efforts by members of the laboratory medicine community. In 1999 a landmark conference in Stockholm agreed a hierarchical structure for setting analytical performance goals. A recent 2014 conference in Milan revised and refined these, suggesting three approaches based on:

1. The effect of analytical performance on clinical outcome (either directly or indirectly)
2. Components of biological variation of the measurand, or
3. “State of the art”

However, whilst the Milan consensus provides a useful framework, it did not consider the interconnectivity of the approaches and the possibility that there may be a unified strategy for combining them. Furthermore, very few examples have been reported of analytical performance goal based upon the effect on clinical outcomes, probably due to the complexity and cost of these approaches. Whether it is pragmatic and appropriate to recommend that manufacturers, clinical scientists and researchers invest significant time and resource in striving to set analytical performance goals against the highest model, clinical outcomes, when a simpler solution may be “fit for purpose” is therefore debatable.

The current statement also does not consider the role of cost-effectiveness, a major component of UK Healthcare decision-making. On an almost monthly basis the NICE Diagnostic Assessment Programme produces Health Technology Assessments and economic models that should in principle be adaptable to evaluating the impact of analytical
performance goals on clinical outcomes and cost-effectiveness. Health economists and decision modelers have a wealth of methodological expertise that may be highly useful in developing analytical performance goals. Similarly a better understanding of analytical and pre-analytical factors may prove useful in Health Technology Assessments. Further methodological work in this area should be pursued.

Considering the demonstrable impact of pre-analytical and analytical factors on biomarker measurements, better utilization of metrological concepts (e.g. Measurement Uncertainty) by medical statisticians may improve diagnostic test study and trial design, particularly in terms of power calculations, optimal cut-off thresholds and monitoring intervals. Furthermore, the specific impact of random (imprecision) and systematic (bias and specificity) factors on trial design should be considered. A recent simulation study of HBA1c measurement convincingly demonstrated that when using a fixed guideline based cut-point, varying the bias and imprecision had very different effects on diagnostic accuracy, with bias having the most severe consequences. 

**Limitations**

A programme with a responsibility for developing a framework built on excellent research innovative methodology, excellent clinical biochemistry and appraisal of analytic and clinical validity of tests and also deliver a substantial RCT of a monitoring regimen, was likely to be challenging. Our literature reviews demonstrate the scale of the challenge. There were likely to be disappointments and limitations. We feel it is appropriate to highlight four of these:

- The delays in set up and recruitment into the study cohorts of Workstream 2 and the ELUCIDATE RCT had a negative impact on our ability to generate large and complete datasets with adequate long follow up to answer questions conclusively. Further follow up is required and will be done.

- The low to moderate compliance in the ELUCIDATE RCT may reduce the effect size of any benefit of ELF, and may possibly render the trial underpowered, though a different package of investigations was clearly delivered between the arms following the diagnosis of cirrhosis.
• The absence of the long follow up data precludes conclusions about the true value of the ELF test to alter the serious consequences of liver disease.

• The absence of long follow up data at this stage has limited opportunities for more methodological research.

• The discovery pipeline in renal and liver disease has not yet provided exciting new molecular biomarkers to evaluate in our cohorts.

Final Comments

In this programme, we have summarised what is known of monitoring tests using prominent examples and rigorous methodological appraisal and systematic overview; developed new approaches to evaluating the analytic and clinical validity of new biomarkers, particularly exploring the requirements for underpinning infrastructures; considering the products of modern proteomics; and delivered an exemplar RCT which has demonstrated changes in the process of care for a vitally important area and growing area of morbidity and mortality in the United Kingdom. Incomplete follow up as a consequence of delays in set up and recruitment have limited our conclusions and these can be remedied by health informatic strategies which we have outlined, planned and will robustly deliver.

The work of this programme was pivotal in our proposal for an NIHR Diagnostic Evidence Cooperative and the methodological, clinical biochemistry and clinical trials conclusions from the Programme will continue to underpin the delivery of diagnostic evidence and the development of novel methods for delivering it in a more timely and cost effective way for many years to come.
Acknowledgements

We are grateful to the contributions of the patients who donated the samples, the clinical and nursing staff who have contributed to the various cohorts of the Programme. In addition, the authors would like to acknowledge the contribution of:

- The members of LIVErNORTH for their contribution towards Patient and Public Involvement study
- Ms Claire Davies, Senior Trial Manager, for help and support at the outset of the trial
- Dr Sue Pavitt for assisting with aspects of translational medicine at the outset of the trial
- Dr Barbara Potrata, Research Fellow, for design and set up at the outset of the ELUCIDATE patient exit study
- ELUCIDATE Data Monitoring Ethics Committee: Dr Shahid Khan, Prof Elaine McColl, Prof Peter Mills, Prof Paula Williamson
- Independent Trial Steering Committee members: Mrs Joan Bedlington (PPI), Prof Simon Dixon (Health Economist, Dr Jonathon Fallowfield (Hepatologist), Mrs Tilly Hale (PPI), Prof James Neuberger (Hepatologist), Dr Christine Patch (Clinical Geneticist), Dr Andrew Roddam (Statistician)
- Philip Akude and Mike Paulden for their contributions to the health economics study with Professor Chris McCabe
- Business Managers & Finance: Sandra Holliday, Donna Johnstone, Wendy Kennedy, Simon Revesai, Tom St. David-Smith
- Sample banking and data management team: Joanne Brown, Andrew Bernard, Narinder Gahir, Damien Hindmarch, Leah Khazin, Norma Lister, Riitta Partanen, Lauren Tate, Emma Tidswell, Sharon Jackson, Gunnar Piho
- The Chair, Professor Phil Kalra and committee members of the CRN Speciality group for Renal Diseases and fellow committee members
- The Chair Dr Stephen Ryder and committee members of the CRN Speciality Group for Hepatology
- All the clinical laboratories that participated in the inter-laboratory study, to Dr Allan Thompson and Mr James Dowd at Siemens Diagnostics and to Helena Baker and colleagues at the Blood Sciences Department at Leeds Teaching Hospitals. Also for
Helena Baker’s involvement in the VEGF stability study which will be extended and published.

- Dr Rebecca Kift and Miss Sophie Hepburn for their work on the analytical performance assessment of commercially available NGAL tests

Author contributions

All authors made a substantial contribution to the concept and design of one or more of the several studies or the acquisition of data, data analysis, or interpretation of data and drafted the manuscript or revised it critically for important intellectual content. All senior programme leaders contributed to all chapters.

Professor Peter J Selby (Programme Director and Clinical Lead Renal Cancer) – Led on the management of the programme with the Programme Manager, Co-Chief Investigator of the ELUCIDATE Trial and co-ordinated the completion of this report.

Professor Rosamonde E Banks (Scientific Lead on Renal Disease Biomarkers) – Senior Programme Leader, contributed to the programme application. Led the Clinical Translation Workstream (WS2) and the preparation of the WS2 Chapters.

Professor Walter Gregory (Senior Lead, ELUCIDATE Director and Principal Statistician) – Senior Programme Leader, contributed to the programme application, led the ELUCIDATE RCT Workstream (WS3) and Chief Statistician on the cohort and RCT design and analysis.

Professor Jenny Hewison (Deputy Programme Director and Lead on Health Services Research) – Senior Programme Leader, contributed to programme application. Led the Methodology Workstream (WS1) and preparation of the WS1 Chapters of this report.

Professor William Rosenberg (Clinical Lead on Hepatology) – Senior Programme Leader and Chief Investigator of the ELUCIDATE Trial. Contributed to the programme application and drafting of this report.
**Professor Douglas G Altman** (Strategic Lead Methodology) – Programme co-applicant, contributed to the programme application and work on evidence based monitoring and impact. Contributed to Chapters 2-6 and 18.

**Professor Jonathan J Deeks** (Strategic Lead Methodology) – Programme co-applicant and senior lead on the monitoring and simulation aspects of the methodological workstream and advised on all aspects of the programme methodology. Contributed to the programme application and drafting of this report (Chapters 2-7).

**Professor Chris McCabe** (Senior Lead on Health Economics) – Programme co-applicant, contributed to the programme application and drafting of this report, including Chapter 8. Provided health economic evaluation and contributed to the methodological considerations in the optimization of monitoring biomarkers of the programme.

**Dr Julie Parkes** (Senior Lead and Senior Lecturer in Public Health) – Programme co-applicant, contributed to the programme application and drafting of this report including Chapters 4 and 22. Contributed to the design of the randomised controlled trial and designated lead for the follow up of patients by NHS Health and Social Care Information centre.

**Dr Catharine Sturgeon** (Strategic Lead on Clinical Biochemistry) – Programme co-applicant contributed to the programme application and drafting of this report including Chapters 13 and 17. Advised on Clinical Biochemistry work and contributed to the verification of the ELF test analytical performance.

**Dr Douglas Thompson** (Operational Lead on Clinical Biochemistry) – Programme co-applicant contributed to the programme application and Clinical Biochemistry work. Contributed to the exploration of verification, validation and pre-analytical variables of biomarkers within the clinical translation workstream (Chapters 10, 13-15 and 17).

**Dr Maureen Twiddy** (Senior Research Fellow) - co-ordinated the patient and public evaluation of the methodology workstream and ELUCIDATE trial and contributed to the drafting of Chapters 9 and 23.
Dr Janine Bestall (Senior Research Fellow) - co-ordinated the patient and public evaluation of the methodology workstream and ELUCIDATE trial and contributed to the drafting of Chapters 9 and 23.

Mrs Joan Bedlington (Patient and Public Contributor) - provided advice and guidance during the ELUCIDATE trail and assisted with patient perspectives study and contributed to the drafting of Chapters 9 and 23.

Mrs Tilly Hale (Patient and Public Contributor) - provided advice and guidance during the ELUCIDATE trial and assisted with patient perspectives study. Sadly Mrs Hale died in December 2015 and contributed to the drafting of Chapters 9 and 23.

Dr Jacqueline Dinnes (Senior Researcher, Methodology) – Theme operational lead. Contributed to the methodology workstream and led on the review of methods for monitoring and examination of case studies. Led on Chapters 2, 3, 4 and 6, contributed to Chapter 5 and the drafting of this report.

Dr Marc Jones (Trial Co-Ordinator) Theme operational lead. Provided overall co-ordination and management of the recruitment, compliance and follow up of ELUCIDATE trial and biomarker cohorts. Contributed to the drafting of this report including Chapters 11, 16, 18 and 20.

Dr Andrew Lewington (Consultant Renal Physician and Honorary Associate Professor) – Clinical operational lead and Chief investigator of the Validation of biomarkers in the diagnosis of early kidney transplant complications and the prediction of chronic kidney transplant dysfunction and contributed to the review and prioritisation of circulating biomarkers in renal transplantation cohort. Contributed content of Chapters 10, 11, 12 and 15 and to the overall drafting of this report.

Dr Michael P Messenger (Principle Healthcare Scientist) – Theme operational lead. Led the set up and co-ordination of centres for the prospective observational cohorts within the Clinical Translation Workstream and contributed to the verification and validation of biomarker assays and the pre-analytical performance of the ELF test. Led on Chapters 11 and 17 and contributed to content of Chapters 12-15 and the overall drafting of the report.
Miss Vicky Napp (Operations Director) – Theme operational lead, provided overall direction and management of the ELUCIDATE Trial and contributed to Chapters 1, 16, 18, 20 and 22 and the overall drafting of this report.

Miss Alice Sitch (Lecturer in Medical Statistics) – Theme operational lead, provided simulation modelling of patient benefit, led on Chapters 5 and 7, contributed to Chapters 4 and 6 and the overall drafting of this report.

Dr Sudeep Tanwar (ELUCIDATE Co-Chief Investigator) – Clinical operational lead and Co-chief Investigator of the ELUCIDATE Trial. Directed the delivery of the latter stages of the RCT and contributed to Chapters 20 and 21 and the overall drafting of this report.

Dr Naveen S Vasudev (Clinical Associate Professor and Honorary Consultant in Medical Oncology) – Clinical operational lead and co-chief investigator of the validation/qualification of prognostic and monitoring biomarkers in RCC study. Contributed to Chapters 10-12, 14 and 15 and the overall drafting of this report.

Dr Paul Baxter (Associate Professor in Biostatistics) – Co-investigator, provided health economic evaluation of the ELUCIDATE Trial and contributed to the methodological considerations in the optimization of monitoring biomarkers. Contributed to Chapter 8 and the overall drafting of this report.

Dr Sue Bell (Senior Trial Manager) – Co-ordinated and managed the ELUCIDATE Trial, led on Chapter 16, contributed to Chapters 18, 20 and 22 and the overall drafting of this report.

Dr David A Cairns (Biostatistician) - provided statistical analysis within the clinical translation workstream and contributed to Chapter 14 and the drafting of this report.

Ms Nicola Calder (Senior Research Officer) - Supported the co-ordination of the biomarker cohorts and contributed to the evaluation of promising renal cancer biomarker candidates. Contributed to the content of Chapters 14 and 17 and overall drafting of the report.
Mr Neil Corrigan (Senior Medical Statistician) - provided statistical support to the ELUDICATE trial and contributed to Chapter 18 and the drafting of the report.

Dr Francesco Del Galdo (Senior Lecturer) - contributed to the verification of ELF test analytical performance in Chapter 17 and to the drafting of this report.

Mr Peter Heudtlass (Medical Statistician) - provided statistical support to the ELUDICATE trial. Contributed to Chapters 20 and 21 and the drafting of this report.

Dr Nick Hornigold (Postdoctoral Researcher) – designed and conducted immune assay work on biomarkers which is reported in chapter 14.

Professor Claire Hulme (Professor of Health Economics) – provided input to the health economic analysis of the ELUCIDATE trial, Chapter 21 and contributed to the drafting of this report.

Mrs Michelle Hutchinson (Biostatistician) - provided statistical analysis within the clinical translation workstream and contributed to Chapters 13 and 14 and the drafting of this report.

Dr Carys Lippiatt (Clinical Lead, Pathology R&D) – provided Clinical Biochemistry input and contributed to the drafting of Chapters 10, 13, 14 and 15 of this report.

Mr Tobias Livingstone (Senior Research Officer) - assisted with exploring the verification, validation and pre-analytical variables of biomarker assays. Contributed to Chapter 13 and the drafting of this report.

Dr Roberta Longo (Research Fellow, Health Economics) - provided health economic evaluation of the ELUCIDATE Trial (Chapter 21) and led on the methodological considerations in the optimization of monitoring biomarkers (Chapter 8) and contributed to the drafting of this report.

Mr Matthew Potton (Data Management Assistant) - provided data management for the ELUCIDATE trial and cohorts within the clinical translation workstream. Contributed to the drafting of this report.
Mrs Stephanie Roberts (Programme Manager) - co-ordinated the programme, including the programme delivery, funding and drafted important parts of the manuscript. Assisted the co-ordination of the ELF analytical validation study (Chapter 17).

Dr Sheryl Sim (Consultant Oncologist) – Input into the Clinical Translation Workstream and the accrual of the renal cancer cohort. Contributed to the drafting of Chapters 11 and 14 of the report.

Dr Sebastian Trainor (Clinical Research Fellow) - Input into the Clinical Translation Workstream and the accrual of the renal cancer cohort. Contributed to the drafting of Chapter 14.

Dr Matthew Welberry Smith (Renal Consultant) - Input into the Clinical Translation Workstream and the accrual of the renal transplant cohort. Discovery of the ACY-1 biomarker. Contributed to the drafting of Chapters 10, 11, 12 and 15.

Professor James Neuberger (Professor of Hepatology) - Chair of the ELUCIDATE Trial Steering Committee and advised on the planning and delivery of the hepatology aspects of the programme, design of the ELUCIDATE analysis and advised on the drafting of the ELUCIDATE reports.

Dr Douglas Thorburn (Consultant Hepatologist) – Top recruiting Hospital PIs and clinical team. Principle Investigator at Royal Free London NHS Foundation Trust for ELUCIDATE. Contributed to delivery of trial and drafting of the report.

Dr Paul Richardson (Consultant Hepatologist) - Top recruiting Hospital PIs and clinical team. Principle Investigator at The Royal Liverpool and Broadgreen University Hospitals NHS Trust for ELUCIDATE. Contributed to delivery of trial and drafting of the report.

Dr John Christie (Consultant Gastroenterologist and Hepatologist) - Top recruiting Hospital PIs and clinical team. Principle Investigator at Royal Devon and Exeter NHS Foundation Trust for ELUCIDATE. Contributed to delivery of trial and drafting of the report.
**Professor Neil Sheerin** (Professor of Nephrology) – Top recruiting Hospital PIs and clinical team. Principle Investigator at Newcastle Upon Tyne Hospitals NHS Foundation Trust for the Renal Transplant Cohort. Contributed to delivery of the cohort and drafting of the report.

**Dr William McKane** (Consultant Nephrologist) - Top recruiting Hospital PIs and clinical team. Principle Investigator at Sheffield Teaching Hospitals NHS Foundation Trust for the Renal Transplant Cohort. Contributed to delivery of the cohort and drafting of the report.

**Mr Paul Gibbs** (Vascular and Transplant Surgeon) - Top recruiting Hospital PIs and clinical team. Principle Investigator at Portsmouth Hospitals NHS Trust for the Renal Transplant Cohort. Contributed to delivery of the cohort and drafting of the report.

**Ms Anusha Edwards** (Kidney Transplant Surgeon) - Top recruiting Hospital PIs and clinical team. Principle Investigator at North Bristol NHS Trust for the Renal Transplant Cohort. Contributed to delivery of the cohort and drafting of the report.

**Professor Naeem Soomro** (Consultant Urologist) - Top recruiting Hospital PIs and clinical team. Principle Investigator at Newcastle Upon Tyne Hospitals NHS Foundation Trust for the Renal Cancer Cohort. Contributed to delivery of the cohort and drafting of the report.

**Mr Adebanji Adeyolu** (Consultant Urological Surgeon) - Top recruiting Hospital PIs and clinical team. Principle Investigator at Stockport NHS Foundation Trust for the Renal Cancer Cohort. Contributed to the delivery of the cohort and drafting of the report.

**Mr Grant D Stewart** (Clinical Senior Lecturer and Honorary Consultant in Urological Surgery) - Top recruiting Hospital PIs and clinical team. Principle Investigator at NHS Lothian for the Renal Cancer Cohort. Contributed to the delivery of the cohort and drafting of the report.

**Mr David Hrouda** (Consultant Urologist) - Top recruiting Hospital PIs and clinical team. Principle Investigator at Charing Cross Hospital, Imperial College Healthcare NHS Trust for the Renal Cancer Cohort. Contributed to the delivery of the cohort and drafting of the report.
Data sharing

Data can be obtained by contacting the corresponding author.
References


179. Primrose JA. Randomised controlled trial to assess the cost-effectiveness of intensive versus no scheduled follow-up in patients who have undergone resection for colorectal cancer with curative intent (The FACS Trial) (Protocol). In: Health Technology Assessment; 2005.


183. Bell KJL, Irwig L, Craig JC, Macaskill P. Use of randomised trials to decide when to monitor response to new treatment. *BMJ* 2008;336:361-5. [http://dx.doi.org/10.1136/bmj.39476.623611.25](http://dx.doi.org/10.1136/bmj.39476.623611.25)


208. Sturgeon CM, Lai LC, Duffy MJ. Serum tumour markers: how to order and interpret them. *BMJ* 2009; 10.1136/bmj.b3527. [http://dx.doi.org/10.1136/bmj.b3527](http://dx.doi.org/10.1136/bmj.b3527)


213. Baker SG. Improving the biomarker pipeline to develop and evaluate cancer screening tests. *JNCI* 2009;101:1116-9. [http://dx.doi.org/10.1093/jnci/djp186](http://dx.doi.org/10.1093/jnci/djp186)


234. Smellie WSA. What is a significant difference between sequential laboratory results? J Clin Pathol 2008;61:419-25. [http://dx.doi.org/10.1136/jcp.2007.074715]


263. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Annals of internal medicine 2006;144:850-5.


270. Cytological surveillance compared with immediate referral for colposcopy in management of women with low grade cervical abnormalities: multicentre randomised controlled trial. BMJ 2009;339:b2546. http://dx.doi.org/10.1136/bmj.b2546


275. Kim DY. Prospective, randomized study of PIVKA-II and AFP measurement every 3 months compared to AFP every 6 months in surveillance program for early detection of hepatocellular carcinoma; 2007.
276. Lepage C. Post-operative monitoring of patients operated for stage II or III colorectal cancer with intent to cure phase III, multicentre prospective trial; 2009.
282. MRC. A framework for development and evaluation of RCTs for complex intervention to improve health; 2000.
283. Isrctn. ISRCTN Register. In.
286. Siemens. ELF Test. In.


330. Womack C, Mager SR. Human biological sample biobanking to support tissue biomarkers in pharmaceutical research and development. *Methods* 2014;70:3-11. doi:10.1016/j.ymeth.2014.01.014


336. Scudellari M. Biobank managers bemoan underuse of collected samples. *Nat Med* 2013;19:253. [http://dx.doi.org/10.1038/nm0313-253a](http://dx.doi.org/10.1038/nm0313-253a)


345. Cummings J, Raynaud F, Jones L, Sugar R, Dive C. Fit-for-purpose biomarker method validation for application in clinical trials of anticancer drugs. *Br J Cancer* 2010;103:1313-7. [http://dx.doi.org/10.1038/sj.bjc.6605910](http://dx.doi.org/10.1038/sj.bjc.6605910)


Jonasch E, Gao J, Rathmell WK. Renal cell carcinoma. BMJ 2014;349:g4797. http://dx.doi.org/10.1136/bmj.g4797


403. NHS blood and transplant factsheet: cost effectiveness of transplantation.


407. NHS blood and transplant activity report.


http://dx.doi.org/10.1136/jclinpath-2014-202683
http://dx.doi.org/10.1016/j.urology.2013.08.052
http://dx.doi.org/10.1111/bju.12767
http://dx.doi.org/10.1111/bju.12275
http://dx.doi.org/10.1245/s10434-015-4614-0
http://dx.doi.org/10.1016/j.eururo.2011.01.034
http://dx.doi.org/10.1155/2014/689795
http://dx.doi.org/10.1038/sj.bjc.6600050
http://dx.doi.org/10.1002/ijc.11000
http://dx.doi.org/10.1007/s00432-005-0019-y
http://dx.doi.org/10.1007/s10147-012-0412-x


[http://ar.iiarjournals.org/content/34/2/921.long](http://ar.iiarjournals.org/content/34/2/921.long)


https://www.jstage.jst.go.jp/article/jmi/55/1%2C2/55_1%2C2_106/_pdf


http://imr.sagepub.com/content/40/2/475.full.pdf


2010;43:4-25.  http://dx.doi.org/10.1016/j.clinbiochem.2009.10.001


10.1186/1471-2407-6-207


Hollmen ME, Kyllonen LE, Inkinen KA, Lalla ML, Salmela KT. Urine neutrophil gelatinase-associated lipocalin is a marker of graft recovery after kidney transplantation. *Kidney Int* 2011;79:89-98. [http://dx.doi.org/10.1038/ki.2010.351](http://dx.doi.org/10.1038/ki.2010.351)


GRAMBSCH PM, THERNEAU TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994;81:515-26. [http://dx.doi.org/10.1093/biomet/81.3.515](http://dx.doi.org/10.1093/biomet/81.3.515)


760. Uzzo RG. Renal masses—to treat or not to treat? If that is the question are contemporary biomarkers the answer? J Urol 2008;180:433-4. http://dx.doi.org/10.1016/j.juro.2008.04.124


657


The Association for Clinical Biochemistry and Laboratory Medicine; 2005.


Appendices
Appendix 1 – Appendices to chapter 11
Details:

Author(s) of Study Site Operating Procedure: Tobias Wind (NIHR bioRTB)
Michael Messenger (NIHR bioRTB)
Carly Rivers (CTRU)

Comments:

The following Site Specific Procedures are for collection, processing, and distribution of samples for the Renal Cell Carcinoma (RCC) NBHR Biomarker Research Tissue Bank (bioRTB). The objective being to validate biomarkers for prognosis and longitudinal monitoring in patients with renal cell carcinoma.

Version Control:

<table>
<thead>
<tr>
<th>Version number</th>
<th>Edited by</th>
<th>Date edit completed</th>
<th>Details of editions made</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>MPM</td>
<td>09/08/11</td>
<td>Section A, para 3, changed &quot;3. 6&quot; to &quot;3-6&quot;; added F08 to Figure 3 and inserted a diamond symbol in Figures 1 and 2.</td>
</tr>
<tr>
<td>3</td>
<td>MPM</td>
<td>20/07/12</td>
<td>Changed process for Forms 04 and 05 to send original and keep a copy. Pg10 corrected text to say &quot;centrifuge both the serum and plasma samples together at room temperature for 10 minutes at 2000 x g (approximately 3000rpm)&quot;. Pg 10 removed statement about recording time serum transferred as only 1 box for both serum and plasma on Form 04.</td>
</tr>
<tr>
<td>4</td>
<td>MPM</td>
<td>23/05/13</td>
<td>Updated Section A &amp; Figure 2 to clarify that a final sample is collected on relapse. Confirmed sample processing times in section 2.2</td>
</tr>
</tbody>
</table>

Contents

Details: .......................................................................................................................... 1
Comments: ....................................................................................................................... 1
Version Control: .............................................................................................................. 1
Contents ............................................................................................................................. 1
Section A Introduction ..................................................................................................... 2
Section B Trial Sample Handling .................................................................................... 6
  1. TISSUE SAMPLES .................................................................................................... 6
  1.1 REQUESTING ....................................................................................................... 6
  1.2 TISSUE SAMPLE PROCESSING ............................................................................ 6
  1.3 COLLECTION OF FORM 05 .................................................................................. 6
  2. BLOOD & URINE SAMPLES .................................................................................. 7
  2.1 COLLECTION PROCEDURE ................................................................................ 8
  2.2 PROCESSING PROCEDURE ................................................................................ 9
  2.3 ON COMPLETION OF PROCESSING .................................................................. 10
  3. SHIPMENT PROCEDURE ....................................................................................... 12
  4. QUERIES .............................................................................................................. 12

T80568/0016_V1.0_UK8056
Section A  Introduction

This Study Site Operating Procedure (SSOP) is applicable to the Principal Investigator, Research Nurse, and any other member of staff at research sites who have responsibilities within the Evaluation of Biomarkers for Prognosis of Renal Cell Carcinoma study for the collection and processing of samples for the Leeds NiHR Biomarker Research Tissue Bank (bRTE).

The objective of the study is to validate/qualify prognostic and longitudinal monitoring markers of RCC using prospectively collected high quality clinical samples from multiple centres. Blood, urine and tissue samples will be requested from eligible patients attending participating centres.

Figures 1-3 illustrate the patient/volunteer pathways, their associated sampling regimes and the forms requiring completion at each stage. During the initial 18 months of the study 600 newly diagnosed RCC patients (all stages and histological types) will be recruited onto the cross-sectional arm of the study. RCC patients in the cross-sectional arm are only required to provide a single blood and urine sample at registration and an FFPE tumour tissue block if undergoing nephrectomy (see Figure 1). In the first year, an additional 200 newly diagnosed RCC patients undergoing nephrectomy as part of their treatment regime will be recruited onto the longitudinal arm of the study (see Figure 2). Patients in the longitudinal arm will have baseline blood and urine samples taken at registration and then between 4-60 days post-registration, but prior to nephrectomy. Following nephrectomy an FFPE block of tumour tissue is obtained, followed by further blood and urine samples at 3-6, 12, 18 and 24 months. Sampling will cease earlier if the patient relapses within this time period. However, a final sample must be collected at relapse, prior to initiation of any treatment for the relapse, see Figure 2. All RCC patients in both arms will be followed up annually for a period of up to 5 years. A blood and urine sample is required for healthy control volunteers at registration, with no follow up data required (see Figure 3). In all study arms clinical data is collected at different stages through the use of several case report forms (CRFs). Details of these forms can be found in Table 1.

Table 1: Details of Trial Forms

<table>
<thead>
<tr>
<th>Form</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F01</td>
<td>RCC Patient Eligibility &amp; Registration Form</td>
</tr>
<tr>
<td>F02</td>
<td>RCC Patient Baseline Assessment Form</td>
</tr>
<tr>
<td>F03</td>
<td>RCC Patient Surgery/Ablation/Pathology Details Form</td>
</tr>
<tr>
<td>F04</td>
<td>Sample Form</td>
</tr>
<tr>
<td>F05</td>
<td>RCC Patient Tissue Sample Form</td>
</tr>
<tr>
<td>F06</td>
<td>RCC Patient Follow-Up Form</td>
</tr>
<tr>
<td>F07</td>
<td>Healthy Volunteer Eligibility &amp; Registration Form</td>
</tr>
<tr>
<td>F08</td>
<td>Healthy Volunteer Baseline Assessment Form</td>
</tr>
</tbody>
</table>
Figure 1: Cross Sectional RCC Patient pathway

Sampling Regime

- Urine
- Bloods
  - Serum
  - EDTA Plasma
  - Buffy Coat
- Tissue Sample

Forms to be Completed

- F01
- F02
- F04*
- F03
- F05
- F06

* Included within the sample packs; ● Only if undergoing nephrectomy
Figure 2: Longitudinal RCC Patient pathway

- Urine
  - Bloods: Serum, EDTA Plasma, Buffy Coat

Registration (baseline)
Pre-Treatment (within 4-60 days of baseline)
Post-Nephrectomy
Follow Up (6-18 months)
Follow Up (12 & 24 months)
Follow Up (36, 48, 60 months)

Forma to be completed:
- F01
- F02
- F04*
- F03
- F05
- F04*
- F06
- F04*
- F05
- F05

* Included within the sample packs; • Only if undergoing nephrectomy; • Take a final sample at relapse.
Figure 3: Healthy Control Volunteer pathway

* Included within the sample packs
Section B  Trial Sample Handling

1. TISSUE SAMPLES

1.1 REQUESTING

Applicable to: Research/Clinical Team

For all patients undergoing nephrectomy, an FFPE tumour tissue block should be collected for the research tissue bank (in addition to those normally used for diagnosis).

- Immediately prior to nephrectomy please complete section A of Form 05 “Tissue request form” requesting an additional formalin-fixed paraffin-embedded (FFPE) tumour tissue block and attach to the standard hospital pathology request form sent with the kidney to pathology.

1.2. TISSUE SAMPLE PROCESSING

Applicable to: Histopathology

On receiving the request for an additional FFPE tumour tissue block (Form 05):

1. Ensure the request is logged on the local system according to normal local procedures.
2. Fix and prepare an additional FFPE tumour tissue block according to local standard operating procedures, but do not designate one for research until all blocks/sections have been reviewed as usual by the diagnostic pathologist.
3. Following the usual microscopic diagnostic examination of tumour tissue blocks, designate one for research use in the NIHR biomarker study.
4. Complete section B of Form 05 and pass the form and FFPE tumour tissue block to the local histopathology link person for the study.
5. Within 1 week of designation, the local histopathology link person for the study should complete section C of form 05 and send a copy alongside the designated FFPE block to the NIHR BioRTB, using the pre-addressed safebox packaging provided.

1.3 COLLECTION OF FORM 05

Applicable to: Research/Clinical Team

At locally agreed intervals a member of the research/clinical team should contact the Histopathology link person to arrange collection of the original version of form 05. Following collection, the original form 05 should be sent to the Leeds CTRU and a copy stored in the investigators site file.
2. BLOOD & URINE SAMPLES

Applicable to: Research Nurse/Clinician and Sample Processing Team

Two types of sample packs will be provided for processing blood samples: one for the RCC patient samples and another for healthy control volunteer samples.

The sample kits will contain the following:

**RCC sample pack:**
- Sample Form (Form 04)
- Tube Kit 01 (31 tubes)
- 31 Tube caps
- Pastettes (x4)
- 7 mL Biju (x2)
- 150 mL Urine Collection Pot
- 50 mL Centrifuge Tube
- 20 mL Barcoded Universal

**Healthy Control sample pack:**
- Sample Form (Form 04)
- Tube Kit 02 (30 tubes)
- 30 tube caps
- Pastettes (x4)
- 7 mL Biju (x2)
- 150 mL Urine Collection Pot
- 50 mL Centrifuge Tube
- 20 mL Barcoded Universal

Please take extra care to observe the following:
- Do not mix any of the contents between packs, as all barcodes are unique.
- Do not move tubes around within the tube kits as their location is pre-defined.
- If the tubes accidentally become re-arranged in the rack ensure that you:
  - a. Put the correct sample type in the correct rack location (see Figure 4)
  - b. Put the correct coloured lids on the sample tubes (yellow=urine; red=serum blue=plasma; white=buffy coat, see Figure 3)
  - c. Tell us exactly what happened on the sample form (Kits received without descriptions of errors will fail quality inspection and be discarded).

*Figure 4: Tube Kit Layout (Buffy Coat tubes are not included in the Healthy Control Sample Packs)*
2.1 COLLECTION PROCEDURE

1. Ensure the consent form has been completed and copies of the form have been placed in the patient notes (RCC patients only), filing the original in the investigator site file.

2. Select the appropriate sample pack and take to the clinic.

3. Record the following information on the sample form:
   - Patient/Volunteer initials
   - Patient/Volunteer date of birth
   - Patient/Volunteer ID
   - Date sample(s) were taken
   - Manufacturer of blood collection tube(s)
   - Times of venepuncture and urine sample
   - Any comments

Urine sample:
1. Collect urine (mid-stream), directly into the urine collection pot provided.
2. Mark the pot with the patient/volunteer ID, date of birth and initials.
3. Record time of urination on the sample form.
4. Place in bag and then back in pack box (provided).

Blood samples:
Please collect 8-10mL of blood into each tube type using the standard blood collection procedure and apparatus for venepuncture used in your hospital, not via a needle and syringe.

**Blood tubes for Serum:**
The tubes used for collection of serum samples should be an 8-10 ml plain clot activator tube (silica activator only)
   - These tubes are typically red top (serum) when sourced from Greiner and Becton Dickinson; but are white when sourced from Sansted.
   - Note: Please do not use tubes containing gel or separators for this sample.

**Blood tubes for Plasma:**
The tubes used for collection of EDTA Plasma samples should be 4-8 ml Potassium EDTA Plasma tube(s)
   - These tubes are typically purple (top) when sourced from Greiner and Becton Dickinson; but are red when sourced from Sanstedt.

**PROCEDURE**
1. Collect blood directly into appropriate tube(s). Mix by inverting gently 5x.
2. Mark the tube(s) with patient/volunteer ID, date of birth and initials.
3. Record time of venepuncture on the sample form (record both times if serum and plasma collected at different times)
4. Place in bag and then back in pack box (provided).

Take all samples for processing immediately to the laboratory within their sample box.
2.2 PROCESSING PROCEDURE

Please refer to Figure 6 for a flow diagram of the sample processing procedure

1. Cross check the IDs on the samples received with the sample form to ensure that the correct blood/urine samples have been received, and that none of the samples are missing.

2. Without removing the clear plastic lid, label the side of the Tube Kit rack with the patient ID, date of birth and initials.

3. Ensure blood samples are left for a minimum of 45 minutes post collection (refer to Sample Form: Time of venepuncture) at room temperature. Process blood samples as soon as possible after this time and freeze within 2 hours of venepuncture (if this is not possible please make a note in the comments section).

4. Urine samples should be processed at room temperature and frozen within 2 hours of collection.

Urine sample:

1. Transfer urine into 50ml centrifuge tube (provided in pack) and label with ID, date of birth and initials.

2. Centrifuge the urine at 2000 x g (approximately 3000rpm in many bench-top centrifuges - needs to be checked as varies with centrifuge type and size) for 10 mins.

3. Using a pastette (supplied) aliquot the urine into the 10 barcoded tubes in the top row (marked U) of the Tube Kit (see Figure 4). Fill each tube to just above the central black line (see Figure 5).

4. Place the yellow lids on these tubes.

5. Transfer the remaining centrifuged urine into the 20 mL bar-coded Universal (Supplied)

6. Record the time the urine samples were transferred on the Sample Form

7. Replace the lid of tube kit whilst processing blood samples.

Figure 5: Tube fill level (Please avoid overfilling all tubes, as the liquid will expand upon freezing.)
Blood Samples:
After a **minimum of 45 minutes** following venepuncture, centrifuge both the serum and plasma samples together at room temperature for 10 minutes at 2000 x g (approximately 3000rpm).

**Serum sample:**
1. Use a pastette (supplied) to remove as much of the serum as possible without disturbing the red cell clot. Dispense the serum into the pooling tube (supplied).
2. Aliquot the serum into the 10 bar-coded tubes in the **middle row (marked S)** of the Tube Kit (see Figure 4). Fill each tube to just above the central black line (see Figure 5).
3. Place the red lids on these tubes

**NOTE:** If only a small blood sample was obtained aliquot serum into fewer tubes and discard any unused tubes.

**Plasma sample:**
1. Use a pastette (supplied) to remove the **upper two thirds** of the plasma to avoid contamination with the buffy coat. Dispense the plasma into the pooling tube (supplied).
2. Aliquot the plasma into the 10 bar-coded tubes in the **lower row (marked P)** of the Tube Kit (see Figure 4). Fill each tube to just above the central black line (see Figure 5).
3. Place the blue lids on these tubes
4. Record the time the plasma samples were transferred on the Sample Form

**NOTE:** If only a small blood sample was obtained aliquot plasma into fewer tubes and discard any unused tubes.

**Buffy Coat sample:**
(For **Healthy Control** samples skip this step and proceed to “On Completion of Processing”)
1. Carefully aspirate the white buffy coat layer from the top of the red blood cells using a pastette (supplied) – don’t worry if some of the remaining plasma and some of the red blood cells are also collected.
2. Transfer into the solitary tube underneath the plasma samples marked B (see Figure 4).
3. Place the white lid on this tube.

*Please refer to Figure 6 for a flow diagram of the sample processing procedure*

2.3 **ON COMPLETION OF PROCESSING:**
1. Immediately store all tubes and Universals in a freezer at a temperature of less than -70°C.
2. Record the freezer location of the tube racks, and what time they were frozen on Sample Form 04.
3. Send original Sample Form 04 to Leeds CTRU and retain a copy in the site file.
Figure 6: Summary flow diagram of sample processing procedure

- Check IDs of samples against the sample form
- Label the Tube Kit rack with the patient ID, date of birth and initials.
- Process urine immediately (within 2 hours)
  - Urine Sample (Collection pot provided)
    - Transfer into 50ml centrifuge tube
    - Centrifuge at 2000 x g for 10 mins
    - Aliquot into the top 10 tubes
    - Place Yellow lids onto tubes
    - Transfer remaining centrifuged urine into the bar-coded universal tube
    - Record the time samples transferred on the Sample Form
- Process blood(s) after 45 minutes, but within 2 hours
  - Serum Sample (Plain Clot Activator Tubes)
    - Centrifuge at 2000 x g for 10 mins
    - Remove serum and dispense into a pooling tube
    - Aliquot into the middle 10 tubes
    - If serum volume is low, aliquot into fewer tubes
    - Place Red lids onto tubes
    - Record the time samples transferred on the Sample Form
  - Plasma Sample (Potassium EDTA Tubes)
    - Centrifuge at 2000 x g for 10 mins
    - Remove upper two thirds of plasma into pooling tube
    - Aliquot into the lower 10 tubes
    - If plasma volume is low, aliquot into fewer tubes
    - Place Blue lids onto tubes
    - Record the time samples transferred on the Sample Form
  - For Buffy Coat, Aspirate white layer and transfer to the single tube at bottom of rack
- Place White lid onto tube
- Freeze tube racks at < -70°C, record the location and time on the Sample Form
- Send original Sample Form to Leeds CTRU, retain copy locally in investigator site file
3. **SHIPMENT PROCEDURE**

Frozen samples will be stored at each site until required to be shipped for biobanking. At this time a coordinator from The Leeds NII-R Biomarker Research Tissue Bank will contact you with a request form including a list of all samples to be shipped and details of the courier who will liaise with you over delivery of the packing materials and pick-up date. All shipping materials will be supplied by the courier and must be used as per the instructions in accordance with UN3373 to avoid possible leakage of materials.

Frozen samples:
1. The dry ice and shipment containers will be provided by a courier.
2. Place the locked sample kit boxes and bagged universal tubes into the thermal shipment container. Please note when packing samples they should not be allowed to warm or thaw out and should be kept on dry ice at all times once removed from the freezer and packed as quickly as possible.
3. Fill the thermal shipment container with dry ice to the top, place lid on container.
4. Sign and date the request form, place form in the box and fold over all flaps. The samples are now ready for transportation.

Monitor sample collection and ensure that the samples have been collected as planned – contact the courier if not.

4. **QUERIES**

If you have any questions, please contact the Study Manager at Leeds CTRU (Tel: 0113 343 1498 Fax: 0113 343 1487 if it relates to any forms or clinical data; and Dr Michael Messenger (Tel: 0113 206 5267 or bioRTB@leeds.ac.uk ) if it relates to queries about sample processing or collection
Title: TRANSLATIONAL MANUAL: RENAL TRANSPLANT SAMPLE HANDLING

Trial Name: Evaluation of Biomarkers for Post Renal Transplant Complications

Version: 4.0  Date: 31.07.2013

Details:

Author(s) of Study Site Operating Procedure: Michael Messenger, Tobias Wind, Damien Hindmarch

Comments:

The following Site-Specific Procedures are for collection, processing, and distribution of samples for the Renal Transplant NBHR Biomarker Research Tissue Bank (bioRTB). The objective being to validate biomarkers for diagnosis, prognosis and longitudinal monitoring in patients with renal transplant complications

Version Control:

<table>
<thead>
<tr>
<th>Version number</th>
<th>Edited by</th>
<th>Date edit completed</th>
<th>Details of editions made</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>MPM</td>
<td>09.03.12</td>
<td>Updated Figure 1</td>
</tr>
<tr>
<td>3</td>
<td>MPM</td>
<td>23.04.13</td>
<td>Updated Introduction &amp; Figure 1 to document increased recruitment figures and clarify sample collection points. Table 1 updated to include new forms F01a, F01b &amp; F09. Confirmed sample processing times in section 2.2</td>
</tr>
<tr>
<td>4</td>
<td>MPM</td>
<td>31.07.13</td>
<td>Revised Introduction &amp; Figure 1 to account for reduction in sample collection. Samples are to be collected daily for first week of hospital stay, then weekly for one month, then at 2, 3 and 6 months from discharge. Also clarified that sampling and patient participation ends following graft failure, transplant nephrectomy or death</td>
</tr>
</tbody>
</table>

Contents:

Details: .............................................................................................................. 1
Comments: ........................................................................................................ 1
Version Control: .............................................................................................. 1
Contents .............................................................................................................. 1
Section A Introduction ....................................................................................... 2
Section B Trial Sample Handling ...................................................................... 4
  1. BLOOD & URINE SAMPLES ........................................................................ 4
     2.1 COLLECTION PROCEDURE ................................................................. 4
     2.2 PROCESSING PROCEDURE .................................................................. 7
     2.3 ON COMPLETION OF PROCESSING .................................................... 8
  3. SHIPMENT PROCEDURE ............................................................................. 10
  4. QUERIES ................................................................................................. 10
Section A  Introduction

This Study Site Operating Procedure (SSOP) is applicable to the Principal Investigator, Research Nurse, and any other member of staff at research sites who have responsibilities within the Evaluation of Biomarkers for Post Renal Transplant Complications a study for the collection and processing of samples for the Leeds NIHR Biomarker Research Tissue Bank (bioRTB).

The objective of the study is to validate diagnostic, prognostic and longitudinal monitoring markers of renal transplant complications using prospectively collected high quality clinical samples from multiple centres. Blood and urine samples will be requested from eligible patients attending participating centres.

Figure 1 illustrates the patient pathway, their associated sampling regimes and the forms requiring completion at each stage. Up to 850 patients on the renal transplant waiting list will be recruited onto the study. Patients are required to provide blood and urine samples at baseline (consent and immediately pre-operatively, where possible), daily during the first week of hospital stay, then weekly for 1 month, then at months 2, 3 and 6 (i.e. 6 months from date of discharge). Patients will be followed up annually for a period of up to 5 years. Sample collection and follow up will end if patients suffer graft failure, transplant nephrectomy or death. Clinical data is collected at different stages through the use of several case report forms (CRFs). Details of these forms can be found in Table 1.

Table 1: Details of Trial Forms

<table>
<thead>
<tr>
<th>Form</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F01a</td>
<td>Eligibility &amp; Registration</td>
</tr>
<tr>
<td>F01b</td>
<td>Baseline Assessment</td>
</tr>
<tr>
<td>F02</td>
<td>Preoperative Assessment</td>
</tr>
<tr>
<td>F03</td>
<td>Post-operative Investigations</td>
</tr>
<tr>
<td>F04</td>
<td>Intra/Post-Operative Investigation</td>
</tr>
<tr>
<td>F05</td>
<td>Follow-up</td>
</tr>
<tr>
<td>F06</td>
<td>Sample Form</td>
</tr>
<tr>
<td>F99</td>
<td>Withdrawal</td>
</tr>
</tbody>
</table>
Figure 1: Renal transplant patient pathway

1. If a patient will be dialysed try to collect blood samples pre-dialysis
2. Complete for the first week (7 days) of hospital stay
3. Form completed at the end of hospital stay
4. Forms included within the sample packs
5. Only collect out of hours (weekend) samples where possible
6. Follow-up schedule starts once patient is discharged
Section B  Trial Sample Handling

1. BLOOD & URINE SAMPLES

Applicable to: Research Nurse/Clinician and Sample Processing Team

Sample packs will be provided for processing blood and urine samples. The sample packs will contain the following:

Renal transplant sample pack:
- Sample Form (Form 06)
- Tube Kit 02 (30 tubes)
- 30 Tube caps
- Pasteles (x4)
- 7 mL Bijou (x2)
- 150 mL Urine Collection Pot
- Medium size bag
- 50 mL Centrifuge Tube
- 20 mL Barcoded Universal

Please take extra care to observe the following:
- Do not mix any of the contents between packs, as all barcodes are unique
- Do not move tubes around within the tube kits as their location is pre-defined.
- In the event that tubes are accidentally moved within the rack, ensure that you:
  a. Put the correct sample type in the correct rack location (see Figure 2)
  b. Put the correct coloured lids on the sample tubes (yellow=urine; red=serum blue=plasma, see Figure 3)
  c. Tell us exactly what happened on the sample form (Kits received without descriptions of errors will fail quality inspection and be discarded).

2.1 COLLECTION PROCEDURE

1. Ensure the consent form has been completed and copies of the form have been given to the patient, placed in the patient notes and the investigator site file.

2. Take a sample kit and record the following information on the sample form:
   - Patient/Volunteer initials
   - Patient/Volunteer date of birth
   - Patient/Volunteer ID
   - Date sample(s) were taken
   - Manufacturer of blood collection tube(s)
   - Times of venepuncture and urine sample
   - Any comments
Figure 2: Tube Kit 02 Layout
Urine sample:
1. Collect urine into the urine collection pot provided.
2. Record whether it was collect directly (mid-stream) or via catheter.
3. Mark the pot with the patient ID, date of birth and initials.
4. Record time of sample collection on form F06 and mark whether taken on the same date as the bloods. If not, record urine sampling date.
5. Place in bag and then back in pack box (provided).

Blood samples:
Please collect 6-10mL of blood into each tube type using the standard blood collection procedure and apparatus used in your hospital, not via a needle and syringe.

Blood tubes for Serum:
The tubes used for collection of serum samples should be plain clot activator tube (silica activator only)
- These tubes are typically red top (serum) when sourced from Greiner and Becton Dickinson; but are white when sourced from Sarstedt.
- Note: Please do not use tubes containing gel or separators for this sample

Blood tubes for plasma:
The tubes used for collection of EDTA Plasma samples should be Potassium EDTA Plasma tube(s)
- These tubes are typically purple (top) when sourced from Greiner and Becton Dickinson; but are red when sourced from Sarstedt.

PROCEDURE
1. Collect blood directly into appropriate tube(s). Mix by inverting gently 5 x.
2. Record whether blood was collected via venepuncture (preferred) or central line.
3. Mark the tube(s) with patient/volunteer ID, date of birth and initials.
4. Record time of venepuncture on the sample form (record both times if serum and plasma collected at different times)
5. Place in bag and then back in pack box (provided).

Take all samples for processing immediately to the laboratory within their sample box.
2.2 PROCESSING PROCEDURE

Please refer to Figure 4 for a flow diagram of the sample processing procedure

1. Cross check the IDs on the samples received with the sample form (F06) to ensure that the correct blood/urine samples have been received, and that none of the samples are missing.

2. Label the Tube Kit rack with the patient ID, date of birth and initials.

3. Ensure blood samples are left for a minimum of 45 minutes post collection (refer to Sample Form: Time of venepuncture) at room temperature. Process blood samples as soon as possible after this time and freeze within 2 hours of venepuncture (if this is not possible please make a note in the comments section).

4. Urine samples should be processed at room temperature and frozen within 2 hours of collection.

Urine sample:
1. Transfer urine into the centrifuge tube (provided) and label with ID, date of birth and initials.
2. Centrifuge the urine at 2000 x g (approximately 3000rpm in many bench-top centrifuges - needs to be checked as varies with centrifuge type and size) for 10 mins.
3. Using a pastette (provided) aliquot the urine into the 10 barcoded tubes in the top row (row A, marked U) of the Tube Kit (see Figure 2). Fill each tube to just above the central black line (see Figure 3).
4. Place the yellow lids on these tubes.
5. Transfer the remaining centrifuged urine into the 20 mL bar-coded Universal tube (provided).
6. Record the time the urine samples were transferred on the Sample Form
7. Replace the lid of tube kit whilst processing blood samples.

Figure 3: Tube fill level (Please avoid overfilling tubes, as the liquid will expand upon freezing.)

[Image of tube fill level]
Blood Samples:
After a minimum of 45 minutes following venepuncture, centrifuge both the serum and plasma samples together at room temperature for 10 minutes at 2000 x g (approximately 3000rpm).

Serum sample:
1. Use a pastette (provided) to remove as much of the serum as possible without disturbing the red cell clot. Dispense the serum into the pooling tube (provided).
2. Aliquot the serum into the 10 bar-coded tubes in the middle row (row C, marked S) of the Tube Kit (see Figure 2). Fill each tube to just above the central black line (see Figure 3).
3. Place the red lids on these tubes

NOTE: If only a small blood sample was obtained aliquot serum into fewer tubes and discard any unused tubes, note the number of aliquots on the sample form.

Plasma sample:
1. Use a pastette (provided) to remove the upper two thirds of the plasma to avoid contamination with the buffy coat. Dispense the plasma into the pooling tube (provided).
2. Aliquot the plasma into the 10 bar-coded tubes in the lower row (row E, marked P) of the Tube Kit (see Figure 2). Fill each tube to just above the central black line (see Figure 3).
3. Place the blue lids on these tubes
4. Record the time the blood samples were transferred on the Sample Form

NOTE: If only a small blood sample was obtained aliquot plasma into fewer tubes and discard any unused tubes, note the number of aliquots on the sample form

Please refer to Figure 4 for a flow diagram of the sample processing procedure

2.3 ON COMPLETION OF PROCESSING:
1. Immediately store all tubes and Universals in a freezer at a temperature of less than -70°C.
2. Record the freezer location of the tube racks, and what time they were frozen on Sample Form 06.
3. Send original Sample Form 06 to Leeds CTRU and retain a copy in the site file.
**Figure 4: Summary flow diagram of sample processing procedure**

- **Check IDs of samples against the sample form**
- **Label the Tube Kit rack with the patient ID, date of birth and initials.**

**Urine Sample** (Collection pot provided)
- **Transfer into centrifuge tube**
- **Centrifuge at 2000 x g for 10 mins**
- **Aliquot into the top 10 tubes**
- **Place Yellow lids onto tubes**
- **Transfer remaining centrifuged urine into the bar-coded universal tube**
- **Record the time samples transferred on the Sample Form**

**Serum Sample** (Plain Clot Activator Tubes)
- **Centrifuge at 2000 x g for 10 mins**
- **Remove serum and dispense into a pooling tube**
- **Aliquot into the middle 10 tubes**
- **If serum volume is low, aliquot into fewer tubes**
- **Place Red lids onto tubes**
- **Record the time samples transferred on the Sample Form**

**Plasma Sample** (Potassium EDTA Tubes)
- **Centrifuge at 2000 x g for 10 mins**
- **Remove upper two thirds of plasma into pooling tube**
- **Aliquot into the lower 10 tubes**
- **If plasma volume is low, aliquot into fewer tubes**
- **Place Blue lids onto tubes**
- **Record the time samples transferred on the Sample Form**

- **Freeze tube racks at <=70°C, record the location and time on the Sample Form**
- **Retain a copy of the sample form (FOE) in the investigator site file and send the original to CTRU**
3. SHIPMENT PROCEDURE

Frozen samples will be stored at each site until required to be shipped for biobanking. At this time a coordinator from the Leeds NIF-RR Biomarker Research Tissue Bank will contact you with a request form including a list of all samples to be shipped and details of the courier who will liaise with you over delivery of the packing materials and pick-up date. All shipping materials will be supplied by the courier and must be used as per the instructions in accordance with UN3373 to avoid possible leakage of materials.

Frozen samples:
1. The dry ice and shipment containers will be provided by a courier.
2. Place the locked sample kit boxes and bagged universal tubes into the thermal shipment container. Please note when packing samples they should not be allowed to warm or thaw out and should be kept on dry ice at all times once removed from the freezer and packed as quickly as possible.
3. Fill the thermal shipment container with dry ice to the top, place lid on container.
4. Sign and date the request form, place form in the box and fold over all flaps. The samples are now ready for transportation.

Monitor sample shipment and ensure that the samples have been collected as planned – contact the courier if not.

4. QUERIES

If you have any questions, please contact the Data Manager at Leeds CTRU (Tel: 0113 343 1498 Fax: 0113 343 1487) if it relates to any forms or clinical data; and Dr Michael Messenger (Tel: 0113 206 5267 or bioRTB@leeds.ac.uk) if it relates to queries about sample processing or collection
Study Site Operating Procedure

Title: TRIAL SAMPLE HANDLING
Trial Name: ELUCIDATE TRIAL
Version: 3.0, Date: 16.11.2010

Details:

Author(s) of Study Site Operating Procedure: Dr. Michael Messenger (NIHR bioRTB)
Claire Davies (CTRU)
Carly Rivers (CTRU)

Comments:

The following Site Specific Procedures are for collection, processing, and distribution of samples for the ELF test and NIHR Biomarker Research Tissue Bank (bioRTB).

Version Control:

<table>
<thead>
<tr>
<th>Version number</th>
<th>Edited by</th>
<th>Date edit completed</th>
<th>Details of editions made:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>CD</td>
<td>22.09.2010</td>
<td>Section B1 amended to permit use of needle and syringe but to advise that notes should be made on the ELF sample form. Section B3 also amended to request that notes be added to the ELF sample form if the serum sample was not processed in the 2 hour window.</td>
</tr>
<tr>
<td>2.0</td>
<td>CD</td>
<td>16.11.2010</td>
<td>Section B5 updated with amended e-mail addresses for CTRU and Biobank lab.</td>
</tr>
</tbody>
</table>

Contents:

- Details .................................................................................................................................................. 1
- Comments: ............................................................................................................................................. 1
- Version Control: .................................................................................................................................. 1
- Contents: .............................................................................................................................................. 1
- Section A: Applicability ......................................................................................................................... 2
- Section B: Trial Sample Handling ....................................................................................................... 3
- 1. SAMPLE KITS AND GENERAL INSTRUCTIONS ........................................................................... 3
- 2. COLLECTION PROCEDURE ............................................................................................................ 4
- 3. PROCESSING PROCEDURE ................................................................................................................ 5
- 4. ON COMPLETION: ............................................................................................................................ 7
Section A  Applicability

This Study Site Operating Procedure (SSOP) is applicable to the Principal Investigator, Research Nurse, and any other member of staff at the research site who have responsibilities within the ELUCIDATE Trial for the collection, processing and despatch of samples for the ELF test and NIHR Biomarker Research Tissue Bank (bioRTB).

All patients require an ELF test sample to be taken at registration and randomisation. In the ELF arm, follow-up samples are to be taken every 6 months. In the control arm a single ELF test will be taken at diagnosis of cirrhosis. A single sample for the NIHR biomarker RTB is only taken at randomisation in addition to the ELF test sample, refer to Figure 1. Note that patients on heparin are not eligible for the study as the ELF test cannot be performed.

The registration samples should be taken non-fasted. Randomisation samples should be taken fasted. Follow-up samples should be taken non-fasted. For the purposes of this trial, a patient is considered fasted if they have had no food (water only) overnight or for 4 or more hours.

Figure 1: Flow diagram of the patient pathway and associated sampling regimes
Section B  Trial Sample Handling

1. SAMPLE KITS AND GENERAL INSTRUCTIONS

Two types of sample kit will be provided for processing blood samples: one for the ELF Test samples and another for the ELF Test and NIHR biomarker RTB samples.

The sample kits will contain the following:

ELF TEST sample kit:
- Sample Form (Form 05)
- ELF Shipping Form
- Pastettes (x2)
- Self-adhesive blood tube label
- 1x purple capped ELF TEST tube
- Pre-paid and addressed Royal Mail Safebox

Randomisation ELF TEST and NIHR Biomarker RTB sample kit:
- Sample Form (Form 09)
- ELF Shipping Form
- 1x purple capped ELF TEST tube
- Pooling Tube
- Pastettes (x3)
- Self-adhesive blood tube labels
- 10 x red capped NIHR bioRTB tubes
- Pre-paid and addressed Royal Mail Safebox

Please do not mix contents between kits as barcodes are unique to each kit/form/sample.

Please collect blood using the standard blood collection procedure and apparatus for venepuncture used in your hospital, preferably not via a needle and syringe. If a needle and syringe must be used then please make a note in the comments section and take care to remove the needle prior to filling the blood tubes. You will need to supply the actual blood collection tubes as below (do not use tubes containing EDTA or heparin):

- The tube used for the ELF TEST serum samples should be a 4-6 ml serum separator tube (SST)
  - SST tubes are typically red or gold top (serum) depending on the manufacturer and contain a gel.

- The tubes used for NIHR Biomarker RTB serum samples should be an 8-10 ml plain clot activator tube (silica activator only)
  - These tubes are typically red top (serum) when sourced from Greiner and Becton Dickinson; but are white when sourced from Sansted.
  - Note: Please do not use tubes containing gel or separators for this sample.
2. **COLLECTION PROCEDURE**

1. Ensure the patient consent form has been completed and copies of the form and the patient information leaflet (PIL) have been placed in the patient notes, filing the original in the investigator site file.

2. Identify what stage of the patient pathway (see Figure 1) the patient is at and what procedures they have consented to. Only collect the **NIHR biomarker RTB sample** if at randomisation and the patient has consented.

3. Select the appropriate **sample kit** and take to the clinic.

4. Record the following information on the **sample form**:
   - Patient Initials
   - Patient Date of birth
   - Patient ID
   - Date sample(s) were taken
   - Patient fasted/non-fasted*
   - If at Randomisation, whether a NIHR bioRTB sample has been taken.
   - Manufacturer of blood collection tube(s)
   - Any comments

*fasted defined as no food (water only) either overnight or for more than four hours

**For ELF-TEST sample:**
1. Collect 4-6 mL blood directly into a **serum separator/gel tube (SST)**. Mix by inverting gently 5x.

2. Stick the **ELF-TEST** self-adhesive blood tube label (provided in sample kit) to the SST tube and mark-up with patient ID, date of birth and initials.

3. Record time of venepuncture on the **SAMPLE kit bag** and **sample form**.

4. Place back in **kit bag** (provided).

**For NIHR Biomarker RTB sample:**
1. Collect 8-10 mL blood directly into the **plain clot activator tube**. Mix by inverting gently 5x.

2. Stick NIHR bioRTB self-adhesive blood tube label (provided in kit) to the plain clot activator tube and mark-up with patient ID, date of birth and initials.

3. If different to ELF test, record time of venepuncture on the **SAMPLE kit bag**. (also make a note in the comments section of the sample form).

4. Place back in **kit bag** (provided).

Take all samples for sample processing immediately to the laboratory within a closed sample bag.
3. PROCESSING PROCEDURE

*Please refer to Figure 2 for a flow diagram of the sample processing procedure*

Cross check the samples received with the sample form to make certain no blood samples are missing. Ensure the sample(s) are left to clot for a minimum of 45 minutes post collection (time of venepuncture) at room temperature. As soon as possible after this time and within 2 hours (if this is not possible please make a note of the time in the comments section), centrifuge at room temperature for 10 minutes at 2000 x g (approximately 3000rpm in many bench-top centrifuges - needs to be checked as varies with centrifuge type and size).

- Process all sample(s) not less than 45 minutes after venepuncture, but within 2 hours.
- Centrifuge at room temperature, for 10 mins at 2000 x g.

**Which samples were received?**

**ELF Test** (SST*/gel TUBES) **AND/ OR** **NIHR bioRTB** (Plain Clot Activator Tubes)

**Transfer serum to purple top tube labelled “ELF TEST” (supplied in kit)**

**Record the time serum transferred to ELF TEST tube on the sample form**

**Write patient date of birth on ELF sample tube and place in supplied safebox™ packaging.**

**Complete the supplied ELF Test shipping form, place in safebox™ and send.**

**Transfer serum to the pooling tube (supplied in kit)**

**Aliquot into red top “NIHR bioRTB” tubes (supplied in kit) and record the time of serum transfer on the sample form**

**Write patient ID on tubes**

**Immediately freeze the samples at -70 to -80°C.”**

**Record on the sample form: - time samples frozen - number of aliquots - location of storage**

Fax/e-mail **Sample Form** to Leeds CTetu, retain original locally in investigator site file.

*Figure 2: Procedure for sample processing*
ELF Test sample: (SST/gel tube)

1. Following centrifugation of the ELF TEST serum separator/gel tube sample, use a disposable pastette (supplied) to remove approximately 1ml of the serum, without disturbing the red cells. Transfer serum to purple capped ELF TEST tube (supplied in kit).
2. Record the time the serum was transferred on the sample form.
3. Write the patient date of birth on the ELF TEST sample tube in permanent marker. (Please do not write patient ID on the ELF sample)
4. Place the ELF TEST sample in the supplied absorbent packaging then into the grip seal bag and finally into the sealable container within the pre-paid safebox® (Do not close safebox®).

**NOTE: Once the safe box is closed it cannot be re-opened.**
5. Complete the supplied “ELF Test Shipping form” and retain a copy in the investigator site file. Place original in the safebox®.
6. Close the safebox® and send via the normal postal system.
7. Record any issues on the sample form.

NIHR bioRTB sample: (plain clot activator tube)

1. Following centrifugation of the NIHR bioRTB plain clot activator tube sample, use a disposable pastette (supplied) to remove as much of the serum as possible without disturbing the red cells. Dispense the serum into the pooling tube (supplied).
2. Using a pastette (supplied), aliquot the serum approximately equally into the 10 “NIHR bioRTB” red capped tubes (supplied). If only a small blood sample was obtained and the resulting serum volume is less than 2 ml, aliquot the serum approximately equally into only 5 tubes and discard the other 5 unused tubes.
3. Record the time the serum was transferred on the sample form.
4. Write the patient ID on the sample tubes in permanent marker.
5. Within 10 minutes of completion store the NIHR bioRTB sample tubes (red caps) in a freezer at a temperature of between -70°C and -80°C.
6. Record the number of tubes frozen, location and what time they were frozen on the sample form.
7. Record any issues on the sample form.

N.B. Any deviations from this procedure must be documented on the sample form.

**NIHR bioRTB Shipment Procedure:**

Frozen samples will be stored until required to be shipped for banking. At this time a coordinator from the Leeds NIHR Biomarker RTB will contact you to arrange for samples to be collected and shipped. Contact details can be found below. The dry ice and shipment containers will be provided by a specialist courier and must be used as per instructions to comply with UN3373 and avoid leakage of materials or personal injury.

On arrival of the courier please do the following:

- Pack the requested samples into the thermal shipment container in the 13x13x5cm storage boxes provided previously by the NIHR bioRTB.
- Fill with dry ice to the top of the thermal shipment container, place lid on unit.
- Sign and date the request form and place in box and fold over all flaps. The samples are now ready for transportation.
4. **ON COMPLETION:**

Fax/email sample form to Leeds CTRU as described in the final CRF and retain original locally in investigator site file. (Fax: 0113 343 1471 or c.davies@leeds.ac.uk)

5. **ANY QUERIES:**

If you have any questions, please contact Claire Davies, Senior Trial Manager at the CTRU (Tel: 0113 343 1498 or ELUCIDATE@leeds.ac.uk)

For queries relating specifically to the **NIHR bioRTB samples please contact Dr Michael Messenger (Tel: 0113 206 5267 or bioRTB@leeds.ac.uk)**
Renal Cancer Marker Study
PATIENT INFORMATION SHEET

Leeds NIHR Biomarker Research Tissue Bank
We would like to invite you to take part in a research study. Before you decide whether you would like to take part, you need to understand why the research is being done and what it would involve for you. Please take time to read the following information carefully and discuss it with your family, friends or your GP if you wish. Please take the opportunity to put any questions you may have to a qualified and experienced person. If you decide that you are happy to take part, please sign the attached consent form and have this witnessed. Keep a copy of this information for future reference.

What kind of research is being done and why?

Testing human tissue samples and fluids such as urine and blood is necessary to understand how the human body works normally and what changes when things go wrong. We are currently carrying out research into several diseases involving the kidney including renal (kidney) cancer. The main purpose of this research is to develop new clinical tests that can identify changes in the proteins, genes or other substances that we can measure ("biomarkers") in patient samples. These biomarker tests may be used to improve patient care such as helping to diagnose disease earlier or in deciding which drug is having the best effect in a patient.

What am I being asked to donate and what procedures are involved?

Both healthy and unhealthy tissue, cells and fluid samples are important to us as we need to compare them. We are asking you to donate:

- Some kidney tissue if you are having an operation to remove a small bit of your kidney (a biopsy) or all of one of your kidneys (a nephrectomy) as part of your routine treatment. When the surgeons do this, for example as part of your normal care to diagnose your condition, the tissue which is removed goes to the
pathology department where a specialist pathologist will examine it. Often there is spare tissue which they don't need for clinical purposes and we are asking you to donate any unused or spare tissue not needed for your clinical tests. This involves no extra procedures for you at all.

- Blood and urine samples. We would ask your consent to take a small amount of blood (normally less than 20 mls or 5 teaspoons) and/or to provide a urine sample for our research. Often at the hospital you will have a blood sample taken from you ("venepuncture") as a standard part of your clinical care. If you consent to an additional sample for research we would wherever possible take this at the same time through the same needle and therefore avoid any additional needle punctures. If it's not possible and we have to get a sample for research at a different time to your routine venepuncture for clinical blood tests it involves exactly the same process. In all cases this will be carried out by a fully trained member of staff. We will also ask you (if appropriate) if you would be willing to give further samples at other times in the future during your hospital visits.

Donating these samples will not make any difference to the tests that are needed for your clinical care.

**Do I have to take part?**

No, the decision of whether or not to take part is completely up to you. Deciding to donate or not has no impact on the type or standard of care you receive, now and in the future.

**What are the benefits or advantages of taking part?**

Research studies usually take many years to complete. You will be contributing to a bank of tissue, cell and fluid samples, which may help to speed up research into
human disease. The results of the research overall may benefit patients with renal cancer in the future. In addition to contributing to generating new knowledge in medical research, it may also decrease the need to rely on testing on animals. However as the research results are about improving care and tests in the future and are not current clinical tests, the results of our experiments on your samples would not be given to you individually.

As an unconditional gift, the benefits of donating tissue, cell and fluid samples are humanitarian rather than personal. You will not receive any financial reward, including from the successful development of any drug or treatment, which might arise from the research and later goes on to make a profit.

What are the risks to me of donating my tissues and fluids?

There are no additional health risks associated with donating samples for research purposes if they are taken as part of a normal diagnostic procedure. If we are taking a blood sample at a different time from your routine tests, the only risks would be minor bruising. If you are a patient and anything in the procedure for obtaining your samples were to go wrong, the normal complaint mechanisms of the NHS are open to you.

What will my samples be used for?

Your samples will be used in various research projects which will involve large-scale analysis of the proteins present in your tissues, cells and fluids to help us understand the biology of your illness and develop new biomarker tests. Samples from some patients may also possibly be involved in studies examining the genetic material (DNA and RNA) and may undergo variety of procedures including whole genome sequencing. This could determine many or all of the features of your DNA but we are interested only in results which are relevant to your illness. None of these results will be passed back to you individually and they will be kept absolutely confidential.
Could any of the results show that I have other illnesses?

It is possible that some of this information may show changes which could be relevant to other illnesses. For example we may find results that show that you are possibly at risk of a genetically determined illness and this may also be relevant to your relatives. As the tests which we carry out are for research purposes only and are not current clinical tests, any results of that kind would need to be considered and investigated properly by a qualified doctor to ensure the information is correct. You have the option of choosing that if such a finding occurs you would like us to keep it absolutely confidential and take no action or we could contact your GP who would then investigate any possibilities with you further using current clinical practice and tests. Also if we are unable to contact you for any reason you can choose whether or not you want us to inform your relatives about these results.

Can I withdraw my consent if I change my mind?

Yes you can if the samples and/or data have not yet been used. Unused data and samples would, after your notice of withdrawal, be disposed of securely and respectfully. If you change your mind and your samples or data have been used, your gift may have already contributed to new knowledge. This cannot be recalled. If you change your mind when you are still in hospital, you can ask a member of your clinical team to inform us on your behalf. If you change your mind later, or you would prefer not to approach us directly, you can write confidentially to our organisation’s Research and Development Dept, who will ensure that your wishes are carried out. A standard letter has been given to you for this purpose.

Who will know I am participating in the research?

The only people who will know your identity are hospital staff and a limited number of staff at the Clinical Trials Research Unit where data is stored on secure computers.
All are bound by a professional duty to protect your privacy. An identification number will be assigned to your samples, which ensures that researchers cannot identify you personally from your donation. This will be used in any other databases where details of your donated samples and associated information are stored.

**Will any of my personal information be used?**

We are asking for your permission for staff to access and use information from your clinical records, including those held electronically. The information we collect will only be that which is relevant to our research and will include general information such as age, gender, any medication you may be on, whether or not you smoke and what kind of diet you eat, as well as information more specifically about your illness such as pathology results, results of routine blood tests and any scan (CT) results and how you respond to different treatments. Access may start at the time you donate your samples and/or be required later e.g. to look at your clinical progress. Before your information is released to researchers, it is anonymised keeping only an identification number. Participants’ identities will not be disclosed either to other researchers or when the results of the research are made public.

**Who is funding the research?**

This study is funded by the National Institute for Health Research (NIHR) but we also receive funding from other sources including Cancer Research UK and the Medical Research Council. Occasionally, we may also receive collaborative grants from companies such as pharmaceutical or diagnostic companies, particularly where we are developing new diagnostic tests in partnership for example. These grants allow us to recover our costs, and any funds we receive in excess of our costs are used to fund further research.
**Are there any other third parties involved in the research?**

We may collaborate with other researchers in the UK or abroad. They may work in universities, hospitals or the private sector. Your tissue will not, however, be sold for profit.

**Who has reviewed the research?**

Our research is reviewed by panels of experts associated with the various funding bodies and within academic research internationally. It is also reviewed by relevant ethics committees. This Research Tissue Bank has been approved by Leeds Research Ethics committee on 15th June 2010.

**Will I get feedback from the research?**

Any findings resulting from the research will be published in scientific or medical journals. Information will be available on the Leeds Teaching Hospitals research website, the NIHR Renal and Liver Biomarkers Programme website (www.biomarkerspipeline.org) and the research group website (www.proteomics.leeds.ac.uk).

**Donating to the wider research community?**

Other research groups, within Leeds Teaching Hospitals, Leeds University (where the sample bank is based) or elsewhere, are also dependent on donations of tissue, cell and fluid samples to make progress. We would like you to consider whether you would like us to restrict the use of your samples to our research group (and those groups who we work with directly in collaboration), or whether you give us permission to share your samples and associated anonymous data with other research groups. Any project is reviewed by the Research Tissue Bank Management committee, to
make sure that it is scientifically sound and that it fits with the consent that you have given. We will not release any samples unless we are satisfied that our committee has approved the project and the research group has agreed to abide by our conditions. Please tell us what you decide in the consent form.

There are costs involved in storing and sending samples, and we may ask external researchers to contribute to those costs, but we will not make a profit.

**Other things to consider**

Your tissue may be used for research that involves:
- Export for use in research outside the UK
- Commercial research e.g. developing new tests

Your tissue will **not** be used for research that involves:
- Research involving therapeutic/reproductive cloning (the latter remaining illegal under the Human Fertilisation and Embryology Authority 2001)
- Research involving human embryos and stem cells
- Research involving animal-human hybrid embryos
- Research into termination of pregnancy or contraception

**Who to contact for Further Information:**

If you would like further information you can either
- Ask the person who has provided this booklet to you
- Contact the Principal Investigator,
  Professor Peter Selby
  Email: p.j.selby@leeds.ac.uk

*Thank you for reading this patient information sheet.*

Leeds NIHR Biomarker RTB RCC Patient Information Booklet Version 1.3 January 2011
Renal Transplant Marker Study
PATIENT INFORMATION SHEET

Leeds NIHR Biomarker Research Tissue Bank
We would like to invite you to take part in a research study. Before you decide whether you would like to take part, you need to understand why the research is being done and what it would involve for you. Please take time to read the following information carefully and discuss it with your family, friends or your GP if you wish. Please take the opportunity to put any questions you may have to a qualified and experienced person. If you decide that you are happy to take part, please sign the attached consent form and have this witnessed. Keep a copy of this information for future reference.

What kind of research is being done and why?

Testing human fluids such as urine and blood is necessary to understand how the human body works normally and what changes when things go wrong. We are currently carrying out research into several diseases involving the kidney including those involving kidney transplantation. The main purpose of this research is to develop new clinical tests that can identify measurable changes in proteins ("biomarkers") in patient samples. These biomarker tests may be used to improve patient care such as diagnosing disease earlier or in deciding which drug is having the best effect in a patient.

What am I being asked to donate and what procedures are involved?

We are asking you to donate:

Blood and urine samples. We would ask your consent to take a small amount of blood (normally less than 20 mls or 5 teaspoons) and/or to provide a urine sample for our research. Often at the hospital you will have a blood sample taken from you ("venepuncture") as a standard part of your clinical care. If you consent to an additional sample for research we would wherever possible take this at the same time.
through the same needle and therefore avoid any additional needle punctures. If it’s not possible and we have to get a sample for research at a different time to your routine venepuncture for clinical blood tests it involves exactly the same process. In all cases this will be carried out by a fully trained member of staff. We will also ask you (if appropriate) if you would be willing to give further samples at other times whilst in hospital or in later hospital visits.

Donating these samples will not make any difference to the tests that are needed for your clinical care.

**Do I have to take part?**

No, the decision of whether or not to take part is completely up to you. Deciding to donate or not has no impact on the type or standard of care you receive, both now and in the future.

**What are the benefits or advantages of taking part?**

Research studies usually take many years to complete. You will be contributing to a bank of clinical samples which may help to speed up research into human disease. The results of the research overall may benefit patients with renal transplant rejection and delayed graft function in the future. In addition to contributing to generating new knowledge in medical research, it may also decrease the need to rely on testing on animals. However as the research results are about improving care and tests in the future and are not current clinical tests, the results of our experiments on your samples would not be given to you individually.

As an unconditional gift, the benefits of donating samples are humanitarian rather than personal. You will not receive any financial reward, including from the successful

Leeds NIHR Biomarker RTB Renal Transplant Patient Information Booklet Version 1.3 January 2011
development of any drug or treatment, which might arise from the research and later goes on to make a profit.

**What are the risks to me of donating my fluids?**

There are no additional health risks associated with donating fluid samples for research purposes if they are taken as part of a normal diagnostic procedure. If we are taking a blood sample at a different time from your routine tests, the only risks would be minor bruising. If you are a patient and anything in the procedure for obtaining your samples were to go wrong, the normal complaint mechanisms of the NHS are open to you.

**What will my samples be used for?**

Your samples will be used in various research projects which will involve large-scale analysis of the proteins present in your fluids to help us understand the biology of your illness and develop new biomarker tests. None of these results will be passed back to you individually and they will be kept absolutely confidential.

**Could any of the results show that I have other illnesses?**

As the tests we carry out are for research purposes only we wouldn’t use the results for clinical purposes as there is not enough information available to allow us to do this.

**Can I withdraw my consent if I change my mind?**

Yes you can if the samples and/or data have not yet been used. Unused data and body fluids would, after your notice of withdrawal, be disposed of securely and respectfully. If you change your mind and your samples or data have been used, your

Leeds NIHR Biomarker RTB Renal Transplant Patient Information Booklet Version 1.3 January 2011
gift may have already contributed to new knowledge. This cannot be recalled. If you change your mind when you are still in hospital, you can ask a member of your clinical team to inform us on your behalf. If you change your mind later, or you would prefer not to approach us directly, you can write confidentially to our organisation’s Research and Development Dept who will ensure that your wishes are carried out. A standard letter has been given to you for this purpose.

**Who will know I am participating in the research?**

The only people who will know your identity are hospital staff and a limited number of staff at the Clinical Trials Research Unit where data is stored on secure computers. All are bound by a professional duty to protect your privacy. An identification number will be assigned to your samples, which ensures that researchers cannot identify you personally from your donation. This will be used in any other databases where details of your donated samples and associated information are stored.

**Will any of my personal information be used?**

We are asking for your permission for staff to access and use information from your clinical records, including those held electronically. The information we collect will only be that which is relevant to our research and will include general information such as age, gender, what kind of diet you eat or whether you smoke, as well as information more specifically about your illness such as pathology results, results of routine blood tests, any scan results and how you respond to different treatments. Access may start at the time you donate your samples and/or be required later e.g to look at your clinical progress. Before your information is released to researchers, it is anonymised keeping only an identification number. Participants’ identities will not be disclosed either to other researchers or when the results of the research are made public.
Who is funding the research?

This study is funded largely by the National Institute for Health Research (NIHR) but we also receive funding from other sources such as Cancer Research UK and the Medical Research Council. Occasionally, we may also receive collaborative grants from companies such as pharmaceutical or diagnostic companies, particularly where we are developing new diagnostic tests in partnership for example. These grants allow us to recover our costs, and any funds we receive in excess of our costs are used to fund further research.

Are there any other third parties involved in the research?

We may collaborate with other researchers in the UK or abroad. They may work in universities, hospitals or the private sector. Your samples will not, however, be sold for profit.

Who has reviewed the research?

Our research is reviewed by panels of experts associated with the various funding bodies and within academic research internationally. It is also reviewed by relevant ethics committees. This Research Tissue Bank has been approved by Leeds Research Ethics committee on 15th June 2010.

Will I get feedback from the research?

Any findings resulting from the research will be published in scientific or medical journals. Information will be available on the Leeds Teaching Hospitals research website, the NIHR Renal and Liver Biomarkers Programme website (www.biomarkerspipeline.org) and on the research group website (www.proteomics.leeds.ac.uk).
Donating to the wider research community?

Other research groups, within Leeds Teaching Hospitals, Leeds University (where the sample bank is based) or elsewhere, are also dependent on donations of body fluids to make progress. We would like you to consider whether you would like us to restrict the use of your samples to our research group (and those groups who we work with directly in collaboration), or whether you give us permission to share your samples and associated anonymous data with other research groups. Any project is reviewed by our Research Tissue Bank management committee, to make sure that it is scientifically sound and that it fits with the consent that you have given. We will not release any samples unless we are satisfied that our committee has approved the project and the research group has agreed to abide by our conditions. Please tell us what you decide in the consent form. There are costs involved in storing and sending samples, and we may ask external researchers to contribute to those costs, but we will not make a profit.

Other things to consider

Your samples may be used for research that involves:

- Export for use in research outside the UK
- Commercial research e.g. developing new tests

Your samples will not be used for research that involves:

- Research involving therapeutic/reproductive cloning (the latter remaining illegal under the Human Fertilisation and Embryology Authority 2001)
- Research involving human embryos and stem cells
- Research involving animal-human hybrid embryos
- Research into termination of pregnancy or contraception
Who to contact for Further Information:

If you would like further information you can either
- Ask the person who has provided this booklet to you
- Contact the Principal Investigator,
  Professor Peter Selby
  Email: p.j.selby@leeds.ac.uk

Thank you for reading this patient information sheet.
HEALTHY VOLUNTEERS IN INFORMATION SHEET

Leeds NIHR Biomarker Research Tissue Bank
We would like to invite you to take part in a research study. Before you decide whether you would like to take part, you need to understand why the research is being done and what it would involve for you. Please take time to read the following information carefully and discuss it with your family, friends or your GP if you wish. Please take the opportunity to put any questions you may have to a qualified and experienced person. If you decide that you are happy to take part, please sign the attached consent form and have this witnessed. Keep a copy of this information for future reference.

What kind of research is being done and why?

Testing human fluids such as urine and blood is necessary to understand how the human body works normally and what changes when things go wrong. We are currently carrying out research into several diseases. These include particularly, but not exclusively, diseases involving the kidney such as renal (kidney) cancer and kidney transplantation. The main purpose of this research is to develop new clinical tests that can identify measurable changes in proteins (“biomarkers”) in patient samples. These biomarker tests may be used to improve patient care such as helping to diagnose disease earlier or in deciding which drug is having the best effect in a patient.

What am I being asked to donate and what procedures are involved?

When we find new biomarkers it is important to know what normal levels are by comparing our results in patients with those in normal healthy volunteers or “controls”. We are therefore asking you to donate a blood and/or urine sample to allow us to do this.
We would ask your consent to take a small amount of blood (normally less than 20 mls or 5 teaspoons) and/or to provide a urine sample for our research. Often at the hospital or at your GP you will have a blood sample taken from you ("venepuncture") as a standard part of your clinical care. To get a sample for research involves exactly the same process. In all cases this will be carried out by a fully trained member of staff.

**Do I have to take part?**

No, the decision of whether or not to take part is completely up to you.

**What are the benefits or advantages of taking part?**

Research studies usually take many years to complete. You will be contributing to a bank of fluid samples, which may help to speed up research into human disease. The results of the research overall may benefit patients with a range of kidney diseases including renal cancer, renal transplantation and acute kidney injury in the future. In addition to contributing to generating new knowledge in medical research, it may also decrease the need to rely on testing on animals.

As an unconditional gift, the benefits of donating blood and urine samples are humanitarian rather than personal. You will not receive any financial reward, including from the successful development of any test or treatment, which might arise from the research and later goes on to make a profit.

**What are the risks to me of donating my fluids?**

There is a risk of minor bruising when taking blood samples. If anything in the procedure for obtaining your samples were to go wrong, the normal complaint mechanisms of the NHS are open to you.
What will my samples be used for?

Your samples will be used in various research projects which will involve large-scale analysis of the proteins present in your fluids to help us understand the biology of diseases and develop new biomarker tests. None of these results will be passed back to you individually and they will be kept absolutely confidential.

Could any of the results show that I have other illnesses?

As the tests we carry out are for research purposes only we wouldn’t use the results for clinical purposes as there is not enough information available to allow us to do this.

Can I withdraw my consent if I change my mind?

Yes you can if the samples and/or data have not yet been used. Unused data and samples would, after your notice of withdrawal, be disposed of securely and respectfully. If you change your mind and your samples or data have been used, your gift may have already contributed to new knowledge. This cannot be recalled. If you change your mind later, or you would prefer not to approach us directly, you can write confidentially to our organisation’s Research and Development Dept who will ensure that your wishes are carried out. A standard letter has been given to you for this purpose.

Who will know I am participating in the research?

The only people who will know your identity are hospital staff and a limited number of staff at the Clinical Trials Research Unit where the data is stored on secure computers. All are bound by a professional duty to protect your privacy. An identification number will be assigned to your samples, which ensures that researchers cannot identify you personally from your donation. This will be used in
any other databases where details of your samples are stored. A limited amount of information will be collected from you by the person obtaining your consent. This includes information such as your age, gender, brief medical history and details of lifestyle factors such as smoking and diet.

**Who is funding the research?**

This study is funded largely by the National Institute for Health Research (NIHR) but we also receive funding from other sources such as Cancer Research UK and the Medical Research Council. Occasionally, we may also receive collaborative grants from companies such as pharmaceutical or diagnostic companies, particularly where we are developing new diagnostic tests in partnership for example. These grants allow us to recover our costs, and any funds we receive in excess of our costs are used to fund further research.

**Are there any other third parties involved in the research?**

We may collaborate with other researchers in the UK or abroad. They may work in universities, hospitals or the private sector. Your samples will not, however, be sold for profit.

**Who has reviewed the research?**

Our research is reviewed by panels of experts associated with the various funding bodies and within academic research internationally. It is also reviewed by relevant ethics committees. This Research Tissue Bank has been approved by Leeds Research Ethics committee on (15th June 2010).

**Will I get feedback from the research?**

Any findings resulting from the research will be published in scientific or medical journals. Information will be available on the Leeds Teaching Hospitals research
website, the NIHR Renal and Liver Biomarkers Programme website (www.biomarkerspipeline.org) and on the research group website (www.proteomics.leeds.ac.uk).

**Donating to the wider research community?**

Other research groups, within Leeds Teaching Hospitals, Leeds University (where the sample bank is based) or elsewhere, are also dependent on donations of body fluids to make progress. We would like you to consider whether you would like us to restrict the use of your samples to our research group (and those groups who we work with directly in collaboration), or whether you give us permission to share your samples and associated anonymous data with other research groups. Any project is reviewed by our Research Tissue Bank management committee, to make sure that it is scientifically sound and that it fits with the consent that you have given. Please tell us what you decide in the consent form. There are costs involved in storing and sending samples, and we may ask external researchers to contribute to those costs, but we will not make a profit.

**Other things to consider**

Your samples may be used for research that involves:
- Export for use in research outside the UK
- Commercial research e.g. developing new tests

Your samples will **not** be used for research that involves
- Research involving therapeutic/reproductive cloning (the latter remaining illegal under the Human Fertilisation and Embryology Authority 2001)
- Research involving human embryos and stem cells
- Research involving animal-human hybrid embryos
- Research into termination of pregnancy or contraception

Leeds NIHR Biomarker RTB Volunteer Information Booklet Version 1.3 January 2011
Who to contact for Further Information:-

If you would like further information you can either
- Ask the person who has provided this booklet to you
- Contact the Principal Investigator,
  Professor Peter Selby
  Email: p.j.selby@leeds.ac.uk

Thank you for reading this information sheet.
Renal Cancer Marker Study Patient Consent Form
Leeds NIHR Biomarker Research Tissue Bank

Please indicate your understanding of the research study and your consent to take part by initialling (NOT ticking) each of the boxes below.

I have read and understand the patient information sheet "Leeds NIHR Biomarker RTB RCC Patient Information Sheet Version 1.3 January 2011" and have had the opportunity to ask questions. These have been answered clearly and satisfactorily and I understand the risks and benefits of donating my samples for research.

I give permission for my fluid and cell samples, and tissue samples which are not needed for diagnosis, to be collected and used in scientific research by the Leeds NIHR Biomarker Programme Group and their collaborators (including commercial companies), including / not including (delete as appropriate) in large scale genomic studies.

If any of the research findings provide other information which may be relevant to me personally or my relatives such as risk of other illnesses, I would / would not (delete as appropriate) like my GP to be contacted to investigate this further. If I can't be contacted I do / do not (delete as appropriate) give permission for you to contact my relatives about this.

Name of Contact: ____________________________________________

I do / I do not (delete as appropriate) give permission for my tissue and fluid samples to be shared with other research groups for projects approved by the Leeds NIHR Biomarker RTB Management Committee.

I agree for my details (which will include my name, date of birth, gender, NHS number, and postcode) to be registered with the Medical Research Information Service (MRIS) or traced via the NHS Information Service or relevant patient registries so that information about my health status may be obtained by researchers if necessary.

I give permission for this information about me, provided by me or found in my medical and other health related records to be supplied to and stored by researchers, including electronically, in an anonymous way that protects my identity. I understand that my anonymised samples and data may be shared on a collaborative basis with researchers in other UK centres and, potentially, centres abroad, including outside the European Economic Area (EEA).

I understand that:
• my participation is voluntary and that I am free to decline to give my consent or to withdraw from the study at any time without having to give a reason and that opting out at any stage has no bearing on my legal rights or subsequent medical treatment.

• if I withdraw consent, any samples and data which have already been used in research before that date cannot be withdrawn but unused samples will be disposed of respectfully and my data will no longer be used.

I understand and agree that I will not personally benefit, financially or medically, from my gift of tissue and fluid samples. This includes if my samples are involved in research leading to a new treatment or medical test.

I confirm that I offer my tissue and fluid samples as an unconditional gift and do not wish to place any restriction on the research that will be carried out on them, beyond the limits stated in the information which I have already read.

I agree to a copy of this Consent Form being sent to the Clinical Trials Research Unit (CTRU).

Patient's signature: ___________________________ Date: __________

Full name of patient (please print): ___________________________

Patient trial ID number: ___________________________

Signature of person taking consent: ___________________________ Date: __________

Full name of person taking consent (please print): ___________________________

Thank you for agreeing to take part in this research.
Renal Transplant Marker Study Patient Consent Form
Leeds NIHR Biomarker Research Tissue Bank

Please indicate your understanding of the research study and your consent to take part by initialing (NOT ticking) each of the boxes below.

Initial:

I have read and understand the patient information sheet "Leeds NIHR Biomarker RTB Renal Transplant Patient Information Sheet Version 1.3 January 2011" and have had the opportunity to ask questions. These have been answered clearly and satisfactorily and I understand the risks and benefits of donating my samples for research.

I give permission for my fluid and cell samples, and tissue samples which are not needed for diagnosis, to be collected and used in scientific research by the Leeds NIHR Biomarker Programme Group and their collaborators (including commercial companies).

I do / I do not (delete as appropriate) give permission for my tissue and fluid samples to be shared with other research groups for projects approved by the Leeds NIHR Biomarker RTB Management Committee.

I agree for my details (which will include my name, date of birth, gender, NHS number, and postcode) to be registered with the Medical Research Information Service (MRIS) or traced via the NHS Information Service or relevant patient registries so that information about my health status may be obtained by researchers if necessary.

I give permission for this information about me, provided by me or found in my medical and other health related records to be supplied to and stored by researchers, including electronically, in an anonymous way that protects my identity. I understand that my anonymised samples and data may be shared on a collaborative basis with researchers in other UK centres and, potentially, centres abroad, including outside the European Economic Area (EEA).

I understand that:

- my participation is voluntary and that I am free to decline to give my consent or to withdraw from the study at any time without having to give a reason and that opting out at any stage has no bearing on my legal rights or subsequent medical treatment.
- if I withdraw consent, any samples and data which have already been used in research before that date cannot be withdrawn but unused samples will be disposed of respectfully and my data will no longer be used.
- I understand and agree that I will not personally benefit, financially or medically, from my gift of tissue and fluid samples. This includes if my samples are involved in research leading to a new treatment or medical test.
- I confirm that I offer my tissue and fluid samples as an unconditional gift and do not wish to place any restriction on the research that will be carried out on them, beyond the limits stated in the information which I have already read.
- I agree to a copy of this Consent Form being sent to the Clinical Trials Research Unit (CTRU).

Patient's signature: ____________________________ Date: __________

Full name of patient (please print): ______________________________________

Patient trial ID number: __________________________

Signature of person taking consent: __________________________ Date: __________

Full name of person taking consent (please print): ______________________________________

Thank you for agreeing to take part in this research.

Leeds NIHR Biomarker RTB Renal Transplant Patient Consent Form Version 1.2 January 2011
Healthy Volunteer Consent Form
Leeds NIHR Biomarker Research Tissue Bank

Please indicate your understanding of the research study and your consent to take part by initialling (NOT ticking) each of the boxes below.

I have read and understand the patient information sheet “Leeds NIHR Biomarker RTB Healthy Volunteer Information Sheet Version 1.3 January 2011” and have had the opportunity to ask questions. These have been answered clearly and satisfactorily and I understand the risks and benefits of donating my samples for research.

I give permission for my fluid and cell samples, and tissue samples which are not needed for diagnosis, to be collected and used in scientific research by the Leeds NIHR Biomarker Programme Group and their collaborators (including commercial companies)

I do / I do not (delete as appropriate) give permission for my tissue and fluid samples to be shared with other research groups for projects approved by the Leeds NIHR Biomarker RTB Management Committee.

I give permission for this information about me, provided by me or found in my medical and other health related records to be supplied to and stored by researchers, including electronically, in an anonymous way that protects my identity. I understand that my anonymised samples and data may be shared on a collaborative basis with researchers in other UK centres and, potentially, centres abroad, including outside the European Economic Area (EEA).

I understand that:

- my participation is voluntary and that I am free to decline to give my consent or to withdraw from the study at any time without having to give a reason and that opting out at any stage has no bearing on my legal rights or subsequent medical treatment.
- if I withdraw consent, any samples and data which have already been used in research before that date cannot be withdrawn but unused samples will be disposed of respectfully and my data will no longer be used.
- I understand and agree that I will not personally benefit, financially or medically, from my gift of tissue and fluid samples. This includes if my samples are involved in research leading to a new treatment or medical test.
- I confirm that I offer my tissue and fluid samples as an unconditional gift and do not wish to place any restriction on the research that will be carried out on them, beyond the limits stated in the information which I have already read.
- I agree to a copy of this Consent Form being sent to the Clinical Trials Research Unit (CTRU).

Volunteer’s signature: __________________________ Date: _______________

Full name of Volunteer (please print): __________________________________________

Volunteer’s trial ID number: ________________________________________________

Signature of person taking consent: __________________________ Date: _______________

Full name of person taking consent (please print): ______________________________

Thank you for agreeing to take part in this research.

Leeds NIHR Biomarker RTB Healthy Volunteer Consent Form Version 1.2 January 2011
Appendix 2 - Appendices to chapter 17
Siemens ELF test - Assessment of between-site agreement – January 2016

Site: XXXXX

Please assay the ten specimens provided in two separate runs following instructions provided by the manufacturer. Results for the ten specimens and for all IQC specimens included in the run should then be entered in the tables below. Please add any additional relevant information about either run.

Results for Run 1.

Date and time of run: ______________________

<table>
<thead>
<tr>
<th>Specimen number</th>
<th>Hyaluronic acid (HA) (µg/L)</th>
<th>PIINP (µg/L)</th>
<th>TIMP-1 (µg/L)</th>
<th>ELF Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>E001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kit control 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kit control 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kit control 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Site: XXXXX
Results for Run 2.
Date and time of run: 

<table>
<thead>
<tr>
<th>Specimen number</th>
<th>Hyaluronic acid (HA) (µg/L)</th>
<th>PIIINP (µg/L)</th>
<th>TIMP-1 (µg/L)</th>
<th>ELF Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>E001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kit control 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kit control 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kit control 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please enter any additional comments below and return your results sheets as soon as possible by e-mail to ukneqas@ed.ac.uk or by fax to 0131 242 6882. Many thanks!
<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Site 1 HA (ug/L)</th>
<th>Site 2 HA (ug/L)</th>
<th>Site 3 HA (ug/L)</th>
<th>Site 4 HA (ug/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E001</td>
<td>87, 83, 85</td>
<td>83, 91, 87</td>
<td>88, 102, 95</td>
<td>88, 86, 87</td>
</tr>
<tr>
<td>E002</td>
<td>54, 51, 52</td>
<td>51, 52, 52</td>
<td>51, 59, 55</td>
<td>53, 51, 52</td>
</tr>
<tr>
<td>E003</td>
<td>76, 76, 76</td>
<td>76, 80, 78</td>
<td>78, 89, 83</td>
<td>70, 83, 76</td>
</tr>
<tr>
<td>E004</td>
<td>35, 36, 36</td>
<td>39, 39, 39</td>
<td>38, 44, 41</td>
<td>37, 36, 36</td>
</tr>
<tr>
<td>E005</td>
<td>51, 51, 51</td>
<td>51, 55, 53</td>
<td>55, 57, 56</td>
<td>50, 51, 50</td>
</tr>
<tr>
<td>E006</td>
<td>68, 65, 66</td>
<td>69, 73, 71</td>
<td>80, 86, 83</td>
<td>70, 72, 71</td>
</tr>
<tr>
<td>E007</td>
<td>41, 43, 42</td>
<td>42, 44, 43</td>
<td>44, 47, 46</td>
<td>41, 43, 42</td>
</tr>
<tr>
<td>E008</td>
<td>97, 100, 98</td>
<td>97, 102, 100</td>
<td>109, 100, 104</td>
<td>98, 99, 99</td>
</tr>
<tr>
<td>E009</td>
<td>65, 62, 64</td>
<td>65, 63, 64</td>
<td>71, 65, 68</td>
<td>66, 65, 65</td>
</tr>
<tr>
<td>E010</td>
<td>85, 95, 90</td>
<td>94, 102, 98</td>
<td>120, 103, 111</td>
<td>101, 101, 101</td>
</tr>
<tr>
<td>QC 1</td>
<td>20, 20, 20</td>
<td>22, 20, 21</td>
<td>25, 23, 24</td>
<td>21, 20, 21</td>
</tr>
<tr>
<td>QC 2</td>
<td>46, 43, 45</td>
<td>48, 47, 48</td>
<td>51, 52, 51</td>
<td>49, 46, 48</td>
</tr>
<tr>
<td>QC 3</td>
<td>175, 161, 168</td>
<td>180, 173, 176</td>
<td>191, 195, 193</td>
<td>169, 160, 164</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Site 5 HA (ug/L)</th>
<th>Site 6 HA (ug/L)</th>
<th>Site 7 HA (ug/L)</th>
<th>Site 8 HA (ug/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E001</td>
<td>84, 82, 83</td>
<td>121, 99, 110</td>
<td>84, 83, 83</td>
<td>83, 85, 84</td>
</tr>
<tr>
<td>E002</td>
<td>52, 49, 51</td>
<td>39, 63, 51</td>
<td>51, 51, 51</td>
<td>51, 53, 52</td>
</tr>
<tr>
<td>E003</td>
<td>73, 69, 71</td>
<td>105, 83, 94</td>
<td>74, 71, 72</td>
<td>73, 75, 74</td>
</tr>
<tr>
<td>E004</td>
<td>35, 37, 36</td>
<td>50, 27, 39</td>
<td>35, 36, 35</td>
<td>36, 36, 36</td>
</tr>
<tr>
<td>E005</td>
<td>47, 52, 49</td>
<td>64, 42, 53</td>
<td>49, 50, 50</td>
<td>50, 52, 51</td>
</tr>
<tr>
<td>E006</td>
<td>71, 66, 68</td>
<td>93, 71, 82</td>
<td>68, 69, 68</td>
<td>73, 73, 73</td>
</tr>
<tr>
<td>E007</td>
<td>41, 42, 42</td>
<td>60, 32, 46</td>
<td>42, 44, 43</td>
<td>43, 43, 43</td>
</tr>
<tr>
<td>E008</td>
<td>100, 101, 100</td>
<td>153, 146, 150</td>
<td>101, 104, 102</td>
<td>105, 103, 104</td>
</tr>
<tr>
<td>E009</td>
<td>60, 69, 64</td>
<td>45, 95, 70</td>
<td>68, 61, 65</td>
<td>65, 66, 66</td>
</tr>
<tr>
<td>E010</td>
<td>102, 105, 104</td>
<td>145, 114, 129</td>
<td>101, 101, 101</td>
<td>95, 109, 102</td>
</tr>
<tr>
<td>QC 1</td>
<td>21, 21, 21</td>
<td>24, 25, 25</td>
<td>20, 20, 20</td>
<td>23, 22, 22</td>
</tr>
<tr>
<td>QC 2</td>
<td>45, 43, 44</td>
<td>62, 52, 57</td>
<td>45, 44, 45</td>
<td>44, 44, 44</td>
</tr>
<tr>
<td>QC 3</td>
<td>156, 167, 162</td>
<td>141, 212, 177</td>
<td>173, 167, 170</td>
<td>167, 167, 167</td>
</tr>
<tr>
<td>Sample ID</td>
<td>Run 1</td>
<td>Run 2</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>E001</td>
<td>15.8</td>
<td>16.2</td>
<td>16.0</td>
<td>0.2</td>
</tr>
<tr>
<td>E002</td>
<td>9.4</td>
<td>9.5</td>
<td>9.4</td>
<td>0.0</td>
</tr>
<tr>
<td>E003</td>
<td>8.5</td>
<td>8.7</td>
<td>8.6</td>
<td>0.1</td>
</tr>
<tr>
<td>E004</td>
<td>12.6</td>
<td>12.8</td>
<td>12.7</td>
<td>0.2</td>
</tr>
<tr>
<td>E005</td>
<td>9.8</td>
<td>10.0</td>
<td>9.9</td>
<td>0.1</td>
</tr>
<tr>
<td>E006</td>
<td>10.2</td>
<td>10.3</td>
<td>10.2</td>
<td>0.1</td>
</tr>
<tr>
<td>E007</td>
<td>10.6</td>
<td>10.4</td>
<td>10.5</td>
<td>0.1</td>
</tr>
<tr>
<td>E008</td>
<td>22.5</td>
<td>22.0</td>
<td>22.2</td>
<td>0.4</td>
</tr>
<tr>
<td>E009</td>
<td>21.8</td>
<td>20.2</td>
<td>21.0</td>
<td>1.1</td>
</tr>
<tr>
<td>E010</td>
<td>10.2</td>
<td>9.9</td>
<td>10.0</td>
<td>0.2</td>
</tr>
<tr>
<td>QC 1</td>
<td>3.5</td>
<td>3.3</td>
<td>3.4</td>
<td>0.1</td>
</tr>
<tr>
<td>QC 2</td>
<td>6.5</td>
<td>6.2</td>
<td>6.3</td>
<td>0.2</td>
</tr>
<tr>
<td>QC 3</td>
<td>12.3</td>
<td>11.7</td>
<td>12.0</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Run 1</th>
<th>Run 2</th>
<th>Mean</th>
<th>SD</th>
<th>%CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>E001</td>
<td>16.3</td>
<td>16.7</td>
<td>16.5</td>
<td>0.2</td>
<td>1.5</td>
</tr>
<tr>
<td>E002</td>
<td>9.7</td>
<td>10.2</td>
<td>9.9</td>
<td>0.3</td>
<td>3.4</td>
</tr>
<tr>
<td>E003</td>
<td>9.3</td>
<td>8.7</td>
<td>9.0</td>
<td>0.4</td>
<td>4.5</td>
</tr>
<tr>
<td>E004</td>
<td>12.8</td>
<td>13.8</td>
<td>13.3</td>
<td>0.7</td>
<td>5.1</td>
</tr>
<tr>
<td>E005</td>
<td>9.8</td>
<td>11.2</td>
<td>10.5</td>
<td>0.9</td>
<td>9.0</td>
</tr>
<tr>
<td>E006</td>
<td>10.4</td>
<td>11.2</td>
<td>10.8</td>
<td>0.6</td>
<td>5.3</td>
</tr>
<tr>
<td>E007</td>
<td>11.1</td>
<td>11.3</td>
<td>11.2</td>
<td>0.1</td>
<td>1.1</td>
</tr>
<tr>
<td>E008</td>
<td>23.6</td>
<td>22.6</td>
<td>23.1</td>
<td>0.7</td>
<td>2.9</td>
</tr>
<tr>
<td>E009</td>
<td>21.1</td>
<td>21.7</td>
<td>21.4</td>
<td>0.4</td>
<td>1.9</td>
</tr>
<tr>
<td>E010</td>
<td>10.9</td>
<td>10.5</td>
<td>10.7</td>
<td>0.2</td>
<td>2.3</td>
</tr>
<tr>
<td>QC 1</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>QC 2</td>
<td>6.2</td>
<td>6.4</td>
<td>6.3</td>
<td>0.1</td>
<td>2.3</td>
</tr>
<tr>
<td>QC 3</td>
<td>11.9</td>
<td>12.9</td>
<td>12.4</td>
<td>0.7</td>
<td>5.9</td>
</tr>
</tbody>
</table>

Site 1 PIIINP (ug/L) Site 2 PIIINP (ug/L) Site 3 PIIINP (ug/L) Site 4 PIIINP (ug/L) Site 5 PIIINP (ug/L) Site 6 PIIINP (ug/L) Site 7 PIIINP (ug/L) Site 8 PIIINP (ug/L)
<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Site 1 TIMP-1 (ug/L)</th>
<th>Site 2 TIMP-1 (ug/L)</th>
<th>Site 3 TIMP-1 (ug/L)</th>
<th>Site 4 TIMP-1 (ug/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run</td>
<td>Mean</td>
<td>SD</td>
<td>%CV</td>
<td>Mean</td>
</tr>
<tr>
<td>Run 1</td>
<td>Run 2</td>
<td>Run 1</td>
<td>Run 2</td>
<td>Run 1</td>
</tr>
<tr>
<td>E001</td>
<td>270 261 265 6.2 2.3</td>
<td>286 280 283 4.1 1.5</td>
<td>266 264 265 1.4 0.5</td>
<td>282 274 278 5.7 2.0</td>
</tr>
<tr>
<td>E002</td>
<td>235 243 239 5.7 2.4</td>
<td>247 246 247 0.6 0.2</td>
<td>234 242 238 5.8 2.4</td>
<td>242 246 244 2.8 1.2</td>
</tr>
<tr>
<td>E003</td>
<td>253 251 252 1.8 0.7</td>
<td>271 268 270 2.5 0.9</td>
<td>254 262 258 5.2 2.0</td>
<td>262 270 266 5.8 2.2</td>
</tr>
<tr>
<td>E004</td>
<td>193 197 195 2.9 1.5</td>
<td>210 202 206 5.5 2.7</td>
<td>193 200 196 5.2 2.6</td>
<td>201 200 200 0.7 0.4</td>
</tr>
<tr>
<td>E005</td>
<td>240 231 235 5.9 2.5</td>
<td>252 250 251 1.5 0.6</td>
<td>236 245 240 5.9 2.4</td>
<td>246 242 244 3.0 1.2</td>
</tr>
<tr>
<td>E006</td>
<td>214 213 213 1.3 0.6</td>
<td>228 227 227 0.6 0.3</td>
<td>213 227 220 10.0 4.5</td>
<td>225 222 224 2.2 1.0</td>
</tr>
<tr>
<td>E007</td>
<td>294 299 297 3.4 1.1</td>
<td>310 314 312 2.8 0.9</td>
<td>290 298 294 5.3 1.8</td>
<td>307 298 302 6.1 2.0</td>
</tr>
<tr>
<td>E008</td>
<td>298 287 293 7.4 2.5</td>
<td>299 306 302 4.7 1.6</td>
<td>292 297 294 4.0 1.3</td>
<td>298 298 298 0.0 0.0</td>
</tr>
<tr>
<td>E009</td>
<td>244 236 240 5.5 2.3</td>
<td>256 253 255 2.1 0.8</td>
<td>243 252 247 6.0 2.4</td>
<td>252 254 253 1.6 0.6</td>
</tr>
<tr>
<td>E010</td>
<td>272 257 264 10.3 3.9</td>
<td>278 274 276 2.6 0.9</td>
<td>266 277 272 7.6 2.8</td>
<td>285 278 281 4.8 1.7</td>
</tr>
<tr>
<td>QC1</td>
<td>99 99 99 0.1 0.1</td>
<td>100 98 99 1.4 1.4</td>
<td>96 95 95 1.3 1.3</td>
<td>98 99 99 0.3 0.3</td>
</tr>
<tr>
<td>QC2</td>
<td>270 277 273 4.6 1.7</td>
<td>287 270 278 12.2 4.4</td>
<td>280 269 274 8.4 3.1</td>
<td>277 285 281 5.2 1.8</td>
</tr>
<tr>
<td>QC3</td>
<td>559 538 549 15.2 2.8</td>
<td>577 577 577 0.2 0.0</td>
<td>560 561 560 0.8 0.1</td>
<td>572 603 588 21.5 3.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Site 5 TIMP-1 (ug/L)</th>
<th>Site 6 TIMP-1 (ug/L)</th>
<th>Site 7 TIMP-1 (ug/L)</th>
<th>Site 8 TIMP-1 (ug/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run</td>
<td>Mean</td>
<td>SD</td>
<td>%CV</td>
<td>Mean</td>
</tr>
<tr>
<td>Run 1</td>
<td>Run 2</td>
<td>Run 1</td>
<td>Run 2</td>
<td>Run 1</td>
</tr>
<tr>
<td>E001</td>
<td>272 278 275 4.5 1.6</td>
<td>279 280 280 1.0 0.4</td>
<td>267 262 265 3.3 1.2</td>
<td>274 278 276 2.8 1.0</td>
</tr>
<tr>
<td>E002</td>
<td>239 238 239 0.9 0.4</td>
<td>246 247 246 0.1 0.1</td>
<td>235 230 232 3.2 1.4</td>
<td>234 240 237 4.0 1.7</td>
</tr>
<tr>
<td>E003</td>
<td>260 267 263 5.0 1.9</td>
<td>266 270 268 3.1 1.2</td>
<td>255 250 253 3.5 1.4</td>
<td>257 261 259 2.6 1.0</td>
</tr>
<tr>
<td>E004</td>
<td>194 202 198 5.2 2.6</td>
<td>206 199 202 4.7 2.3</td>
<td>190 187 189 2.1 1.1</td>
<td>196 200 198 2.5 1.3</td>
</tr>
<tr>
<td>E005</td>
<td>241 245 243 2.9 1.2</td>
<td>252 256 254 2.5 1.0</td>
<td>230 239 234 6.5 2.8</td>
<td>245 239 242 4.4 1.8</td>
</tr>
<tr>
<td>E006</td>
<td>221 223 222 1.8 0.8</td>
<td>224 233 228 6.0 2.6</td>
<td>217 216 216 1.3 0.6</td>
<td>219 223 221 3.3 1.5</td>
</tr>
<tr>
<td>E007</td>
<td>294 291 293 2.3 0.8</td>
<td>301 310 305 6.3 2.1</td>
<td>289 285 287 3.3 1.1</td>
<td>301 309 305 5.6 1.8</td>
</tr>
<tr>
<td>E008</td>
<td>293 295 294 1.0 0.3</td>
<td>300 302 301 1.2 0.4</td>
<td>287 291 289 2.8 1.0</td>
<td>281 293 287 8.4 2.9</td>
</tr>
<tr>
<td>E009</td>
<td>250 248 249 1.6 0.6</td>
<td>255 253 254 1.6 0.6</td>
<td>246 240 243 4.0 1.7</td>
<td>256 249 253 4.9 2.0</td>
</tr>
<tr>
<td>E010</td>
<td>272 271 272 1.2 0.4</td>
<td>279 281 280 1.0 0.4</td>
<td>263 263 263 0.1 0.1</td>
<td>275 263 269 8.4 3.1</td>
</tr>
<tr>
<td>QC1</td>
<td>94 101 98 5.1 5.2</td>
<td>97 100 99 2.2 2.2</td>
<td>96 96 96 0.1 0.1</td>
<td>100 101 100 0.3 0.3</td>
</tr>
<tr>
<td>QC2</td>
<td>272 274 273 1.3 0.5</td>
<td>278 275 276 2.4 0.9</td>
<td>269 267 268 1.4 0.5</td>
<td>272 286 279 9.6 3.4</td>
</tr>
<tr>
<td>QC3</td>
<td>565 579 572 10.3 1.8</td>
<td>582 584 583 1.2 0.2</td>
<td>573 560 567 9.0 1.6</td>
<td>583 579 581 2.4 0.4</td>
</tr>
</tbody>
</table>
Sample
ID
E001
E002
E003
E004
E005
E006
E007
E008
E009
E010
QC 1
QC 2
QC 3

Site 1 ELF score
Run Run
1
2
Mean
10.4 10.3 10.3
9.5
9.5
9.5
9.8
9.8
9.8
9.3
9.3
9.3
9.5
9.5
9.5
9.7
9.7
9.7
9.5
9.5
9.5
10.8 10.8 10.8
10.3 10.2 10.3
10.0 10.1 10.0
7.6
7.5
7.6
9.2
9.1
9.1
11.1 10.9 11.0

SD

%CV

0.03
0.01
0.01
0.02
0.01
0.03
0.02
0.00
0.08
0.04
0.03
0.06
0.08

0.3
0.1
0.1
0.2
0.1
0.3
0.2
0.0
0.8
0.4
0.4
0.6
0.8

Site 5 ELF score
Sample
ID
E001
E002
E003
E004
E005
E006
E007
E008
E009
E010
QC 1
QC 2
QC 3

Run
1
10.4
9.5
9.8
9.3
9.4
9.8
9.5
10.8
10.2
10.2
7.5
9.1
10.9

Run
2
10.4
9.5
9.7
9.4
9.6
9.8
9.5
10.8
10.4
10.2
7.6
9.1
11.1

Site 2 ELF score
Run Run
1
2
Mean
10.4 10.5 10.4
9.5
9.5
9.5
9.8
9.9
9.8
9.4
9.4
9.4
9.6
9.6
9.6
9.8
9.9
9.8
9.5
9.6
9.6
10.8 10.8 10.8
10.3 10.3 10.3
10.1 10.2 10.2
7.7
7.6
7.6
9.2
9.2
9.2
11.1 11.1 11.1

SD

%CV

0.04
0.00
0.02
0.01
0.03
0.04
0.06
0.05
0.02
0.05
0.05
0.03
0.03

0.4
0.0
0.2
0.1
0.3
0.4
0.6
0.5
0.2
0.5
0.6
0.3
0.2

Site 6 ELF score
Mean
10.4
9.5
9.8
9.4
9.5
9.8
9.5
10.8
10.3
10.2
7.6
9.1
11.0

SD

%CV

0.0
0.0
0.1
0.1
0.1
0.0
0.0
0.0
0.1
0.0
0.0
0.0
0.1

0.0
0.1
0.6
0.8
1.4
0.1
0.1
0.1
0.9
0.1
0.5
0.2
0.8

Run
1
10.7
9.3
10.1
9.6
9.8
10.1
9.8
11.2
10.0
10.5
7.7
9.4
10.9

Run
2
10.5
9.7
9.9
9.1
9.4
9.8
9.3
11.1
10.6
10.3
7.7
9.2
11.2

Mean
10.6
9.5
10.0
9.4
9.6
10.0
9.6
11.1
10.3
10.4
7.7
9.3
11.1

Site 3 ELF score
Run Run
1
2
Mean
10.4 10.5 10.4
9.5
9.6
9.6
9.8
9.9
9.9
9.4
9.5
9.4
9.6
9.7
9.6
9.9 10.0
9.9
9.5
9.6
9.6
10.9 10.8 10.8
10.4 10.3 10.4
10.3 10.2 10.3
7.7
7.7
7.7
9.2
9.2
9.2
11.1 11.2 11.1

SD

%CV

0.1
0.1
0.1
0.1
0.0
0.1
0.0
0.1
0.0
0.1
0.0
0.0
0.0

0.6
1.2
0.9
1.1
0.4
0.6
0.5
0.7
0.4
0.6
0.6
0.0
0.3

Site 7 ELF score
SD

%CV

0.1
0.3
0.1
0.4
0.3
0.2
0.4
0.0
0.5
0.1
0.1
0.1
0.2

1.1
3.0
1.4
3.9
2.7
1.5
3.8
0.1
4.4
1.4
0.7
1.3
2.2

Run
1
10.4
9.5
9.8
9.4
9.6
9.8
9.5
10.8
10.3
10.2
7.6
9.1
11.0

Run
2
10.4
9.5
9.8
9.4
9.6
9.9
9.6
10.8
10.3
10.3
7.6
9.1
11.0

Mean
10.4
9.5
9.8
9.4
9.6
9.8
9.5
10.8
10.3
10.2
7.6
9.1
11.0

Site 4 ELF score
Run Run
1
2
Mean
10.4 10.4 10.4
9.5
9.5
9.5
9.7
9.9
9.8
9.3
9.3
9.3
9.5
9.5
9.5
9.8
9.8
9.8
9.4
9.5
9.5
10.7 10.7 10.7
10.3 10.2 10.3
10.2 10.1 10.2
7.6
7.6
7.6
9.2
9.1
9.2
11.0 11.0 11.0

SD

%CV

0
0
0.1
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.1
0.0

0
0
1.4
0.2
0.1
0.1
0.3
0.1
0.4
0.3
0.2
0.6
0.3

SD

%CV

0.03
0.01
0.04
0.02
0.02
0.01
0.04
0.01
0.05
0.01
0.00
0.00
0.02

0.3
0.1
0.4
0.2
0.2
0.1
0.4
0.1
0.4
0.1
0.0
0.0
0.2

Site 8 ELF score
SD

%CV

0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.1
0.0
0.0
0.0

0.1
0.1
0.2
0.0
0.1
0.1
0.2
0.1
0.0
0.8
0.0
0.0
0.0

Run
1
10.4
9.5
9.8
9.3
9.5
9.8
9.5
10.8
10.3
10.2
7.6
9.1
11.0

Run
2
10.3
9.5
9.7
9.3
9.5
9.8
9.5
10.8
10.3
10.2
7.6
9.1
11.0

Mean
10.3
9.5
9.8
9.3
9.5
9.8
9.5
10.8
10.3
10.2
7.6
9.1
11.0

726


Appendix 3 – summary of changes to the original ELUCIDATE protocol

<table>
<thead>
<tr>
<th>Protocol version</th>
<th>Date approved</th>
<th>Summary of amendment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version 1.0</td>
<td>02/02/2010</td>
<td>n/a</td>
</tr>
</tbody>
</table>
| Version 2.0      | 25/03/2010    | • Onset of grade 3 or 4 encephalopathy added as an endpoint so patients can be censored at this timepoint  
• Mortality due to HCC expanded to include further liver related mortalities  
• Unresectable HCC moved from 2ndry to 1ry endpoint |
| Version 3.0      | 30/04/2010    | • Timing of optional biobank sample changed from registration to randomisation |
| Recruitment halted between 24/12/2010 and 29/03/2011 whilst protocol amendment approved. 43 patients registered up until 24/12/2010. |
| Version 4.0      | Not approved due to concerns regarding information provision for re-approached patients who had previously failed the eligibility criteria. | • ELF thresholds updated to 8.4 (previously 11.0) for eligibility and 9.5 (previously 12.5) for a diagnosis of cirrhosis |
| Version 5.0      | 14/03/2011    | • ELF thresholds updated to 8.4 (previously 11.0) for eligibility and 9.5 (previously 12.5) for a diagnosis of cirrhosis  
• Additional co-primary endpoint added – time from randomisation to first severe complication  
• Original primary endpoint amended to: time from diagnosis of cirrhosis (by ELF or clinical means) to |
<table>
<thead>
<tr>
<th>Version 6.0</th>
<th>30/05/2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Eligibility criteria re-worded for clarity (patient population unchanged)</td>
<td></td>
</tr>
<tr>
<td>• Time from diagnosis of cirrhosis to first severe complication changed from a co-primary endpoint to a secondary endpoint</td>
<td></td>
</tr>
<tr>
<td>• Definition of hepatocellular cancer as a severe complication amended to HCC beyond the Milan criteria (previously beyond the extended Milan criteria)</td>
<td></td>
</tr>
<tr>
<td>• Definition of liver-related mortality (severe complication) amended to also include death due to spontaneous bacterial peritonitis, and death due to encephalopathy</td>
<td></td>
</tr>
<tr>
<td>• Secondary endpoint changed to specific liver-related morbidity (previously mortality - typographical error)</td>
<td></td>
</tr>
<tr>
<td>• Clarified that recruitment will be 24 months*, with an additional 30 months follow-up, and an additional 39 months long-term follow-up for the primary endpoint (taking it to 5 years after the end of the programme grant).</td>
<td></td>
</tr>
</tbody>
</table>

incidence of first severe complication

• Follow-up period extended to 5 years beyond the end of the NIHR programme grant

• Introduction of patient guidelines for QoL questionnaire completion

• Amendment to sample size and power calculations (sample size remained unchanged)

• Minimisation categories for baseline ELF amended to ≥ 8.4 to < 9.5, ≥ 9.5 to < 11.5 ≥ 11.5 to < 12.5 and ≥ 12.5; (previously 11-11.49, 11.5-11.99, 12-12.49, 12.5+)

• Analysis details amended to reflect above changes
<table>
<thead>
<tr>
<th>Version 7.0</th>
<th>04/03/2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Sample size and statistical analysis sections amended to take account of extended recruitment and follow-up duration and the new single primary endpoint</td>
<td></td>
</tr>
<tr>
<td>• Recruitment closure ahead of the pre-planned 1000 patients</td>
<td></td>
</tr>
<tr>
<td>• Addition of a new secondary endpoint (process outcomes), namely treatment with beta blockers/band ligation of varices, use of endoscopy and ultrasound/alpha-fetoprotein tests, and treatment to normalise liver function tests in patients with Hepatitis B and Hepatitis C.</td>
<td></td>
</tr>
<tr>
<td>• Clarification that if an ELF sample has been at room temperature for more than 2 days before receipt at the central laboratory, a repeat sample will be required.</td>
<td></td>
</tr>
<tr>
<td>• Change to the acceptable time window between registration and randomisation.</td>
<td></td>
</tr>
<tr>
<td>• Clarification that if &gt;12 weeks has passed between the registration ELF sample and randomisation, a repeat blood sample should be taken for ELF testing, to ensure that the patient remains eligible for randomisation.</td>
<td></td>
</tr>
<tr>
<td>• Clarification that variation of plus or minus one month around the visit due date is permitted.</td>
<td></td>
</tr>
<tr>
<td>• Timing of follow-up visits following a diagnosis of cirrhosis amended to 3 months post cirrhosis diagnosis and 6 monthly thereafter (previously 3 monthly from diagnosis of cirrhosis)</td>
<td></td>
</tr>
</tbody>
</table>
| • Post cirrhosis monitoring assessments (OGD, USS, AFP) previously described as mandatory at 3 months post diagnosis of cirrhosis are now only mandatory at this time point if they have not been
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>performed within a specified time window prior to diagnosis of cirrhosis.</td>
</tr>
<tr>
<td></td>
<td>• Clarification that screening / non-registration logs after the closure of the trial will no longer be required.</td>
</tr>
<tr>
<td></td>
<td>• Confirmation that some surplus samples from the central laboratory will be sent off for quality assurance testing.</td>
</tr>
<tr>
<td></td>
<td>• Clarification that baseline samples are acceptable within 1 month prior to registration (previously 14 days).</td>
</tr>
<tr>
<td></td>
<td>• Clarification that variation of follow-up visits within 1 month before or after the scheduled visit date is permitted.</td>
</tr>
<tr>
<td></td>
<td>• Definition of end of trial amended to date the last patient's last data item is collected (previously last follow-up visit).</td>
</tr>
<tr>
<td></td>
<td>• Sample size section updated to include the power based on an assumption that approximately 700 patients will have been randomised when the trial closes.</td>
</tr>
<tr>
<td></td>
<td>• Interim analysis section amended to include the new process outcomes endpoint and the follow-up until July 2014.</td>
</tr>
</tbody>
</table>
Appendix 4 - Original Sample size calculation for the ELUCIDATE Trial

Note that when these figures were derived, prior to the start of the trial, the ELF threshold for defining cirrhosis was set at 11, and therefore there would have been many less patients defined as cirrhotic by ELF. The threshold was changed to 9.5 after the recruitment of 3 patients in March 2011 when it was realised that the threshold of 11 was too stringent.

The RCT is testing the hypothesis that if we monitor patients with chronic liver disease using the ELF score, we will detect liver cirrhosis earlier, and as a result there will be fewer severe complications as well as improvements in other important patient outcomes. We are therefore aiming to show that the incidence of severe complications following cirrhosis is less in the ELF arm. This is the primary end point on which the trial is powered.

The trial will recruit over 18 months with an additional 30 months follow up. Previous studies have led us to anticipate that in the standard arm, at 36 months, we will observe severe complications of the order of 2% variceal bleeds, 1% major bleeds and ¾% inoperable hepatocellular cancers, giving a 3 ¾% incidence of potentially preventable undesirable clinical endpoints.¹

Previous data has illustrated that approximately 20% of patients will have varices suitable for therapy.² Such therapy has a large effect on the progression of these varices (for instance a reduction from 37% progressing to 11%); on the risk of bleeding from these varices (reduced from 30% to 14% in patients with medium to large varices); and on mortality at 36 months (reduced from 7% to 2% over 24 months).² Based on this prior data we hypothesise that we could reduce the incidence of the undesirable clinical endpoints of cirrhosis by a half, or even two-thirds in the ELF arm.

We anticipate that we will observe approximately twice as many patients developing cirrhosis in the ELF arm (approx 20% of patients) compared to the standard arm (approx 10% of patients) over the 18 months of recruitment and 30 months of follow up. Using this accrual and follow up rate, along with the expected incidence rate, we can calculate the expected number of events in each arm using the method described by Collett for sample size estimates based on exponential survival distributions (more details of which can be seen in the SAP).³ This gives expected numbers with severe complications of 19 and 6 in the control and experimental arms respectively, assuming a 2/3 reduction in severe complications; or 12 and 6 events respectively assuming a reduction in severe complications of one half.

With 1000 patients randomised, we would have > 99% power to detect this difference in numbers of patients encountering severe complications, subsequent to being detected with cirrhosis, with a 5%
type I error, 18 months of recruitment and 30 months follow up. Sample size calculations follow as below.

If it is assumed for simplicity that cirrhosis will be detected at uniform intervals throughout the 4 years of the trial, then we can estimate a pair of actuarial incidence curves (assuming exponentiality in generating these incidence curves) that gives us the 19 and 6 events that we originally estimated, with 18 months of recruitment and 30 months of follow-up. These hypothetical curves are presented below.

The exponential parameters give medians of 2.85 years in the control arm and 27 years in the ELF arm (consistent with such a big, two-thirds, reduction in severe complications). Doing sample size calculations with these figures gives a total sample size say $n_s$ (within cirrhosis patients) of 150 (>99% power). For this calculation we are assuming that cirrhosis patients are identified over an 18 month period with a further 30 months of follow up, and that twice as many are identified in the ELF arm (20% of the total in the ELF arm vs. 10% of the total in the control arm). The power calculation for severe complications allows for this. If the total sample size of cirrhotic patients = $n_c$, then $n_c = 20\% \left(1/2n_c\right) + 10\% \left(1/2n_c\right) = 15\% (n_c)$. So $n_c = 100/15 n_c$. We therefore multiply 150 by 100/15 to get the total sample size of monitored patients which comes to 1000 (>99% power).
Note that we have been plotting the likely incidence of severe complications following detection of cirrhosis, so that, even though the curve of the cumulative numbers detected with cirrhosis may have a shape which is concave then linear, with a slow start because of the initial 18-month accrual, the incidence of severe complications following cirrhosis might reasonably be expected to follow the exponential pattern assumed, where the likelihood of a severe complication that follows detection of cirrhosis is unrelated to the time followed up after detection of cirrhosis.

There is a possibility that the trial itself will affect the control arm (contamination) positively. We have therefore gone on to look at this contamination issue. Suppose the control arm gets more successful interventions etc. and the difference in severe complications is reduced by only a half (instead of two-thirds). So we assume that there are in fact only 12 events in the control arm instead of 19, as in the following graph.

The exponential parameters in this case give medians of 5.5 years in the control arm and 27 years again in the ELF arm. Doing sample size calculations with these figures gives a total sample size (within cirrhosis patients) of 143 (80% power) or 191 (90% power). We again multiply these by 100/15 to get the total sample size of monitored patients to be 953 (80% power) or 1273 (90% power).

This scenario might not be appropriate if we had a sizeable number of patients getting severe complications before their cirrhosis is detected. However, if this happened we would treat any patients having severe complications without prior detection of cirrhosis as having a 0 time-to-incidence of
severe complications following cirrhosis on the incidence curve. If this was similar in both arms the curves would simply start at the same probability (> 0) on the y axis, and subsequent divergence of the curves would be more acute than previously to cause the overall 2-fold difference we are looking for.

The trial as its current size is thus well powered (at least 80% even under the assumption that we only see a one half reduction in the incidence of severe complications) to show that this ELF monitoring policy will be of real clinical benefit.


