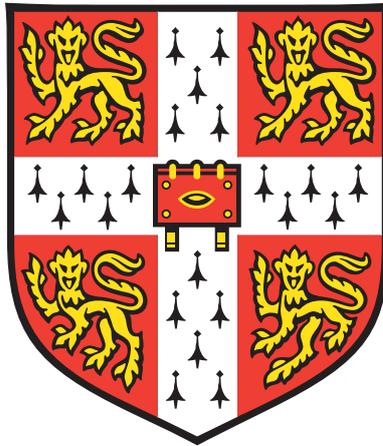


University of Cambridge



Modelling non-linear exposure–disease relationships  
in a large individual participant meta-analysis  
allowing for the effects of exposure measurement error

Alexander Daniel Strawbridge

Robinson College

November 2011

*This dissertation is submitted for the degree of Doctor of Philosophy*



*This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the Acknowledgements.*



## Acknowledgements

The simulation study of chapter 3 and its extension in chapter 5 to consider MacMahon's method features in the paper *Effects of classical exposure measurement error on the shape of exposure–disease associations* [1] which is jointly authored by Ruth Keogh and Ian White and will appear in the first issue of the journal *Epidemiologic Methods*. I contributed to the design of, and conducted, the simulation study, produced the figures, and provided detailed comments on drafts of the manuscript.

Simulation study 3 of chapter 5 features in the paper *Correcting for bias due to misclassification when error-prone continuous exposures are categorized* which will also appear in *Epidemiologic Methods* [2] and is jointly authored by Ruth Keogh and Ian White. I conducted the simulation study, devised the group-SIMEX method, and provided detailed comments on the manuscript.

I would like to express my deepest gratitude to my supervisor Ian White for his continued help and guidance throughout my research; and to my advisors Ruth Keogh, Emanuele Di Angelantonio, Angela Wood and Simon Thompson. I would like to thank the Emerging Risk Factors Collaboration and the constituent studies for providing me with the data that has provided the motivation for this thesis. I thank the Medical Research Council (MRC) for providing funding for this research project. I would also like to thank all of the staff at the MRC Biostatistics Unit in Cambridge, and the other PhD students, for making the period during which I have completed this thesis so memorable. Finally, I would like to thank Becky for being there for me, and making me smile.



## Summary

This thesis was motivated by data from the Emerging Risk Factors Collaboration (ERFC), a large individual participant data (IPD) meta-analysis of risk factors for coronary heart disease (CHD). Cardiovascular disease is the largest cause of death in almost all countries in the world, therefore it is important to be able to characterise the shape of risk factor–CHD relationships. Many of the risk factors for CHD considered by the ERFC are subject to substantial measurement error, and their relationship with CHD non-linear. We firstly consider issues associated with modelling the risk factor–disease relationship in a single study, before using meta-analysis to combine relationships across studies.

It is well known that classical measurement error generally attenuates linear exposure–disease relationships, however its precise effect on non-linear relationships is less well understood. We investigate the effect of classical measurement error on the shape of exposure–disease relationships that are commonly encountered in epidemiological studies, and then consider methods for correcting for classical measurement error. We propose the application of a widely used correction method, regression calibration, to fractional polynomial models. We also consider the effects of non-classical error on the observed exposure–disease relationship, and the impact on our correction methods when we erroneously assume classical measurement error.

Analyses performed using categorised continuous exposures are common in epidemiology. We show that MacMahon’s method for correcting for measurement error in analyses that use categorised continuous exposures, although simple, does not provide the correct shape for non-linear exposure–disease relationships. We perform a simulation study to compare alternative methods for categorised continuous exposures.

Meta-analysis is the statistical synthesis of results from a number of studies addressing similar research hypotheses. The use of IPD is the gold standard approach because it allows for consistent analysis of the exposure–disease relationship across studies. Methods have recently been proposed for combining non-linear relationships across studies. We discuss these methods, extend them to P-spline models, and consider alternative methods of combining relationships across studies.

We apply the methods developed to the relationships of fasting blood glucose and lipoprotein(a) with CHD, using data from the ERFC.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Epidemiology . . . . .	2
1.3	Measurement error . . . . .	4
1.3.1	The effects of measurement error . . . . .	4
1.3.2	Correction for measurement error . . . . .	5
1.4	Meta-analysis . . . . .	6
1.5	Overview of dissertation . . . . .	7
1.5.1	Chapter structure . . . . .	8
<b>2</b>	<b>Modelling non-linear exposure–disease relationships</b>	<b>9</b>
2.1	The Cox model . . . . .	9
2.1.1	Introduction to survival analysis . . . . .	9
2.1.2	Cox proportional hazards model . . . . .	11
2.1.3	Checking the proportional hazards assumption . . . . .	14
2.1.4	Checking the model fit . . . . .	15
2.1.5	Alternatives to the Cox model . . . . .	16
2.2	Functional forms for the linear predictor . . . . .	17
2.2.1	Grouped exposure analysis . . . . .	17
2.2.2	Polynomials . . . . .	19
2.2.3	Fractional polynomials . . . . .	20
2.2.4	Splines . . . . .	21
2.2.5	Other methods for modelling continuous non-linear exposure–disease relationships . . . . .	28
2.2.6	Discussion on choice of the linear predictor . . . . .	29
2.2.7	Confounding . . . . .	33
2.3	Conclusion . . . . .	34
<b>3</b>	<b>The effects of exposure measurement error</b>	<b>35</b>
3.1	Measurement error and misclassification . . . . .	35
3.1.1	What is the ‘truth’? . . . . .	37

3.1.2	Error models . . . . .	38
3.1.3	The effects of measurement error . . . . .	41
3.2	Shape of the exposure–disease relationship . . . . .	43
3.3	Methods for simulation . . . . .	45
3.3.1	Generating survival data . . . . .	45
3.3.2	Determining the degree of non-linearity . . . . .	46
3.4	Simulation Study . . . . .	47
3.4.1	Data generation . . . . .	47
3.4.2	Statistical methods evaluated . . . . .	49
3.4.3	Results . . . . .	51
3.5	Conclusion . . . . .	54
<b>4</b>	<b>Methods of correcting for exposure measurement error</b>	<b>59</b>
4.1	An extra piece of information . . . . .	60
4.2	Classification of methods . . . . .	61
4.3	Correction methods for continuous exposures . . . . .	61
4.3.1	Regression Calibration . . . . .	61
4.3.2	Structural fractional polynomials . . . . .	66
4.3.3	Corrected score function . . . . .	69
4.3.4	SIMulation EXtrapolation—SIMEX . . . . .	71
4.3.5	Multiple imputation . . . . .	77
4.3.6	Moment reconstruction . . . . .	78
4.3.7	Density estimation of the true exposure . . . . .	79
4.3.8	Bayesian approaches . . . . .	80
4.4	Correction methods for grouped data . . . . .	80
4.4.1	MacMahon’s method . . . . .	81
4.4.2	MisClassification SIMEX—MCSIMEX . . . . .	81
4.4.3	Group-SIMEX . . . . .	82
4.4.4	Natarajan’s method . . . . .	83
4.5	Calculation of confidence intervals . . . . .	84
4.6	Conclusion . . . . .	84
<b>5</b>	<b>Performance of correction methods</b>	<b>87</b>
5.1	Evaluation of MacMahon’s method . . . . .	87
5.1.1	MacMahon’s method . . . . .	87
5.1.2	Simulation procedure . . . . .	88
5.1.3	Results . . . . .	88
5.1.4	Discussion . . . . .	88
5.2	Correction methods for categorised continuous exposures . . . . .	91
5.2.1	Methods . . . . .	93

5.2.2	Simulation procedure . . . . .	97
5.2.3	Results . . . . .	98
5.2.4	Discussion . . . . .	98
5.3	Comparison of structural fractional polynomials & P-splines . . . . .	105
5.3.1	Methods . . . . .	105
5.3.2	Simulation procedure . . . . .	108
5.3.3	Results . . . . .	108
5.3.4	Discussion . . . . .	110
5.4	Conclusion . . . . .	111
<b>6</b>	<b>Application — Emerging Risk Factors Collaboration</b>	<b>117</b>
6.1	ERFC — Emerging Risk Factors Collaboration . . . . .	117
6.2	Background — FBG and CHD . . . . .	119
6.3	ERFC fasting blood glucose dataset . . . . .	120
6.4	Modelling — Ignoring the effects of measurement error . . . . .	123
6.4.1	Grouped exposure analysis . . . . .	123
6.4.2	Fractional polynomial and P-spline models . . . . .	124
6.5	Measurement error in FBG . . . . .	126
6.6	Correcting for measurement error . . . . .	130
6.6.1	MacMahon’s method . . . . .	130
6.6.2	SIMEX . . . . .	130
6.6.3	Structural fractional polynomial and P-spline models for FBG . . . . .	133
6.7	Discussion . . . . .	135
6.8	Conclusion . . . . .	139
<b>7</b>	<b>Non-classical measurement error</b>	<b>141</b>
7.1	Introduction . . . . .	141
7.1.1	Introduction of non-linearity . . . . .	142
7.2	Diagnosing non-classical error . . . . .	144
7.3	Simulation study . . . . .	145
7.3.1	Simulation procedure . . . . .	145
7.3.2	Results . . . . .	147
7.3.3	Discussion . . . . .	154
7.4	Correcting for non-classical measurement error . . . . .	158
7.4.1	Methods proposed for correcting for non-classical error . . . . .	158
7.4.2	Mixture regression calibration modelling . . . . .	159
7.4.3	Application — ERFC FBG data . . . . .	160
7.5	Conclusion . . . . .	161
<b>8</b>	<b>Meta-analysis</b>	<b>165</b>

8.1	Evidence synthesis . . . . .	165
8.2	Advantages/disadvantages of meta-analysis . . . . .	166
8.3	Univariate meta-analysis . . . . .	167
8.3.1	Fixed effect meta-analysis . . . . .	167
8.3.2	Random effects meta-analysis . . . . .	168
8.3.3	DerSimonian and Laird estimate of $\tau^2$ . . . . .	168
8.3.4	Other measures of $\tau^2$ . . . . .	170
8.4	Graphs for meta-analysis . . . . .	170
8.4.1	Forest plot . . . . .	170
8.4.2	Funnel plot . . . . .	171
8.4.3	Galbraith plot . . . . .	171
8.5	Multivariate meta-analysis . . . . .	174
8.6	Individual participant data meta-analysis . . . . .	176
8.6.1	Advantages and disadvantages of IPD meta-analysis . . . . .	177
8.6.2	Approaches to IPD meta-analysis . . . . .	178
8.7	IPD meta-analysis for continuous exposure measures . . . . .	178
8.7.1	Multivariate meta-analysis of point estimates . . . . .	179
8.7.2	Continuous approach . . . . .	179
8.8	Application of methods to the ERFC FBG data . . . . .	189
8.8.1	Relationship uncorrected for the effects of measurement error . . . . .	189
8.8.2	Relationship corrected for the effects of measurement error . . . . .	193
8.9	Discussion . . . . .	195
8.10	Conclusion . . . . .	196
<b>9</b>	<b>Analysis of ERFC Lipoprotein(a) dataset</b>	<b>201</b>
9.1	Background — Lp(a) and CHD . . . . .	201
9.2	The ERFC Lp(a) dataset . . . . .	202
9.2.1	Measurement error . . . . .	204
9.3	Analysis of the Lp(a)–CHD relationship . . . . .	205
9.3.1	Grouped exposure analysis . . . . .	205
9.3.2	Continuous exposure analysis . . . . .	209
9.4	Discussion . . . . .	215
9.5	Conclusion . . . . .	216
<b>10</b>	<b>Discussion</b>	<b>217</b>
10.1	Dissertation summary . . . . .	217
10.2	Modelling non-linear exposure–disease relationships . . . . .	219
10.3	Measurement error . . . . .	220
10.3.1	Effects of measurement error . . . . .	220
10.3.2	Correction for measurement error . . . . .	221

10.3.3	Final remarks on measurement error . . . . .	223
10.4	Meta-analysis . . . . .	224
10.5	Limitations . . . . .	225
10.6	Areas for future work . . . . .	226
10.6.1	Modelling non-linear exposure–disease relationships . . . . .	226
10.6.2	Effects of measurement error . . . . .	226
10.6.3	Measurement error correction . . . . .	227
10.6.4	Meta-analysis . . . . .	227
10.7	Conclusion . . . . .	228
<b>A</b>	<b>Appendix to chapter 4</b>	<b>229</b>
A.1	Calculation of basis functions for structural P-splines when $X W$ is normally distributed . . . . .	229
A.2	Calculation of $E(X^p \log X)$ . . . . .	231
A.3	Proof that $\lim_{\zeta \rightarrow -1} \text{Var}(\hat{\beta}_b(\zeta)) = 0$ . . . . .	231
<b>B</b>	<b>Appendix to chapter 5</b>	<b>233</b>
B.1	Simulation study 2 . . . . .	233
B.1.1	Results for linear regression . . . . .	233
B.1.2	Results for multiple imputation and moment reconstruction when the exposure is grouped into quintiles . . . . .	233
B.2	Simulation study 3: RFrMSE . . . . .	234
<b>C</b>	<b>Appendix to chapter 6</b>	<b>245</b>
C.1	ERFC FBG data — study names . . . . .	245
<b>D</b>	<b>Appendix to chapter 7</b>	<b>247</b>
D.1	Example measurement error plots for scenarios 2,4,5,8-10 using log-exposure . . . . .	247
<b>E</b>	<b>Appendix to chapter 9</b>	<b>249</b>
E.1	ERFC $L_p(a)$ data — study names . . . . .	249
<b>F</b>	<b>R code</b>	<b>251</b>
F.1	Code to accompany chapter 2 . . . . .	251
F.1.1	P-spline plotting function . . . . .	251
F.2	Code to accompany chapter 4 . . . . .	253
F.2.1	Calculation of $E((x - k)_+^p)$ for structural P-splines . . . . .	253
F.3	Code to accompany chapter 5 . . . . .	254
F.3.1	Multiple imputation with true exposure observed in a subset . . . . .	254
F.3.2	Moment reconstruction with true exposure observed in a subset . . . . .	255
F.3.3	group-SIMEX . . . . .	257

F.3.4	Plots of SIMEX extrapolation curves . . . . .	259
F.4	Code to accompany chapter 8 . . . . .	261
F.4.1	DerSimonian and Laird based multivariate meta-analysis . . . . .	261
F.4.2	REML based multivariate meta-analysis . . . . .	264

<b>References</b>		<b>267</b>
-------------------	--	------------

# List of Figures

2.1	B-spline and power basis functions. . . . .	24
2.2	Different functional forms for the linear predictor in a Cox proportional hazards model. . . . .	30
3.1	Grouped exposure analysis showing the effect of random measurement error on the observed exposure–disease relationship. . . . .	56
3.2	Fractional polynomial analysis showing the effect of random measurement error on the observed exposure–disease relationship. . . . .	57
3.3	P-spline analyses showing the effect of random measurement error on the observed threshold and non-linear threshold shaped exposure–disease relationships. . . . .	58
4.1	[Illustration of the SIMEX method.]Illustration of the SIMEX method for a true U-shaped relationship given by $y = \frac{x^2}{2}$ . One thousand standard normal values were generated for the true exposure, normally distributed measurement error with variance 0.5 was added to form the observed exposure. Quadratic relationship between outcome and exposure was assumed in the SIMEX model. $B = 200$ and rational linear extrapolant were used. . . . .	73
5.1	Grouped exposure analysis of the exposure–disease relationship using MacMahon’s method for correcting for the effects of exposure measurement error. . . . .	90
5.2	Structural P-spline analysis showing the measurement error corrected exposure–disease relationship. . . . .	112
5.3	Structural fractional polynomial analysis showing the measurement error corrected exposure–disease relationship. . . . .	113
5.4	Gradient of structural fractional polynomial and P-spline models for measurement error corrected threshold exposure–disease relationship. . . . .	114
5.5	rMSE for structural fractional polynomial and P-spline simulations. . . . .	115
6.1	Histograms of FBG and log-FBG for all studies combined . . . . .	123
6.2	Grouped exposure analysis of the observed FBG–CHD relationship for increasing levels of adjustment for confounders. . . . .	125

6.3	Fractional polynomial and P-spline models for the observed FBG–CHD relationship (with 95% confidence intervals given by dotted lines in the colour corresponding to each method). . . . .	126
6.4	Estimated RDRs for FBG by study. . . . .	127
6.5	Graphs as proposed by Ruppert & Carroll for testing measurement error assumptions for studies with repeat measurements of FBG—PRHHP study. . . .	129
6.6	Measurement error corrected grouped exposure analysis of the FBG–CHD relationship using MacMahon’s method. . . . .	131
6.7	Comparison of the uncorrected and MacMahon method corrected group exposure analyses. . . . .	131
6.8	SIMEX corrected fractional polynomial and P-spline models for the FBG–CHD relationship. . . . .	132
6.9	SIMEX extrapolation plots for the SIMEX corrected fractional polynomial model in figure 6.8. . . . .	134
6.10	Illustration of data transformation given in equation 6.1 for structural fractional polynomials. . . . .	136
6.11	Structural fractional polynomial and P-spline models for the FBG–CHD relationship. . . . .	137
6.12	Summary of the fractional polynomial and P-spline models for the FBG–CHD relationship considered in chapter 6. . . . .	140
7.1	Example measurement error plots for a simulated sample of 1,000 individuals under scenarios 1-5, $\sigma_u^2 = 1$ . Loess smooths of the data (red lines) are shown to aid trend identification in the left hand column. . . . .	149
7.2	Example measurement error plots for a simulated sample of 1,000 individuals under scenarios 6-10, $\sigma_u^2 = 1$ . Loess smooths of the data (red lines) are shown to aid trend identification in the left hand column. . . . .	150
7.3	P-spline analysis showing the observed linear exposure–disease relationship under each of the ten scenarios. . . . .	152
7.4	P-spline analysis showing the observed asymptotic exposure–disease relationship under each of the ten scenarios. . . . .	153
7.5	P-spline and fractional polynomial analyses of the observed threshold exposure–disease relationship under scenario 10. . . . .	154
7.6	Structural fractional polynomial analysis of the corrected linear exposure–disease relationship under each of the ten scenarios. . . . .	155
7.7	Structural fractional polynomial analysis of the corrected asymptotic exposure–disease relationship under each of the ten scenarios. . . . .	156
7.8	Structural P-spline and fractional polynomial analyses of the corrected threshold exposure–disease relationship under scenario 10. . . . .	157

7.9	Structural fractional polynomial and P-spline models for the FBG–CHD relationship where the distribution of true FBG given the observed is modelled using a two-component normal mixture model. . . . .	162
8.1	Forest plot of the slope parameter for a linear FBG–CHD relationship. . . . .	172
8.2	Funnel plot of the slope parameter for a linear FBG–CHD relationship. . . . .	173
8.3	Galbraith plot for the slope parameter for a linear FBG–CHD relationship. . . . .	173
8.4	Example 1 — Pooled exposure–disease relationship under each of the three pooling rules for simulated data. . . . .	185
8.5	Example 2 — Pooled exposure–disease relationship under each of the three pooling rules for simulated data. Note that in this example the pointwise and pointwise differential methods give identical results. . . . .	186
8.6	Example 3 — Pooled exposure–disease relationship under each of the three pooling rules, with two choices of reference value. . . . .	188
8.7	Multivariate meta-analysis of a grouped exposure analysis of the observed FBG–CHD relationship. . . . .	189
8.8	P-spline models of the FBG–CHD relationship under each of the three model selection rules. . . . .	191
8.9	P-spline models for the FBG–CHD relationship in individual studies using the overall model selection rule, and the corresponding weights the relationships receive in a pointwise random effects meta-analysis. . . . .	192
8.10	Meta-analysis of the FBG–CHD relationship using fractional polynomial models chosen by the overall model selection rule and pooled using each of the three pooling rules. . . . .	193
8.11	Exploration of how heterogeneity in the FBG–CHD relationship between studies varies with FBG. . . . .	198
8.12	Meta-analysis of the relationship between FBG and CHD, uncorrected and corrected for the effects of measurement error . . . . .	199
9.1	Histograms of $L_p(a)$ and $\log-L_p(a)$ for all studies combined. . . . .	204
9.2	Mean-variance association of $\log-L_p(a)$ for each of the six studies with repeat $L_p(a)$ measurements. Red lines give LOESS smooth of points to aid trend identification. . . . .	206
9.3	Normal Q–Q plots of within-person means of $\log-L_p(a)$ for each of the six studies with repeat $L_p(a)$ measurements. . . . .	207
9.4	Normal Q–Q plot of differences in $\log-L_p(a)$ between baseline and repeat measurements for the ULSAM study. . . . .	209
9.5	Group based analysis of observed $L_p(a)$ adjusted for age and sex (and other conventional cardiovascular risk factors). The group containing those with the lowest $L_p(a)$ measurements is taken as the reference. . . . .	210

9.6	Fractional polynomial and P-spline analyses of observed Lp(a) adjusted for age, sex, and other conventional cardiovascular risk factors. . . . .	210
9.7	Fractional polynomial and P-spline analyses of the Lp(a)–CHD relationship for each study. . . . .	212
9.8	Structural P-spline analysis of the Lp(a)–CHD relationship adjusted for age, sex, and other conventional cardiovascular risk factors. . . . .	213
9.9	Structural P-spline analysis of the Lp(a)–CHD relationship for each study. Darkness of lines related to weights in a pointwise meta-analysis of the individual studies. . . . .	214
9.10	Higgin’s $I^2$ across the range of Lp(a) for pointwise and derivative pointwise pooled structural P-spline models of the Lp(a)–CHD relationship. . . . .	214
B.1	Boxplots of RFrMSE for structural fractional polynomial and P-spline models for the exposure–disease relationship. . . . .	243
D.1	Example measurement error plots under scenarios 2, 4, 5, 8-10, $\sigma_u^2 = 1$ using log-exposure. . . . .	247

# List of Tables

3.1	Models considered for the relationship between true exposure $X$ and log-hazard in the simulation study. . . . .	48
3.2	Parameter values for the true exposure–disease relationships used in the simulation study. . . . .	49
3.3	Estimated power to detect non-linearity under each non-linear exposure–disease relationship shape and estimated type I error in a test of the null hypothesis of linearity when the true association is linear, using grouped exposure, fractional polynomial, and P-spline analyses. . . . .	55
5.1	Estimates of the slope parameter in a logistic regression of $Y$ on $X_C$ across simulations using the true exposure, naive method, and different correction methods when $X$ is observed in a subset. (Scenario (1a)) . . . . .	101
5.2	Estimates of the slope parameter in a logistic regression of $Y$ on $X_C$ across simulations using the true exposure, naive method, and different correction methods when $W_2$ is observed in a subset. (Scenario (1b)) . . . . .	102
5.3	Estimates of the slope parameter in a logistic regression of $Y$ on $X_C$ across simulations using the true exposure, naive method, and different correction methods when $X$ is observed in a subset. (Scenario (2a)) . . . . .	103
5.4	Estimates of the slope parameter in a logistic regression of $Y$ on $X_C$ across simulations using the true exposure, naive method, and different correction methods when $W_2, W_3$ are observed in a subset. (Scenario (2b)) . . . . .	104
5.5	Estimated power to detect non-linearity under each shape for the exposure–disease relationship, using structural fractional polynomial and P-spline analyses. . . . .	110
6.1	Summary of the ERFC FBG data by study. . . . .	122
7.1	Description of each of the scenarios considered in the simulation study. . . . .	147
9.1	Summary of the ERFC Lp(a) data by study. . . . .	203
9.2	Regression calibration models for log-Lp(a) for each study with repeats, and overall model obtained via random effects meta-analysis. . . . .	208

B.1	Estimates of the slope parameter in a linear regression of $Y$ on $X_C$ across simulations using the true exposure, naive method, and different correction methods when $X$ is observed in a subset. (Scenario (1a)) . . . . .	235
B.2	Estimates of the slope parameter in a linear regression of $Y$ on $X_C$ across simulations using the true exposure, naive method, and different correction methods when $W_2$ is observed in a subset. (Scenario (1b)) . . . . .	236
B.3	Estimates of the slope parameter in a linear regression of $Y$ on $X_C$ across simulations using the true exposure, naive method, and different correction methods when $X$ is observed in a subset. (Scenario (2a)) . . . . .	237
B.4	Estimates of the slope parameter in a linear regression of $Y$ on $X_C$ across simulations using the true exposure, naive method, and different correction methods when $W_2, W_3$ are observed in a subset. (Scenario (2b)) . . . . .	238
B.5	Estimates of the slope parameter in a logistic regression of $Y$ on quintile groups of exposure across simulations using the true exposure, naive method, and different correction methods when $X$ is observed in a subset. (Scenario (1a)) . . .	239
B.6	Estimates of the slope parameter in a logistic regression of $Y$ on quintile groups of exposure across simulations using the true exposure, naive method, and different correction methods when $W_2$ is observed in a subset. (Scenario (1b)) . .	240
B.7	Estimates of the slope parameter in a logistic regression of $Y$ on quintile groups of exposure across simulations using the true exposure, naive method, and different correction methods when $X$ is observed in a subset. (Scenario (2a)) . . .	241
B.8	Estimates of the slope parameter in a logistic regression of $Y$ on quintile groups of exposure across simulations using the true exposure, naive method, and different correction methods when $W_2, W_3$ are observed in a subset. (Scenario (2b)) . . . . .	242

## Notation

We shall use the following notation throughout this dissertation:

$W$	Observed exposure
$X$	True exposure
$U$	Measurement error
$Z$	Covariates measured without error
$Y$	Outcome
$W_{ij}$	Observed exposure for the $i$ th individual on the $j$ th occasion
$h(\cdot)$	Hazard function
$\lambda$	Regression dilution ratio
$\zeta$	SIMEX parameter
$\alpha$	Parameters of measurement model
$\beta$	Parameters of regression model
$S(\cdot, \cdot)$	Spline basis function
$g^2(\cdot)$	Measurement error variance function
$\tau^2$	Between study variance
$R(\cdot)$	Risk set
$D(\cdot)$	Set of failures
$\tau_j$	failure time $j$
$d_j$	$ D(\tau_j) $ the number of failures at time $\tau_j$
$\ell(\cdot, \cdot)$	Log-likelihood
$U(\cdot)$	Score function
$df$	Degrees of freedom
$I(X > C)$	An indicator variable that takes the value 1 if $X > C$ and 0 otherwise.

This is intended as a reference and the above notation, along with all additional notation, shall be properly introduced within the main body of the document.

## Abbreviations

All abbreviations will be introduced in the text the first time that they are used. However, we include a summary of abbreviations for reference:

AIC	Akaike's information criterion
APCSC	Asia Pacific Cohort Studies Collaboration
BIC	Bayesian information criterion
BMI	Body mass index
CHD	Coronary heart disease
ERFC	Emerging Risk Factors Collaboration
FBG	Fasting blood glucose
FFQ	Food Frequency Questionnaire
FP1, FP2	Fractional polynomial of degree 1/2
GCV	Generalised cross validation
gSIMEX	Group-SIMEX
IFG	Impaired fasting glucose
IPD	Individual participant data
Lp(a)	lipoprotein(a)
MCMC	Markov chain Monte Carlo
MCSIMEX	Misclassification SIMEX for categorised exposures
MI	Multiple Imputation
MR	Moment reconstruction
MVRS	Multivariable regression splines
RDR	Regression dilution ratio
REML	Restricted maximum likelihood
RFrMSE	Reference free root mean squared error
rMSE	Root mean squared error
SAT	Scholastic aptitude test
SBP	Systolic blood pressure
SIMEX	Simulation-extrapolation for continuous exposures

Abbreviations for the studies used in the analysis of the ERFC FBG data in chapters 6, 7 and 8 are given in appendix C. Abbreviations for the studies used in the analysis of the ERFC Lp(a) data in chapter 9 are given in appendix E.

# Chapter 1

## Introduction

In this brief chapter we discuss the motivation for this thesis; the characterisation of non-linear exposure–disease relationships in the Emerging Risk Factors Collaboration (ERFC), a large individual participant data (IPD) meta-analysis of risk factors for coronary heart disease (CHD). We then give some background on the field of epidemiology and discuss the importance of accurately characterising the shape of exposure–disease relationships. After this we give short introductions to two of the major themes investigated in this dissertation: measurement error and meta-analysis. We finish this chapter with an overview of the material to be covered in the remainder of this dissertation.

### 1.1 Motivation

We are in the midst of a global cardiovascular epidemic [3]. Although cardiovascular disease is typically associated with the Western world, cardiovascular disease is the largest cause of death in almost all countries in the world. CHD caused approximately 1 out of every 6 deaths in the United States in 2006; an average of 1 death every 38 seconds [4]. In the United Kingdom cardiovascular disease is the main cause of mortality, causing almost 200,000 deaths every year. Amongst the most developed countries in the world, only Ireland and Finland have a higher rate of CHD than the UK [5]. In recent years death rates from cardiovascular disease have declined as the result of better management of risk factors, and new medical procedures, yet the burden of disease remains high [4, 5]. The burden of cardiovascular disease is not just on the individual but on society through direct increases in healthcare costs, and the indirect costs of lost productivity due to morbidity and premature mortality [6]. Although great strides have been made in understanding the aetiology of CHD over the past sixty years, there is still much that is not known about CHD. An increased understanding of the causes of CHD can hopefully lead to treatment of risk factors so that populations can enjoy longer healthy lives.

This dissertation is motivated by the ERFC [7], a collaboration of studies contributing IPD on lipid, inflammatory and other markers of CHD in over 1.1 million, predominantly white

## 1. Introduction

---

Western, participants in over 100 prospective studies of cardiovascular diseases. The first CHD event (fatal or non-fatal) is the primary event of interest of which there are over 69,000 incidents within the 11.7 million person years at risk. CHD events include both non-fatal myocardial infarctions and CHD related deaths. The ERFC is investigating the risk factor–CHD relationship for many risk factors including adiposity markers, C-reactive protein [8], fasting blood glucose (FBG) and diabetes [9, 10], lipoprotein-a (Lp(a)) [11], lipids and apolipoproteins [12], and renal function. We shall focus on two of these risk factors in this thesis, FBG (chapters 6, 7, 8) and Lp(a) (chapter 9).

The ERFC aims to better characterise risk factor–CHD relationships. It is able to do this because it has observations from such a large number of studies, and data at the individual participant level. This gives us better power to characterise the shape of the risk factor–CHD relationship than each study individually, and allows us to model the studies in a consistent way. Many risk factors considered by the ERFC are subject to substantial measurement error and are known, or are suspected to have, a non-linear relationship with CHD, including FBG and Lp(a). The ERFC is part of a growing trend within epidemiological research to use the extra power obtained from combining studies that have tackled similar research hypotheses. Other examples include the Asia Pacific Cohort Studies Collaboration (APCSC) [13] and the Collaborative Group on Hormonal Factors In Breast Cancer [14].

An aim of the ERFC is to ‘help advance biostatistical methodology for the analysis of observational data from multiple studies’ [7], and we hope that this dissertation makes a contribution towards achieving this. This thesis originated from the observation the ERFC made that ‘characterizing the shape of the underlying exposure–disease relationship, while taking into account possibly heterogeneous measurement error, is not well studied, especially in the context of IPD meta-analysis’ [15]. This dissertation aims to tackle the problems of characterising the exposure–disease relationship; provide suitable methods for correcting for exposure measurement error, considering the possibility of heteroscedastic or non-normal measurement error; and also provide methods for combining non-linear exposure–disease relationships across studies.

## 1.2 Epidemiology

Epidemiology is the study of patterns of health and disease at the population level. Epidemiology is used in public health research to identify risk factors for disease which can then be used to inform policy decisions and evidence-based medicine. Modern day epidemiology started with John Snow, the ‘father of epidemiology’ [16], who showed the relationship between water sources and outbreaks of cholera in London in the 1850’s [17]. Since the 19th century there has been a huge growth in epidemiology and perhaps one of the most notable achievements of epidemiology was in the 1950’s when the relationship between tobacco smoking and lung cancer was shown [18–20].

---

Epidemiologists use studies of groups of individuals to investigate relationships between exposure and disease. The two main types of study are case-control studies and cohort studies. In a case-control study individuals are sampled according to disease status. We therefore have two groups of individuals, the cases and the controls, which we can compare. In a cohort study we typically follow a well defined population who are all disease free at the beginning of the study, until either all individuals become diseased or until the study ends.

Since the early 20th century there have been numerous prospective cohort studies of various facets of population health. Many of these have focused on cardiovascular disease including the Framingham Heart Study of the epidemiology of cardiovascular disease in the US [21] which started in 1948 and is still going today. The Framingham Heart Study was the first to show the effects of smoking and high cholesterol on heart disease [22, 23]. Since then the relationship of risk factors such as age, high cholesterol, tobacco smoking and high blood pressure with CHD has been consistently found. However, high levels of these risk factors are not present in a large number of CHD cases [24]. In recent years there has been much interest in trying to identify other risk factors for CHD, and since it seems unlikely that a single risk factor with a large effect has been overlooked, the focus has been on identifying a number of risk factors, each of which has a modest effect on CHD risk. To show statistical significance of modest effect sizes requires large sample sizes and therefore, large meta-analyses which combine evidence from multiple studies such as the ERFC and APCSC are required.

Many epidemiological studies have considered only a linear relationship between exposure and disease or have chosen to dichotomise the exposure due to small study size. There are increasing numbers of large studies and meta-analyses thereof where we have sufficient power not only to discern whether there is a relationship between exposure and disease, but also to be able to investigate the shape of this relationship. In epidemiology it is important to be able to describe the shape of the exposure–disease relationship accurately because this can have a significant impact on the policy decisions that are made as a result of research, which will in turn, impact on individuals’ health outcomes. It may be that as a result of understanding the shape of the relationship we are able to identify sets of individuals who would benefit from targeted health interventions. For example in occupational epidemiology we may wish to identify an exposure level beyond which the risk of disease becomes unacceptably high, so that action may be taken to prevent individuals from being exposed above this level. We may identify a set of high-risk individuals which clinicians would like to target to reduce their risk, or we may find that lower exposure at the population level could reduce overall risk, this being achieved through public health initiatives. As an example, if we consider the relationship between blood pressure and CHD then there will be a set of high risk individuals with high blood pressure who will be treated using drugs such as ACE inhibitors, whereas the general population can be given information on lifestyle changes that will help reduce their blood pressure such as taking regular exercise; eating a healthy, balanced diet; drinking less alcohol; and smoking less.

## 1. Introduction

---

Misestimation of the exposure–disease relationship can lead to inappropriate interventions. We need to be concerned with misestimating all levels of risk and not only large risks since a large number of people exposed to a small risk may generate many more cases than a small number exposed to a high risk [25]. By underestimating small risks we may, for example, make it appear uneconomically viable to intervene, which could in turn result in adverse health outcomes for some individuals. Overestimation of risk can have consequences too; for example individuals may unnecessarily be given medication to reduce exposure levels, from which some individuals may suffer from side effects, and the cost of the medication may have been better spent on modifying other risk factors.

Statistics plays a key role in epidemiology. Firstly, statistics may be used in the design of epidemiological studies to inform the required sample size to be able to detect a given effect size. Once the data have been collected then statistics can be used to analyse the data and quantify statistically the significance of results, allowing for differences in characteristics between individuals. Well conducted statistical analysis can help us differentiate between a true exposure–disease relationship, and chance findings. Case-control studies can be modelled using logistic regression where the outcome is a binary indicator denoting whether each individual was diseased at the end of the study or not. In cohort studies Cox proportional hazards models [26] are usually appropriate, because the outcome is a composite of the disease status and censoring time, although logistic regression is also common, especially when there is little censoring. Statistical methods can be used to allow for certain deficiencies in the data collection such as missing data and measurement error. Despite the long relationship between statistics and epidemiology, epidemiology constantly provides new statistical challenges, requiring continual methodological development by statisticians.

### 1.3 Measurement error

#### 1.3.1 The effects of measurement error

Measurement error is ubiquitous in the measurement of exposures whether it be the measurement of biological exposures such as serum creatinine [27], or blood glucose [28]; environmental exposures such as radon [29] or air pollution [30]; dietary exposures such as energy intake [31] or alcohol consumption [32]; or a plethora of other quantities of interest. In fact almost all measurements of physical quantities will be subject to some degree of measurement error.

If for example a study were interested in investigating the relationship between blood pressure and a disease, then they may want to estimate an individual’s long term average, or usual, blood pressure since an individual’s blood pressure is known to vary from day to day. However, if a single blood pressure measurement were to be taken using a sphygmomanometer then the reading obtained will differ from the subject’s usual blood pressure in two ways. Firstly, there will be measurement error due to the measuring device and secondly, day to day variation.

---

These two sources of error: measurement error and within-person variation, are similar in that both prevent us from measuring what we would like to be able to, usual blood pressure, and in fact they have a similar effect on the observed exposure–disease relationship. Therefore, in this dissertation we shall use the term measurement error to mean both error arising from the measurement process and within-person variation.

The effects of measurement error on exposure–disease relationships were neatly summarised by Carroll *et al.* [33] as:

- bias in parameter estimation in models,
- loss of power in detecting associations between variables,
- masking of features of the data, hence making it difficult to ascertain the true shape of the exposure–disease relationship.

In this dissertation we shall see each component of this ‘*triple whammy*’ of effects numerous times. The effects of exposure measurement error have been known for a long time; for example in 1904 Spearman [34] noticed how random measurement error reduces the correlation between two variables; he was the first to refer to this as ‘*attenuation*’ otherwise known as regression dilution. Over time people have become increasingly aware that when exposure measurement error is present, the observed exposure–disease relationship is not the same as the true relationship; usually, but not always, biasing it towards the null.

Many risk factors for CHD are subject to measurement error such as measurements of blood pressure [35–37] and cholesterol [36–39], as well as the two risk factors we consider in this dissertation: FBG [28, 40] and Lp(a) [11]. Emberson *et al.* [41] state that ‘The importance of risk factors for CHD can be greatly underestimated by using a single baseline measure in prospective study analyses’, and thereby not taking into account the effects of measurement error. They go on to say that ‘Studies that wish to estimate associations between disease risk and usual exposure levels need to take regression dilution effects into account. Failure to do so can lead to serious misinterpretation of the importance of CHD risk factors’.

### **1.3.2 Correction for measurement error**

Because of the ‘*triple whammy*’ of measurement error it is essential to correct for its effects in analyses if we are interested in obtaining the true relationship between exposure and disease. In order to correct for the effects of measurement error, additional data are required. These additional data will typically be obtained from a subset of the study’s participants. They may consist of exposure measurements that contain substantially less measurement error than the original measurement, this might be the case if this superior measurement is costly; alternatively, as in the constituent studies of the ERFC, some individuals may be recalled so that another exposure measurement using the original method can be taken.

## 1. Introduction

---

In the case of a linear model with linear predictor there exists a simple and effective approach to correcting for the effects of exposure measurement error, regression calibration, which we describe in chapter 4. This method is also approximately correct for Cox and logistic regression [28, 42]. Because regression calibration is simple and easy to use it has been widely used [43, 44]. However, when the exposure–disease relationship is non-linear there is no simple method of correcting for measurement error. Numerous methods have been proposed for correcting for the effects of exposure measurement error when the exposure–disease relationship is non-linear, and we discuss some of these in chapter 4, however it would be fair to say that no single method is commonly used in practice.

Correction for measurement error can cause large differences in our conclusions. For example, Emberson *et al.* [39] showed that without adjustment for the effects of exposure measurement error 50% of CHD events in middle aged men can be attributed to serum total cholesterol, blood pressure and cigarette smoking, however this became 80% on adjustment for the effects of measurement error. The difference between corrected and uncorrected analyses can have major implications for public health policy and the direction of future research. Appropriate correction is essential, otherwise inflated or misleading results may be obtained [45]. What is an appropriate correction for measurement error is sometimes a contentious issue; for example there was much debate about the methods used for correcting the relationship between 24-hour sodium excretion in urine and blood pressure, for measurement error in the Intersalt study [45–49].

Many books have been written on the topic of measurement error including Fuller’s *Measurement Error Models* [50], Carroll *et al.*’s *Measurement Error in Nonlinear Models* [33], Gustafson’s *Measurement Error and Misclassification in Statistics and Epidemiology* [51] and Buonaccorsi’s *Measurement error: models, methods, and applications* [52]. Despite the fact that the effects of exposure measurement error and the necessity of correcting for it have been known for a long time, research into this area is still very active [1, 53–55].

### 1.4 Meta-analysis

Meta-analysis is the ‘analysis of analyses’ [56]; it allows us to combine the results of multiple studies addressing a set of related research hypotheses. Many attribute the first meta-analysis to Pearson in 1904 [57], where he investigated the effectiveness of enteric fever inoculations. Meta-analysis has become an increasingly popular research tool; the Cochrane Collaboration alone has published over 4000 meta-analyses. Apart from the use of meta-analysis for medical outcomes, it has also been used in the fields of education [58], psychology [59], and economics [60].

Meta-analysis allows us to make conclusions from the data that we may not have been able to from any study individually. This is due to the power gained from a larger sample size,

---

especially when the effect size of interest is small. Meta analysis is a fast growing research area with current areas of activity including multivariate and network meta-analysis.

Meta-analyses traditionally have focused on combining the published results (i.e. data at the aggregate level) from previous studies. However, increasingly meta-analyses are being conducted on the IPD from each study. IPD meta-analysis is the gold standard approach to combining data across multiple studies because it allows us to analyse the data consistently across studies; we discuss in detail the advantages and disadvantages of IPD meta-analysis in chapter 8. Stewart and Parmar [61] state that ‘Whenever possible, a meta-analysis of updated individual patient data should be done because this provides the least biased and most reliable means of addressing questions that have not been satisfactorily resolved’ by individual studies. There has been a rapid growth in the number of published papers that use IPD meta analysis; Riley *et al.* [62] conducted a systematic review of the number of IPD meta-analyses that had been conducted between 1991 and 2008 which showed half a dozen were being performed per year at the beginning of the period, increasing to 49 per year between 2005 and 2009.

Meta analysis can be contentious, for example Eysenck [63] described a meta-analysis of the outcome of psychotherapy [64] as ‘an exercise in mega-silliness’ and Egger, Schneider and Davey Smith argue that meta-analysis should ‘not be a prominent component of reviews of observational studies’ [65]. As we discuss in chapter 8, a good meta-analysis should be well conducted and any deficiencies highlighted [66].

In this dissertation we are concerned with IPD meta-analysis not of point estimates of effect size as is usually the subject of meta-analyses, but of the shape of non-linear exposure–disease relationships across the exposure range; a topic that has previously received little attention [67].

## 1.5 Overview of dissertation

Following an introduction to the analysis of time to event data and measurement error, we look at the effects of exposure measurement error in a single study, on a range of epidemiologically plausible shapes for the exposure–disease relationship using simulation. Having seen the effects, we then go on to consider methods that have been proposed for correcting for exposure measurement error as well as developing our own, and we assess the performance of these methods, again by simulation. We then apply the methods to an example dataset looking at the relationship between FBG and CHD in the ERFC, although we ignore possible heterogeneity in the shape of association between studies at this point. We then consider the effects of non-standard error structures, and the implications of erroneously assuming the standard classical measurement error model on our correction methods. Following this, we return to consider how heterogeneity in the shape of the risk factor–disease relationship between studies, which we previously ignored, can be accounted for using the tools of meta-analysis. We then apply the methods that have been developed throughout the dissertation to another risk factor consid-

ered by the ERFC, Lp(a). Finally, we summarise our conclusions, make suggestions for future work and offer some recommendations for practice.

### 1.5.1 Chapter structure

Chapter 2 introduces the Cox proportional hazards model and looks at different forms for the linear predictor when modelling non-linear exposure–disease relationships focusing on grouped exposure analyses, fractional polynomials and splines.

Chapter 3 introduces the concept of exposure measurement error. We investigate the effect that random exposure measurement error has on a range of epidemiologically plausible non-linear exposure–disease relationships using a simulation study.

Chapter 4 looks at methods for correcting for the effects of exposure measurement error in non-linear models. We use a common correction method, regression calibration, to develop a method for correcting fractional polynomial analyses.

Chapter 5 consists of three simulation studies. The first extends the simulation study of chapter 3 to look at the performance of MacMahon’s method for correcting for measurement error in grouped exposure analyses. The second compares the performance of existing and proposed methods for correcting for measurement error in categorised continuous exposures. The third considers the performance of fractional polynomial and P-spline methods, described in chapter 4, at correcting for exposure measurement error, again extending the simulation of chapter 3.

Chapter 6 focuses on applying fractional polynomial and P-spline methods to the FBG–CHD relationship using ERFC data.

Chapter 7 extends the simulation study of chapter 3 to consider scenarios where the true exposure is subject to non-normal and/or heteroscedastic measurement error, or the true exposure is non-normal, looking at the effect these scenarios have on the observed exposure–disease relationship, and the corrected relationship when we erroneously assume the standard classical measurement error model. Secondly, we consider the use of mixture models to robustify our fractional polynomial and P-spline models.

Chapter 8 starts with an introduction to the basic concepts of meta-analysis. We then consider approaches to carrying out a 2-stage IPD meta analysis of non-linear exposure–disease relationships. We conclude by reanalysing the ERFC FBG data from chapter 6 using the methods developed.

Chapter 9 is an application of the methods that we have developed throughout this dissertation to another risk factor for CHD from the ERFC, Lp(a).

Chapter 10 gives conclusions, areas for future work, and recommendations for practice.

# Chapter 2

## Modelling non-linear exposure–disease relationships

In this chapter we look at survival analysis and methods for modelling a continuous exposure–disease relationship when we have time to event data. We consider the Cox proportional hazards model focusing on: grouped exposure, fractional polynomial and spline functions for the linear predictor. This chapter shall set the foundations for us to consider the effects of exposure measurement error on the observed exposure–disease relationship, and how we may correct for these effects, in the rest of this dissertation.

### 2.1 The Cox model

#### 2.1.1 Introduction to survival analysis

Survival analysis is concerned with modelling the time until an event (or failure) occurs. Events of interest may be health related outcomes such as death, heart attack or stroke, or cancer; or in industrial applications failure of a part within a machine. Sometimes we may only be interested in modelling the rate at which failures occur, but often we want to use a set of explanatory variables to explain differences in survival between individuals.

The survival function  $S(t)$  gives the probability that the observed failure time  $T$  is greater than some specified time  $t$ ,

$$S(t) = \Pr(T > t).$$

The survival function can be estimated parametrically, using for example a Weibull or exponential distribution, or using the non-parametric Kaplan–Meier product–limit estimator [68]. If we let  $n_i$  be the number of individuals under observation at time  $t_i$ , and  $d_i$  the number of deaths

## 2. Modelling non-linear exposure–disease relationships

---

at time  $t_i$ , then the Kaplan–Meier estimator is

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}.$$

This is a step function where the steps occur at the unique failure times. As the sample size increases the survival function will approach the true survival curve. The variance of  $\hat{S}(t)$  can be calculated using Greenwood’s formula [69].

The hazard function,  $h(t)$ , gives the instantaneous risk of experiencing an event at time  $t$ , conditional on survival to time  $t$ ,

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P[(t \leq T \leq t + \Delta t) | T \geq t]}{\Delta t} = -\frac{S'(t)}{S(t)}$$

and can be interpreted as the number of events per population at-risk per unit time. This leads to the cumulative hazard which is

$$H(t) = \int_0^t h(u) du.$$

This is the number of events per population at-risk over the time interval  $[0, t)$ .

Throughout this dissertation we use the proportional hazards model because it allows us to make no assumptions about the form of the baseline hazard. Under the proportional hazards model we assume that the hazard for any individual is a constant (over time) proportion of the hazard for any other individual i.e. for individuals  $i$  and  $j$  and some constant  $k$ ,

$$\frac{h_i(t)}{h_j(t)} \equiv k \quad \forall t.$$

Survival data usually exhibit one or more forms of censoring. Censoring is when we do not know either the time when an individual joined the risk set or their event time. The different forms of censoring are:

- Left censoring—When we only know that an individual’s lifetime is less than some value i.e.  $T < t$ .
- Right censoring—When we only know that the event took place after some time,  $T > t$ .
- Interval censoring—When we only know that the event took place in some interval  $t_1 < T < t_2$ .

The likelihood function, and hence parameter estimation, is complicated by having to take account of the censoring exhibited by the data.

Right censoring and interval censoring are most common in epidemiological studies. Right censoring will occur due to individuals being lost to follow up, or the study ending, whilst in-

---

terval censoring can occur due to the event of interest occurring between follow-up visits. Censoring is usually assumed to be non-informative in epidemiological cohort studies i.e. knowing that the observation was censored tells us nothing about the expected survival time for that individual.

## 2.1.2 Cox proportional hazards model

The Cox model [26] for modelling time to event data has proved highly influential within the field of statistics and has been widely used in a diverse range of applications. The model was a particular boon for medical statistics as it allowed a way of combining knowledge of event data and censoring. This has led to the original 1972 paper [26] receiving almost nearly 20,000 citations (January 2011) on Google Scholar; the second most cited statistics paper of all time [70]. The paper also introduced the idea of partial likelihood whereby we isolate the component of the full likelihood that contains the parameters of interest.

We shall be using the Cox proportional hazards model throughout this dissertation to model exposure–disease relationships. Under this model we assume that the hazard for individual  $i$  with covariate vector  $\mathbf{X}_i(t)$  at time  $t$ ,  $h_i(t)$ , is

$$h_i(t|\mathbf{X}_i) = h_0(t)e^{\beta^T \mathbf{X}_i}.$$

The Cox model is nonparametric as we do not specify the baseline hazard and only estimate the parameter vector  $\beta$ . This makes the Cox model robust against misspecification of the baseline hazard although Cox and Oakes [71] note that by specifying the baseline hazard you tend to get very similar results.

As we do not specify the baseline hazard we can only make inference about the difference in the ratio of the hazard between individuals. The hazard ratio for individual  $i$  relative to individual  $j$  is given by

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)e^{\beta^T \mathbf{X}_i}}{h_0(t)e^{\beta^T \mathbf{X}_j}} = \frac{e^{\beta^T \mathbf{X}_i}}{e^{\beta^T \mathbf{X}_j}} = \exp \{ \beta^T (\mathbf{X}_i - \mathbf{X}_j) \}.$$

For example, if the estimated coefficient for age in a Cox model was 0.0815 then this would imply an 8.1% increase in the event rate for each 1-year older in age compared with the reference. So, if our reference was age 50 then a 45 year-old who was identical in their other covariates would have about 33.5% lower risk of experiencing an event than a 50 year-old, whilst a 55 year-old would have about a 50% higher risk.

The Cox model relies on two assumptions about the underlying process:

- Proportional hazards assumption: the hazard for any individual is a constant (over time) proportion of the hazard for any other individual.
- The effect of risk factors is multiplicative.

## 2. Modelling non-linear exposure–disease relationships

---

A number of tests for checking the proportional hazards assumption have been suggested—we shall consider Schoenfeld residuals later in this section.

We now introduce some notation for survival models. For each individual  $i$  we observe the triplet  $(t_i, \delta_i, \mathbf{X}_i)$  where:

- $t_i$  is the length of the period of observation.
- $\delta_i$  is an indicator taking the value 1 if individual  $i$  experiences an event at time  $t_i$  and 0 otherwise.
- $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$  is a vector of explanatory covariates.

If we let:

- $\boldsymbol{\tau} = (\tau_1, \dots, \tau_k)$  be the set of unique event times where  $\tau_i < \tau_j$  if  $i < j$ ,
- $D(\tau_j)$  be the set of all individuals experiencing an event at time  $\tau_j$ ,
- $d_j = |D(\tau_j)|$ , the number of events at time  $\tau_j$ ,
- $R(\tau_j)$  be the set of individuals at risk immediately before event time  $\tau_j$ ,

then the partial likelihood for the Cox model is given by

$$\ell(\boldsymbol{\beta}) = \prod_{j=1}^k \frac{e^{\sum_{i \in D(\tau_j)} \boldsymbol{\beta}^T \mathbf{X}_i}}{(\sum_{i \in R(\tau_j)} e^{\boldsymbol{\beta}^T \mathbf{X}_i})^{d_j}}.$$

The log partial likelihood is given by

$$\log \ell(\boldsymbol{\beta}) = \sum_{j=1}^k \left( \sum_{i \in D(\tau_j)} \boldsymbol{\beta}^T \mathbf{X}_i - d_j \log \left( \sum_{i \in R(\tau_j)} e^{\boldsymbol{\beta}^T \mathbf{X}_i} \right) \right)$$

and the differentiated log-partial-likelihood,  $U(\boldsymbol{\beta})$  (the score function) has elements given by

$$U(\boldsymbol{\beta})_l = \frac{\partial \log \ell(\boldsymbol{\beta})}{\partial \beta_l} \tag{2.1}$$

$$= \sum_{j=1}^k \left( \sum_{i \in D(\tau_j)} \mathbf{X}_{il} - d_j \bar{X}_{il} \right). \tag{2.2}$$

where  $\bar{X}_{il} = \sum_{i \in R(\tau_j)} w_{ij} X_{il}$ , and  $w_{ij} = \frac{\exp(\boldsymbol{\beta}^T \mathbf{X}_i)}{\sum_{s \in R(\tau_j)} \exp(\boldsymbol{\beta}^T \mathbf{X}_s)}$

By setting  $U(\boldsymbol{\beta})$  equal to zero and solving for  $\boldsymbol{\beta}$  we obtain maximum (partial) likelihood estimates of the regression parameters  $\boldsymbol{\beta}$ .

---

The information matrix is the  $p \times p$  matrix with elements given by

$$\mathbf{I}_{lm}(\beta) = -\frac{\partial^2 \log \ell(\beta)}{\partial \beta_l \partial \beta_m} \quad (2.3)$$

$$= \sum_{j=1}^k \sum_{i \in R(\tau_j)} d_j w_{ij} (X_{il} - \bar{X}_{jl})(X_{im} - \bar{X}_{jm}). \quad (2.4)$$

$\hat{\beta}$  is asymptotically distributed  $N(\beta, E(I^{-1}(\beta)))$ . However, to calculate the expectation requires us to know the censoring distribution for all lives, including those that failed. Therefore, the inverse of the observed information matrix,  $I^{-1}(\hat{\beta})$ , is usually used instead.

The partial likelihood equations can be solved using the Newton-Raphson algorithm, with the starting value  $\hat{\beta}^{(0)} = \mathbf{0}$ . The update equation is

$$\hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} + I^{-1}(\hat{\beta}^{(n)})U(\hat{\beta}^{(n)})$$

with the algorithm terminating when the difference in the log partial-likelihood is less than a given tolerance level  $\epsilon$  i.e. when

$$|\ell(\hat{\beta}^{(n+1)}) - \ell(\hat{\beta}^{(n)})| < \epsilon.$$

The `coxph` function in the `survival` [72] package of R uses  $\epsilon = 10^{-9}$ . Standard Wald or likelihood ratio tests can be performed to ascertain the significance of model parameters.

More than one event may occur at the same time e.g. two deaths in the same day. Our formulae above have used the adjustment for ties given by Peto [73] and Breslow [74]. However, the adjustment is conservative because the denominator of the partial likelihood will count those individuals who have experienced an event more than once. Efron proposed an alternative whereby the terms corresponding to individuals who experienced events are weighted so that the terms contain an average of the people who experienced an event at time  $\tau_j$ . For example suppose we have five individuals ( $i = 1 \dots 5$ ) under observation immediately prior to time  $\tau_j$  of which three individuals ( $i = 1, 2, 3$ ) die at time  $\tau_j$ . Let individual  $i$  have hazard ratio  $h_i$  then the partial-likelihood contribution of the events at time  $\tau_j$  will be

$$\left( \frac{\frac{1}{3}(h_1 + h_2 + h_3)}{h_1 + h_2 + h_3 + h_4 + h_5} \right) \left( \frac{\frac{1}{3}(h_1 + h_2 + h_3)}{\frac{2}{3}h_1 + \frac{2}{3}h_2 + \frac{2}{3}h_3 + h_4 + h_5} \right) \left( \frac{\frac{1}{3}(h_1 + h_2 + h_3)}{\frac{1}{3}h_1 + \frac{1}{3}h_2 + \frac{1}{3}h_3 + h_4 + h_5} \right).$$

This approach is exact when the covariate values of the tied individuals are the same. The alternative to this is a full likelihood approach which takes into account the possibility that the  $d_j$  events could have taken place in any of  $d_j!$  orders. This is computationally complex to calculate. R uses Efron's method as default.

## 2. Modelling non-linear exposure–disease relationships

---

We can allow for possible differences in the baseline hazard between groups of individuals, for example trial centers in a multi-center trial, by stratification. Under a stratified model we allow each strata to have a different baseline hazard but we assume that the effect of the explanatory variables is the same across all strata. For an individual  $i$  in strata  $k$  the hazard is given by

$$h_i(t) = h_{0k}(t)e^{\beta^T \mathbf{X}_i}.$$

The log-partial likelihood for the stratified model will be the sum of the partial likelihoods of each strata. Hence, the score and information matrix are similarly the sum of the score, and information matrices, of each strata respectively.

Cox models are not easy to fit using a Bayesian approach, however it can be achieved. Clayton [75] uses the counting process formulation of the Cox model to fit frailty models using Markov chain Monte Carlo. These models can be fit for example in WinBUGS [76]. Bayesian analyses are computationally expensive because of the need for Gibb's sampling. A large dataset is used in this thesis, for which a Bayesian analysis will be impractical, and therefore this approach is not considered further.

### 2.1.3 Checking the proportional hazards assumption

Proportionality of hazards is one of the main assumptions of the Cox model. An alternative model is that the coefficient  $\beta(t)$  is time varying, so that we have the model

$$h(t|\mathbf{X}_i) = h_0(t)e^{\beta(t)^T \mathbf{X}_i}.$$

An example of when this model might be appropriate is in a clinical trial when the effectiveness of a drug decreases the longer the patient has been taking it. We can test whether the coefficients  $\beta$  are constant over time using the Schoenfeld residuals [77].

Let  $\mathbf{X}_{(\tau_j)}$  be the covariate vector for the individual failing at time  $\tau_j$ . The expected value of  $\mathbf{X}_{(\tau_j)}$  is the weighted mean of the covariates over those at risk immediately prior to  $\tau_j$  i.e.

$$\bar{\mathbf{x}}(\boldsymbol{\beta}, \tau_j) = \frac{\sum_{i \in R(\tau_j)} \mathbf{X}_i e^{\beta^T \mathbf{X}_i}}{\sum_{i \in R(\tau_j)} e^{\beta^T \mathbf{X}_i}}$$

and the Schoenfeld residual,  $s_{\tau_j}$ , for the  $j$ th failure is

$$s_{\tau_j} = \mathbf{X}_{(\tau_j)} - \bar{\mathbf{x}}(\hat{\boldsymbol{\beta}}, \tau_j).$$

If we have multiple events at time  $\tau_j$  then we get  $d_j$  residuals. From the definition of  $\hat{\boldsymbol{\beta}}$  we have that

$$\sum_{i=1}^n (\mathbf{X}_{(\tau_j)} - \bar{\mathbf{x}}(\hat{\boldsymbol{\beta}}, \tau_j)) = 0.$$

---

Grambsch and Therneau [78] showed that we can use scaled Schoenfeld residuals

$$s_{\tau_j}^* = \mathbf{V}^{-1}(\hat{\boldsymbol{\beta}}, \tau_j) s_{\tau_j}$$

where  $\mathbf{V}(\hat{\boldsymbol{\beta}}, \tau_j)$  is the conditional weighted variance of the covariate vectors of individuals alive immediately prior to event time  $\tau_j$ ,

$$\mathbf{V}(\hat{\boldsymbol{\beta}}, \tau_j) = \frac{\sum_{i \in R(\tau_j)} e^{\boldsymbol{\beta}^T \mathbf{X}_i} (\mathbf{X}_i(\tau_j) - \bar{\mathbf{x}}(\boldsymbol{\beta}, \tau_j))^T (\mathbf{X}_i(\tau_j) - \bar{\mathbf{x}}(\boldsymbol{\beta}, \tau_j))}{\sum_{i \in R(\tau_j)} e^{\boldsymbol{\beta}^T \mathbf{X}_i}}$$

to approximate the value of the parameter vector at time  $\tau_j$ . Let  $s_{\tau_j l}^*$  be the element of  $s_{\tau_j}^*$  that corresponds to the  $l$ th parameter, and  $\beta_l(\tau_j)$  the  $l$ th parameter in the model with time varying coefficients, then we have

$$E(s_{\tau_j l}^*) + \hat{\beta}_l \approx \beta_l(\tau_j).$$

By plotting  $s_{\tau_j l}^*$  against time we can see whether the proportional hazards assumption holds. If the assumption holds we would expect to see a horizontal line indicating no change in the parameter value over time. We could also check this by looking for a zero slope in the regression of the residuals against time. However, the plot allows us to visualise the association and we can investigate any outlying observations. Within a stratified model the residuals should be calculated using the variance matrix from within each strata. Methods to get rid of time dependence include adding an interaction between the covariate and time, and stratification.

For further discussion about checking for non-proportionality in Cox models and methods for correcting for it, see Therneau and Grambsch [79].

### 2.1.4 Checking the model fit

In order to check the model fit we can use the Martingale residuals [80],  $\hat{M}_i(t)$ , where

$$\hat{M}_i(t) = \delta_i - \hat{H}(t, \mathbf{X}_i, \hat{\boldsymbol{\beta}}),$$

$$\delta_i = \begin{cases} 1 & \text{if individual } i \text{ experiences an event} \\ 0 & \text{if individual } i \text{ is right censored} \end{cases}$$

and  $\hat{H}(t_i, \mathbf{X}_i, \hat{\boldsymbol{\beta}}_i)$  is the estimated cumulative hazard. The residuals are the difference over  $[0, t]$  between the observed number of events minus the expected number given the model. The martingale residual for the  $i$ th individual is defined as  $\hat{M}_i = \hat{M}_i(\tau_i)$ . We have one residual per observation.

The martingale residuals have the following properties

- $\sum_i \hat{M}_i = 0$

## 2. Modelling non-linear exposure–disease relationships

---

- Asymptotically  $E(\hat{M}_i) = \text{Cov}(\hat{M}_i, \hat{M}_j) = 0$ .

Hence we would expect that the residuals be equally distributed about zero if the model was a good fit to the data, although not normally distributed as under linear regression. We can plot the Martingale residuals for each covariate against the covariate values to visually assess this. Any systematic bias in the residuals is a sign that the model is not a good fit to the data and that the functional form of the relationship should be revisited. It can be helpful to add a smoother to the plot to aid assessment. Influential observations can be found by plotting  $i$  against  $\beta_{j,(-i)}$ , where  $\beta_{j,(-i)}$  is the  $j$ th parameter value obtained when observation  $i$  is excluded from the model.

### 2.1.5 Alternatives to the Cox model

Many approaches have been proposed for modelling survival data other than the Cox proportional hazards model. A logistic model, where we use the event indicator  $\delta_i$  as the outcome, is one of the simplest ways of modelling survival data and produces odds ratios which are the ratio between the number of events that occur in a given time period between the individual of interest and some reference. However, the problem with the logistic model is that it does not allow for censoring which may be substantial, especially during long follow up in epidemiological cohort studies.

Under the Cox model we do not specify the baseline hazard. An alternative approach is to specify the baseline hazard in a fully parametric model using an exponential or Weibull model. Under the exponential model we have that  $h_0(t) = \rho$  i.e. the baseline hazard is constant over time, whilst under the Weibull model

$$h_0(t) = \rho t^{s-1}$$

where  $\rho, s > 0$ . Note that this reduces to the exponential model when  $s = 1$ .

Another approach is to model the survival times using an accelerated failure time model. Under this model the effect of covariates is to multiply the time to event,  $T_i$ , by a constant such that

$$\log T_i = \beta^T \mathbf{X}_i + \epsilon_i$$

where  $\epsilon_i$  is an error term. Assuming a normal distribution for the error term will give a log-normal model, another possible choice is the extreme value distribution.

The Cox model assumes proportional hazards; an alternative to this is an additive hazards model. Under an additive hazards model we assume that [81],

$$h(t|\mathbf{X}) = h_0(t) + \beta^T \mathbf{X}.$$

---

The hazard of experiencing an event can be interpreted as the hazard that you experience the event due to the exposure of interest plus the hazard of experiencing the event from all other sources.

Despite these alternative approaches the Cox proportional hazards model remains the standard model for survival data.

## 2.2 Functional forms for the linear predictor

In many situations a simple linear form for the predictor may be appropriate. However, many exposure–disease relationships are non-linear. In this section we look at different methods for exploring and modelling a non-linear exposure–disease relationship using the Cox model. All the models for the exposure–disease relationship considered in this section may also include additional covariates; however we exclude them here for clearer presentation.

### 2.2.1 Grouped exposure analysis

Turner, Dobson and Pocock [82] recently carried out a survey of five major epidemiological journals and found that in the period considered 86% of articles employed categorisation of a continuous variable in their analyses. Typically the continuous exposure will be partitioned into  $K$  categories and models fit using the categorised exposure instead of the original continuous exposure. The number of categories will depend on the exposure of interest and the sample size. Cutpoints may be chosen to be equally spaced, according to quantiles of the observed data, or with respect to clinically relevant cut-points. Turner, Dobson and Pocock [82] give recommendations for those categorising a continuous exposure. The choice of cutpoints should be considered carefully with respect to the problem under consideration and alternative cutpoints should be considered for sensitivity testing.

For defined groups  $k = 1, \dots, K$  define  $K$  indicators as

$$X_i^{(k)} = \begin{cases} 1 & \text{if the exposure for individual } i \text{ is in the } k\text{th category} \\ 0 & \text{otherwise} \end{cases} .$$

Under the grouped exposure analysis the log-hazard at time  $t$  is modelled as

$$\log h(t|X^{(2)}, \dots, X^{(K)}) = \log h_0(t) + \beta_1 X^{(2)} + \beta_2 X^{(3)} + \dots + \beta_{K-1} X^{(K)} \quad (2.5)$$

where the hazard ratios are relative to a baseline hazard of  $k = 1$ .

This approach is easy to implement and makes no assumption about the shape of the exposure–disease relationship. A plot of the log-hazard within each category against the mean exposure value within that category allows us to explore the shape of the exposure–disease relationship.

## 2. Modelling non-linear exposure–disease relationships

---

Groups also allow us to easily include a zero exposed group which can be difficult to include under other modelling approaches.

The results obtained from a grouped exposure analysis are usually plotted with confidence intervals that are relative to a reference group which has zero variance, with the other groups sharing a common component due to the variation in the reference group. If the choice of reference group contains few events then this variation can be large and obscure the differences between the non-reference groups. This can be somewhat alleviated by choosing the reference group to be that with the greatest number of observations. In practice, however, it may be desirable to have a different reference value than that with the greatest number of observations.

Floating absolute risks (also known as quasi-variances [83]) developed by Easton, Peto, and Babiker [84] have been developed to display the uncertainty of the risk within groups without reference to another group. Their use has been controversial, for example Greenland *et al.* [85] commented that floating absolute risks do not have the coverage properties of the confidence intervals from which they are constructed which could lead to misinterpretation by those unfamiliar with the procedure. Plummer [86] points out that the confidence interval is for  $\gamma = \alpha_i - \hat{\alpha}_1$  where  $\alpha_i$  is the *absolute* risk for group  $i$ .

Easton *et al.* proposed two approaches to calculating floating absolute risks. The first method, augmented likelihood, is specific to conditional logistic regression hence we describe their alternative heuristic method here. The general aim is to find  $\{q_k\}_{k=1\dots K}$  such that for all contrasts,  $c$  [83],

$$\widehat{\text{Var}}(c^T \hat{\beta}) \simeq \sum_{k=1}^K c_k^2 q_k.$$

We can achieve this for simple contrasts  $\beta_j - \beta_k$  by minimising

$$\sum_{j < k} p(q_j + q_k, v_{jk})$$

over  $\{q_k\}$ , where  $p(q, v)$  is a suitable penalty function and  $v_{jk} = \text{Var}(\hat{\beta}_j) + \text{Var}(\hat{\beta}_k) - 2 \text{Cov}(\hat{\beta}_j, \hat{\beta}_k)$ . Penalty functions include the squared-error penalty [84] which aims to control the absolute error

$$(q_j + q_k, v_{jk}) = (q_j + q_k - v_{jk})^2.$$

Controlling the absolute error is not preferable because it can lead to large relative errors for categories with small variance. The penalties suggested by Ridout [87],

$$p(q_j + q_k, v_{jk}) = \left( \frac{q_j + q_k}{v_{jk}} \right)^{-\frac{1}{2}} - 1 + \frac{1}{2} \log \left( \frac{q_j + q_k}{v_{jk}} \right)$$

and Firth and Menezes [83],

$$p(q_j + q_k, v_{jk}) = \left( \log \left( \frac{q_j + q_k}{v_{jk}} \right) \right)^2$$

instead control the relative error.

## 2.2.2 Polynomials

If the exposure–disease relationship is non-linear then we can add polynomial terms to our model to try and capture this

$$\log h(t|X) = \log h_0(t) + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k.$$

Polynomial terms are often included in an *ad hoc* fashion without adequate consideration of alternative powers which may provide a better fit. Polynomials tend to give functions that extrapolate badly. A more principled approach is to consider a range of transformations from Tukey’s ladder of powers [88]. Durazo-Arvizu *et al.* [89] used this approach to choose a suitable transformation of BMI when looking at the association between BMI and all-cause mortality.

An alternative to selecting a power from Tukey’s ladder is to use a Box-Tidwell transformation [90],

$$X^{(p)} = \begin{cases} X^{(p)} & \text{if } p \neq 0 \\ \log X & \text{if } p = 0 \end{cases}. \quad (2.6)$$

The logarithmic transformation for the case  $p = 0$  is obtained by setting  $p_0 = 0$  in the Taylor expansion of  $X^p$  about  $p_0 = p + \epsilon$  for small  $\epsilon$ ,

$$X^p = X^{p_0} + (p - p_0)X^{p_0} \log X + O((p - p_0)^2). \quad (2.7)$$

The optimal value of  $p$  may be found using the following procedure [90]:

1. Let  $p = 1$
2. Fit  $\log h(t|X) = \log h_0(t) + \beta_1 X^p$
3. Fit  $\log h(t|X) = \log h_0(t) + \beta_1^* X^p + \beta_2^* X^p \log X$
4. Set  $p \leftarrow \frac{\beta_2^*}{\beta_1^*} + p$
5. Repeat steps 2–4 until convergence.

Although this procedure usually converges within three iterations, it unfortunately does not

## 2. Modelling non-linear exposure–disease relationships

---

always converge.

### 2.2.3 Fractional polynomials

Fractional polynomials were introduced by Royston and Altman [91] as a method for flexibly modelling continuous exposure–disease relationships using parametric modelling. They have become widely used with software for fitting fractional polynomial models available in SAS [92], STATA [93] and R [94]. For example Baglietto *et al.* used fractional polynomials to model the relationship between alcohol consumption and all cause mortality and Abbas *et al.* [95] used them to model the relationship between serum 25-hydroxyvitamin D and risk of post-menopausal breast cancer.

A fractional polynomial of degree  $m$  is given by

$$\log h(t|X) = \log h_0(t) + \sum_{j=1}^m \beta_j H_j(X)$$

where

$$H_j(X) = \begin{cases} X^{(p_j)} & \text{if } p_j \neq p_{j-1} \\ H_{j-1}(X) \log X & \text{if } p_j = p_{j-1} \end{cases} \quad (2.8)$$

and,  $X^{(p_j)}$  is the Box-Tidwell transformation given in equation 2.6. The definition of  $H_j(X)$  is based on the first order Taylor expansion given in equation 2.7.

It is suggested that  $\mathbf{p} = \{p_1, \dots, p_m\}$  only contains members from the set  $\varphi = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ . This set gives a wide-range of epidemiologically plausible shapes for the exposure–disease relationship whilst keeping the set of powers to be considered small; and contains many commonly used transformations such as quadratic, square root and logarithmic as well as the linear model. Usually a fractional polynomial of degree 1 (FP1) or degree 2 (FP2) is found to provide an adequate fit to the data. Logarithmic transformations are often applied to exposures, but since the log transformation is considered in the model selection procedure, it is suggested that the data are not log transformed *a priori*. If a logarithmic transformation is preferable then this will be selected by the model selection procedure.

It is necessary to choose between the set of possible fractional polynomial models. There are a total of 36 models up to FP2 using the set of powers  $\varphi$ . Ambler and Royston [96] propose a closed test procedure for model selection for fixed  $m$  and significance level  $\alpha$ . For a fractional polynomial of maximum permissible degree 2 the algorithm proceeds, starting with the best fitting fractional polynomial of degree 2, through the following steps:

1. Inclusion—For covariate  $X$  the best fitting degree 2 fractional polynomial is tested for possible omission of  $X$  from the model. If  $X$  is significant (in terms of its likelihood

ratio statistic) then continue, otherwise  $X$  is dropped from the model and we proceed to the next covariate.

2. Non Linearity—Test the best fitting fractional polynomial of degree 2 against the model which is linear in  $X$ . If this test is non-significant then the model is linear in  $X$  and we can continue to the next covariate, otherwise continue.
3. Simplification—Test the best fitting fractional polynomial of degree 2 against the best fitting fractional polynomial of degree 1. If the test is significant choose the fractional polynomial of degree 2, otherwise choose the fractional polynomial of degree 1.

The number of degrees of freedom for each of the above tests is 4 for  $m = 2$  and 2 for  $m = 1$ . That is 1 degree of freedom for each parameter and 1 degree of freedom for each power, although this is an overstatement since the powers are chosen from the limited set  $\wp$ . Ambler [96] showed that this overstatement of the number of degrees of freedom leads to underestimated type I error rates when the exposure–disease relationship is truly null or linear.

Covariate values must be positive in a fractional polynomial analysis since the set of power transformations  $\wp$  contains fractional powers. If a covariate does take negative values then we must transform the data. One proposal [91] is

$$X' = X - X_{\min} + \gamma \quad (2.9)$$

where  $\gamma$  is chosen to be the smallest difference between successive ordered observations although Royston and Altman [91] comment that the fractional polynomial selected is relatively insensitive to the exact choice of  $\gamma$ . The choice of origin is important as the selected fractional polynomial powers can be highly correlated with the choice of origin [91] and can affect the type I error rate of the procedure [96]. If the distance between the mean of the covariate and the origin is large then we are increasingly likely to select fractional polynomial functions that are approximately quadratic. The origin–shift transformation was proposed to deal with this problem [97],

$$X^* = \delta + (1 - \delta) \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where  $\delta$  is small. The transformation shifts and scales the values so that they lie in the range  $[\delta, 1)$ . A value of  $\delta = 0.2$  is recommended as  $X^*$  needs to be sufficiently bounded away from zero to ensure that small exposure values do not create extremely large values under negative power transformations.

## 2.2.4 Splines

Spline functions are used in many applications such as computer graphics and in numerical analysis. In statistics their use in regression models has increased markedly in recent years. Splines can combine a good local fit with desirable global properties such as continuity for a

## 2. Modelling non-linear exposure–disease relationships

---

specified number of derivatives. We shall describe spline bases which are the building blocks of splines, the two main types of splines that are used in regression analyses — regression splines and smoothing splines — and the issues regarding knot selection and number of degrees of freedom of spline models.

A spline on the interval  $[a, b]$  is a piecewise polynomial function and the endpoints of the subintervals on which each of the polynomial functions is defined are called knots. The degree of a spline is the power of the highest order polynomial term. Hence, a degree 1 spline consists of linear segments between each knot.

### Spline basis functions

A set of spline basis functions  $\mathfrak{S}(X) = \{S_1(X), \dots, S_m(X)\}$  on  $[a, b]$  are a set of locally defined polynomial functions (i.e. the functions are zero outside a specified range) which can be combined in a linear combination to form a spline function

$$S(X; \boldsymbol{\beta}) = \sum_j \beta_j S_j(X).$$

We can fit a spline model by generating a spline basis,  $\mathfrak{S}(X)$ , for our covariate  $X$  and including  $\mathfrak{S}(X)$  as covariates in our regression models. There are several different ways of generating spline bases. We consider the widely used power spline and B-spline bases.

The power spline basis of degree  $q$  with knots  $t_0, \dots, t_K$  (which may or may not coincide with the observed  $x_i$ ) is defined using a truncated power function

$$S_j(X) := (X - t_j)_+^q = \begin{cases} (X - t_j)^q, & X \geq t_j \\ 0, & X < t_j \end{cases} \quad (2.10)$$

and has the following properties:

- It is a spline of degree  $q$ ,
- At  $t_j$  the derivatives up to order  $(q - 1)$  are zero and continuous,
- The  $q$ th derivative does not exist at  $x = t_j$ .

We can then fit the spline regression model

$$\log h(t|X) = \log h_0(t) + \beta_1 X + \dots + \beta_q X^q + \sum_{j=1}^q \beta_{q+j} (X - t_q)_+^q.$$

Although the power spline basis is easy to generate there are two key problems with using this basis [98]:

- If the number of knots,  $k$ , is ‘large’ then the value of the spline  $S(x; \beta)$  at  $x$  can involve a lot more than  $q$  of the  $\beta$  coefficients. This means that calculations performed using this basis can be computationally expensive
- If the knots,  $t_0, \dots, t_k$ , are very non-uniformly distributed then some of the basis functions may become nearly linearly dependent on one another. This can lead to ill conditioned systems of equations for calculating the coefficients.

Basis-spline, or B-spline, bases are the most commonly used because they do not suffer from the problems mentioned above. They are called basis-splines since every spline function of a given degree, smoothness, and knots, can be represented as a linear combination of B-splines of that same degree and smoothness.

For a spline basis on  $k$  knots the  $k - 1 - q$  basis B-splines of degree  $q$  can be obtained using the Cox-de Boor recursion [98],

$$b_{j,0}(t) := \begin{cases} 1 & \text{if } t_j \leq t < t_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

$$S_j(t) := b_{j,q}(t) = \frac{t - t_j}{t_{j+q} - t_j} b_{j,q-1}(t) + \frac{t_{j+q+1} - t}{t_{j+q+1} - t_{j+1}} b_{j+1,q-1}(t).$$

Each B-spline has the following properties [99]:

- It consists of  $q + 1$  polynomial pieces, each of which is of degree  $q$ ,
- At the points where the polynomial pieces join the first  $q - 1$  derivatives are continuous,
- The B-spline basis function is positive on a domain spanned by  $q + 2$  knots and zero elsewhere (which is a major advantage over power splines for large bases),
- Except at the boundaries it overlaps with  $2q$  polynomial pieces of its neighbours.

A B-spline regression model is of the form

$$\log h(t|X) = \log h_0(t) + \sum_{j=0}^{k-q-1} \beta_j b_{j,q}(X).$$

If the knots are evenly spaced then the basis is described as uniform. Uniform B-spline bases are computationally advantageous since all the basis functions are simply shifted copies of each other. At each point the sum of the B-spline basis functions equals 1, hence why one of the basis functions is omitted from the model. Figure 2.1 shows examples of degree 3 power and B-spline bases.

Regression spline models are fitted to a set of spline basis functions on  $k$  knots (typically

## 2. Modelling non-linear exposure–disease relationships

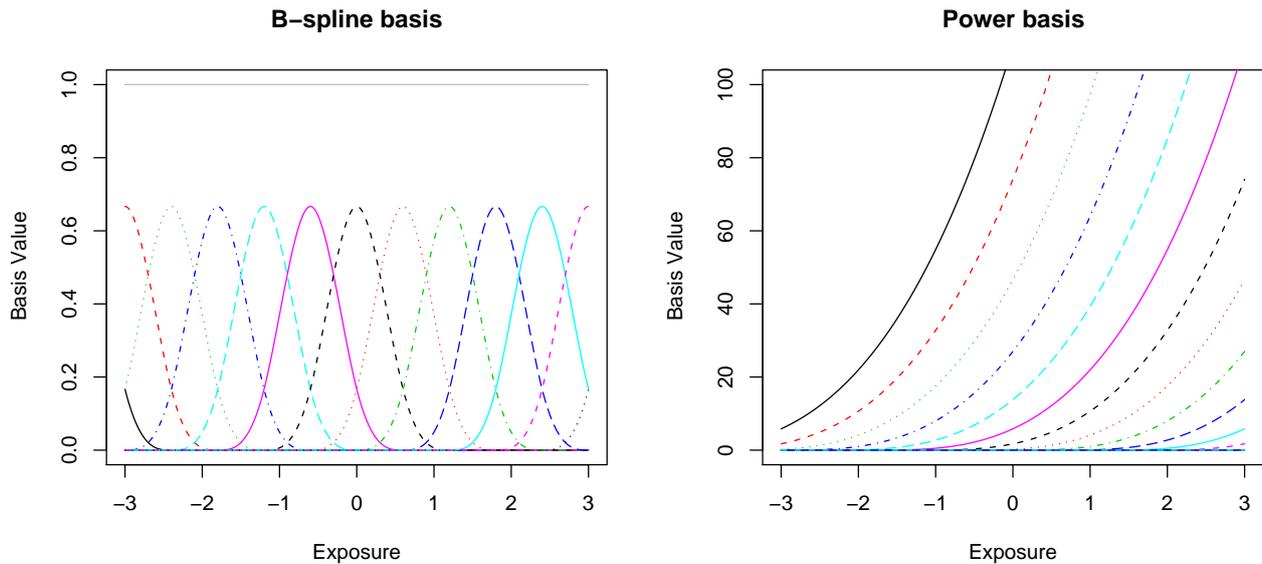


Figure 2.1: B-spline (with sum of basis functions [grey line]) and power basis functions.

$k \ll n$ ). The advantage of these models is that they can easily be fitted in standard software packages and will return a small set of parameters. Since a small number of knots are used to effectively control the roughness of the fitted function we lose some of the flexibility that spline functions offer. If a very small number of knots are used this will produce models that are unrealistic for the true underlying exposure–disease relationship. Choosing the number and location of knots is a nontrivial problem and a lot of the regression spline literature focuses on this challenge.

### Cubic splines

One of the most commonly used forms of regression spline are restricted cubic splines, otherwise known as natural cubic splines. Restricted cubic splines have continuous first and second derivatives and are linear outside the first and last knots. The R function `ns` in the `splines` package will produce a B-spline basis for a restricted cubic spline.

Heinzl and Kaider [100] demonstrate the use of a restricted cubic spline in the relationship of the drug 6-mercaptopurine on remission of acute leukaemia and Desquilbet and Mariotti [101] the association between cardiovascular disease and cholesterol.

### Smoothing splines

Smoothing splines have  $n$  knots, one at each of the observed  $X_i$ . A penalty is imposed on the roughness of the fitted function. The penalty is introduced into the estimation of  $\beta$  by penalising the (partial) log-likelihood of the model. Denote the unpenalised log-likelihood for the regression spline model by  $\ell(x; \beta)$  and the penalised log-likelihood by  $\ell_p(x; \beta, \theta)$  where  $\theta$  is the smoothing parameter which controls the balance between smoothness and adherence to

the data. Then

$$\ell_p(x; \boldsymbol{\beta}, \theta) = \ell(x; \boldsymbol{\beta}) + \theta \int_{x_{min}}^{x_{max}} (S''(x; \boldsymbol{\beta}))^2 dx.$$

As  $\theta$  tends to infinity the spline function becomes a line and as  $\theta$  tends to zero we obtain the unpenalised likelihood. Amongst all spline functions this penalty function is minimised by the set of natural cubic splines with knots at the distinct values of  $x$  [102]. We can express the penalty  $\int_{x_{min}}^{x_{max}} (S''(x; \boldsymbol{\beta}))^2 dx$  as  $\boldsymbol{\beta}^T \mathbf{D} \boldsymbol{\beta}$  where the  $(i, j)$ th element of  $\mathbf{D}$  is given by

$$D_{i,j} := \int_{x_{min}}^{x_{max}} S_i''(x) S_j''(x) dx.$$

This matrix representation of the penalty is convenient for computational reasons and is used by Reinsch [103] in his algorithm for fitting smoothing splines.

A problem with smoothing splines is that as  $n$  increases so does the size of the basis at the same rate, and hence the number of parameters that have to be fitted. Even for relatively modest sample sizes the computation involved in fitting smoothing spline models can be large; this is especially so for the calculation of standard errors. For example the `smooth.spline` function in R does not by default use all the data points as knots if the number of observations is greater than 49.

## P-splines

P-splines are an alternative to smoothing splines and combine the smaller number of knots of regression splines with the penalty on the roughness of the fitted function of smoothing splines. They can be fit using the `pspline` function of the `survival` package in R, although implementations do not exist in STATA or SAS. P-splines were used by Eisen *et al.* [104] to model the relationship between silica exposure and risk of lung cancer mortality and by Hillege *et al.* [105] to model the relationships of urinary albumin excretion with cardiovascular and non-cardiovascular mortality.

P-splines [99] minimise

$$\ell_p(x; \boldsymbol{\beta}, \theta) = \ell(x; \boldsymbol{\beta}) + \theta \boldsymbol{\beta}^T \mathbf{D}_2^T \mathbf{D}_2 \boldsymbol{\beta}$$

where  $\ell(x; \boldsymbol{\beta})$  is the unpenalised log-likelihood for the regression spline model constructed from a uniform B-spline basis for  $X$ , and  $\mathbf{D}_2$  is the matrix of second differences which is explained below. The penalty is such that it penalises non-linearity in the sequence of  $\boldsymbol{\beta}$  coefficients. Although this is the penalty used in practice the penalty could be any symmetric, non-negative definite matrix. P-splines are computationally much easier to fit than smoothing splines because we have a relatively small number of knots. If the knots are evenly spaced then the only dependence of the penalty on the underlying data is through the use of the range of  $X$

## 2. Modelling non-linear exposure–disease relationships

---

to generate the spline basis.

The matrix of second differences,  $D_2$ , is generated using the difference operator,  $\nabla$ . For an  $n \times p$  matrix  $X$ , the differencing operator  $\nabla$  is defined by

$$\nabla X_{i,j} = X_{i+1,j} - X_{i,j}, \quad 1 \leq i \leq n-1$$

which gives an  $(n-1) \times p$  matrix.

$D_k$ , the  $(n-k) \times n$  matrix of  $k$ th differences, is given by  $\nabla^k I_n$  where  $I_n$  is the  $n \times n$  identity matrix and  $\nabla^k$  is the  $\nabla$  operator applied  $k$  times (i.e.  $\nabla^2 X = \nabla \nabla X$ ). For example,  $D_2$  for  $n = 6$  is calculated as

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{\nabla} \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \xrightarrow{\nabla} \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{pmatrix}.$$

Because of the penalisation used with smoothing and P-splines the effective degrees of freedom of the model is not equal to the number of coefficients estimated in the model. Suppose we have a Cox model where the exposure of interest is modelled using a penalised spline with  $r$  basis functions and that we have  $s$  additional covariates. Let  $\beta = (\beta_u, \beta_p)$  be the parameter vector of this model where  $\beta_p$  is the  $1 \times r$  vector of coefficients of the penalised spline, and  $\beta_u$  is the  $1 \times s$  vector of coefficients relating to the additional covariates. We maximise the penalised partial log likelihood,  $\ell_p$  where [79],

$$\ell_p(\beta; \theta) = \ell(\beta) - g(\beta_p; \theta)$$

where  $g(\beta_p; \theta)$  penalises the appropriate likelihood terms corresponding to the penalised terms. Then the information matrix,  $H$ , for this model is given by

$$H = I - \begin{pmatrix} 0 & 0 \\ 0 & -g'' \end{pmatrix} = I + G$$

where  $I$  is the information matrix for the unpenalised log likelihood. Gray [106] suggests that the variance-covariance matrix for the parameter estimates is

$$V := H^{-1} I H^{-1}$$

with the effective degrees of freedom given by  $\text{trace}(\mathbf{H}\mathbf{V})$ ,

$$\text{trace}(\mathbf{H}\mathbf{V}) = \text{trace}(\mathbf{H}\mathbf{H}^{-1}(\mathbf{H} - \mathbf{G})\mathbf{H}^{-1}) = (p + q) - \text{trace}(\mathbf{G}\mathbf{H}^{-1}).$$

Under the null hypothesis the Wald statistic  $(\boldsymbol{\beta}_u, \boldsymbol{\beta}_p)^T \mathbf{H}^{-1}(\boldsymbol{\beta}_u, \boldsymbol{\beta}_p)$  will be asymptotically equivalent to  $\sum_{i=1}^p e_i Z_i^2$  where the  $e_i$  are the eigenvalues of the matrix  $\mathbf{H}\mathbf{V}$  and the  $Z_i$  are independent identically distributed standard normal random variables. In unpenalised regression  $\sum e_i = \dim((\boldsymbol{\beta}_u, \boldsymbol{\beta}_p)) = p + q$  and the Wald test statistic has a  $\chi_{\sum e_i}^2$  distribution. In penalised models  $\sum e_i$  equals  $\text{trace}(\mathbf{H}\mathbf{V})$  the effective degrees of freedom of the model. However, strictly the Wald statistic has a distribution with mean  $\sum e_i$  and variance  $2 \sum e_i^2$ . Since  $2 \sum e_i^2 < 2 \sum e_i$  the conservative standard Chi-square test can be used in practice with the effective degrees of freedom above.

There are several methods for trying to find an ‘optimal’ value of the smoothing parameter. These include [107]:

- Akaike’s Information Criterion (AIC)

$$AIC(\theta) = -2\ell_p(\boldsymbol{\beta}; \theta) + 2\text{trace}(\mathbf{H}\mathbf{V}) \quad (2.11)$$

AIC tends to undersmooth and lead to highly variable choices of smoothing parameter so Hurvich *et al.* [108] provide an alternative ‘corrected’ AIC or cAIC

$$cAIC(\theta) = -2\ell_p(\boldsymbol{\beta}; \theta) + \frac{r(\text{trace}(\mathbf{H}\mathbf{V}) + 1)}{r - (\text{trace}(\mathbf{H}\mathbf{V}) + 2)} \quad (2.12)$$

Both AIC and cAIC are available within the `pspline` function.

- Bayesian Information Criterion (BIC)

$$BIC(\theta) = -2\ell_p(\boldsymbol{\beta}; \theta) + \text{trace}(\mathbf{H}\mathbf{V}) \log n \quad (2.13)$$

- Generalised Cross validation (GCV)

$$GCV(\theta) = \frac{\ell_p(\boldsymbol{\beta}, \theta)}{(n - \text{trace}(\mathbf{H}\mathbf{V}))^2} \quad (2.14)$$

Malloy *et al.* [107] carried out a simulation study investigating the choice of  $\theta$  for model selection in Cox models with penalised splines and found that no method was universally superior.

An alternative to using a model selection procedure is to fix the effective degrees of freedom, which implies a set value of the smoothing parameter. We can find the value of the smoothing parameter by adjusting  $\theta$  until the specified number of degrees of freedom is met. There is no theoretical justification for fixing the degrees of freedom *a priori* however a choice of 4 degrees

## 2. Modelling non-linear exposure–disease relationships

---

of freedom seems to work well in practice and is the default choice in the `pspline` function. Govindarajulu *et al.* [109] found that 4 degrees of freedom were preferable to selection by AIC.

How many knots to choose (i.e. the size of  $k$ ) and the location of the knots is an issue that needs to be considered. By reducing the number of knots the aim is to reduce the complexity of the computation required in fitting the splines but we must maintain the spline’s flexibility to fit the data adequately well. Gray [106] suggests that the number of knots should be at least twice the number of degrees of freedom of the model. In R’s [110] `pspline` function the number of knots is set to  $2\frac{1}{2}$  times the degrees of freedom. Ruppert [111] notes that the cost associated with using too few knots is much greater in terms of bias and mean average squared error than using too many and hence in most practical situations it is not an issue to worry about.

Many methods have also been given for the positioning of the knots including uniform knot spacing, quantiles and more complicated methods such as that of Friedman and Silverman [112] which uses a forward stepwise procedure. Uniform knot spacing is computationally advantageous in forming the B-spline basis whilst sample quantiles have the advantage that they allow the spline greater flexibility in the regions where there is most data.

Likelihood ratio tests for non-linearity can be performed for P-splines. If we consider the power basis representation of a P-spline for  $q = 3$ ,

$$\log h(t|X) = \log h_0(t) + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \sum_{j=1}^9 \beta_{j+3} (X - t_j)_+^3$$

we can clearly see that the linear model is nested within the spline model and occurs when  $\beta_2 = \beta_3 = \dots = \beta_{12} = 0$ . Since we can find a power basis and B-spline representation of the same space we can use a likelihood ratio test (using the unpenalised likelihood) to test for non-linearity [113].

### 2.2.5 Other methods for modelling continuous non-linear exposure–disease relationships

The methods described above are by far the most widely used, although other methods have been proposed. Change-point models are sometimes used when a threshold is hypothesised. These models usually involve two linear segments of differing slopes with a knot connecting them. Much focus is placed on the location of the knot; finding the correct location is a difficult problem. Change-point models are a special case of regression splines with only one internal knot.

Binder and Sauerbrei [114] have proposed a procedure which they call FP+L which combines a fractional polynomial fit with local polynomials that pick-up any local features of the data

---

that the FP model missed. Royston and Sauerbrei [115] have also devised a model selection procedure similar to that for fractional polynomials for choosing an appropriate restricted spline model; they call this procedure multivariable regression splines (MRVS). Under this approach instead of finding the optimal powers for the model as in a fractional polynomial analysis, they find the optimal set of knots from a fixed set.

Royston [116] proposed using the power-exponential transformation  $\exp(kx^p)$  for  $x > 0$  which produces monotonic exposure–disease relationships and gives the Box-Tidwell family of transformations when  $p = 0$  and when  $k = 0$ . This class of model is useful if a true monotonic relationship is strongly supported by expert subject knowledge.

### 2.2.6 Discussion on choice of the linear predictor

Figure 2.2 shows an example of each of the main methods described above applied to a simulated dataset of 50,000 individuals with exposure values generated from a normal distribution mean 10, variance 1. Survival times were generated for each individual according to the true quadratic exposure–disease relationship shown by the dot-dashed line, with approximately 10% of individuals experiencing events during  $0 < t < 10$ , when all other observations were censored.

We can see from figure 2.2 that a grouped exposure analysis does not provide a realistic model for the exposure–disease relationship which is, in reality, continuous. It is also a poor fit to the true relationship when compared to many of the other models considered in the figure. Grouped exposure analyses are sensitive to the choice and number of cutpoints. Grouping can lead to masking of interesting features of the data such as a threshold or features in the tails of the exposure distribution because the effect is averaged over all observations in the group. We can see in figure 2.2 that the location of the nadir of the relationship is unclear. Categorisation can also lead to a loss of efficiency over parametric analyses [117].

Groups are usually interpreted as representing the average-risk within groups, even though this is usually not true. Greenland [118] argues that this average-risk interpretation of groups, within a grouped exposure analysis is questionable and proposes that continuous alternatives should be considered such as fractional polynomials and splines [119]. Weinberg [120] agrees that these alternatives should play a larger rôle in epidemiologic investigations but questions whether Greenland’s harsh view on the inadequacy of grouping data is warranted.

The results of a grouped exposure analysis may not lie on the curve given by the true underlying exposure-disease relationship if the exposure–disease relationship is non-linear. Suppose that the underlying continuous exposure–disease relationship is given by

$$\log(h(t|X)) = \log h_0(t) + \beta g(X)$$

## 2. Modelling non-linear exposure–disease relationships

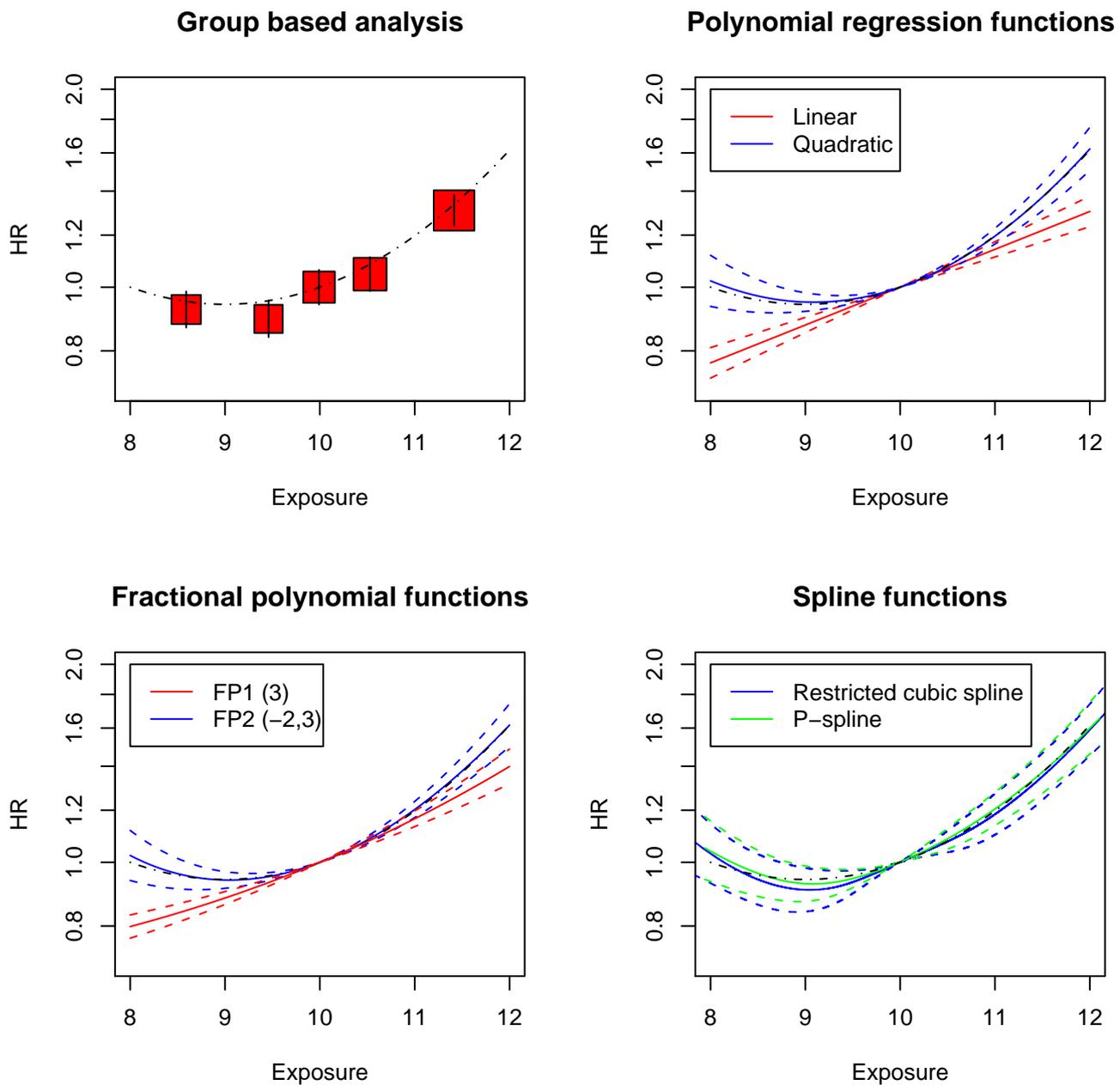


Figure 2.2: Different functional forms for the linear predictor in a Cox proportional hazards model. Dot-dashed line represents the true exposure–disease relationship

where  $g(X)$  is a function of  $X$ , and that we fit the grouped exposure analysis of equation 2.5 to this relationship. The true log-hazard ratio at the average exposure within the  $k$ th group is  $\beta g(E(X_i|X_i^{(k)} = 1))$ . However, under the grouped exposure analysis we observe  $\beta E(g(X_i)|X_i^{(k)} = 1)$ . The two are only equal if the true exposure–disease relationship is linear, or if the risk is constant within groups.

For example, consider a true quadratic relationship where  $\log h(t|X) = \log h_0(t) + \beta X^2$  then

$$E(X_i^2|X_i^{(k)} = 1) = E(X_i|X_i^{(k)} = 1)^2 + \text{Var}(X_i|X_i^{(k)} = 1)$$

so the response for the  $k$ th group will lie above the response for the true relationship at the group mean because of the variance term. The difference in response between the  $k$ th group and the true relationship will be  $\beta \text{Var}(X_i|X_i^{(k)} = 1)$  since the true response for  $E(X_i|X_i^{(k)} = 1)$  is  $\beta E(X_i|X_i^{(k)} = 1)^2$ . This variance is typically not constant across groups, so the difference between the plotted response for each group and the true relationship will be the difference between the variance of that group and that of the group with the smallest variance. This does not seem to be a widely acknowledged phenomenon.

For a normally distributed exposure the differential in variance between the groups most distant from the mean compared with those close to the mean can be large. If one were using quintiles of the data the variance of the top 20% of observations is over 10 times that of those in the central 20% (i.e. those observations between the 40th and 60th percentiles).

If the cut points are at the quintiles of the standard normal distribution then,

Quintile	$(-\infty, -1.400]$	$(-1.400, -0.532]$	$(-0.532, 0.000]$	$(0, 1.400]$	$(1.400, \infty)$
Variance of exposure within quintile	0.219	0.028	0.021	0.028	0.219

If instead of using quintiles of the data we cut the data at the 2.75, 27, 73 and 97.25 percentiles then under the standard normal distribution,

Exposure group	$(-\infty, -2.302]$	$(-2.302, -1.102]$	$(-1.102, 0.000]$	$(0, 1.102]$	$(2.302, \infty)$
Variance of exposure within group	0.119	0.119	0.119	0.119	0.119

which would eliminate the difference between the response at the means of the grouped exposure analysis, and the true underlying quadratic exposure–disease relationship. However, these groupings are not practical since 46% of the exposure will lie in the central group. Also, within small samples you are unlikely to achieve equal variance within groups using these cutpoints because quantiles in the tails can be highly variable.

Despite these downsides a grouped exposure analysis may be useful for initial exploration of the exposure–disease relationship in a Cox model. In Cox models we cannot easily visualise the data, unlike with linear regression where we can use a scatterplot, because the outcome consists of the event indicator and time.

In panel 2 of figure 2.2, we see the quadratic model accurately picks up the true exposure–disease relationship. This is not surprising since the true relationship is quadratic but in general polynomial functions can have undesirable features at the extremes of the exposure distribution such as turning points and are still quite restrictive in the range of shapes they can produce; for example they cannot reproduce asymptotic or threshold relationships. Royston *et al.* [121] give an example of a quadratic model for the relationship between all-cause mortality and number of cigarettes smoked per day which gave decreasing risk for those who consumed more than

## 2. Modelling non-linear exposure–disease relationships

---

about 30 cigarettes per day; a result which is epidemiologically implausible.

A key advantage of polynomials is the ease with which they can be fitted. However, choosing an appropriate polynomial is not straightforward. Fractional polynomials essentially formalise and automate the selection of suitable transformations. In panel 3 of figure 2.2 we see that the FP2 model accurately recreates the true exposure–disease relationship, however it does so using the powers  $(-2, 3)$ . The confidence intervals for fractional polynomials do not take account of the uncertainty in the choice of  $p$  nor the wider problem of model uncertainty [121]. This problem has, however, been considered by Faes *et al.* [122] who apply model averaging to fractional polynomial models.

Instead of selecting powers from the limited set  $\wp$  we could instead select the optimal Box-Tidwell transformation. This is essentially a continuous version (in terms of powers) of fractional polynomials. Presumably this would have the correct type I error rate for testing non-linearity. This type of model can be fit using `boxtid` in STATA, where a fractional polynomial model is fitted to give appropriate starting value(s) for the search.

In panels 3 and 4 of figure 2.2 we see that both the restricted cubic splines and P-spline curves (each with 4 (effective) degrees of freedom) give good fits to the true relationship. The main drawback of splines is that they cannot be written in a compact mathematical form. This means that the relationship can typically only be presented in graphical form. Also, it is not clear which model selection criterion should be used to produce a model that fits the data adequately but is not too ‘wiggly’ and there are few implementations that allow users to fit and plot the models easily. We give R code for plotting P-spline functions in appendix F.

Govindarajulu *et al.* [109] compared the results from restricted cubic splines, fractional polynomials and penalized splines applied to silica exposure and lung cancer mortality. They found that the two spline methods gave similar curves but that a fractional polynomial gave a very different shape for the exposure–disease relationship. This is probably because the exposure is heavily skewed, with the majority of individuals having low exposure values leading to significant uncertainty in the tail regions. Govindarajulu *et al.* [123] then compared the performance of P-splines, restricted cubic splines and fractional polynomials in a simulation study and found that model fit was best for P-splines for almost all shapes of the exposure–disease relationship considered; although fractional polynomials and restricted cubic splines were less biased (in terms of area between the true and fitted curves), on average.

Traditional goodness of fit tests, such as the likelihood ratio test, are a global measure of fit. Often in epidemiological studies we are interested in retaining local features of the data such as features in the tail of the distribution which (fractional) polynomials and grouped exposure analyses can ignore. Royston [124] suggests comparing a smoothing spline with a relatively large number of degrees of freedom to the lower-dimensional fitted function chosen to ensure that important features of the exposure–disease relationship are not missed.

---

An advantage to using a grouped exposure analysis is that you can then produce a table of hazard ratios for individuals within each interval of exposure. Royston *et al.* [121] give advice on how to present the results from analyses of continuous exposure–disease relationships in a similar way. A table combined with a graph of the continuous exposure–disease relationship and its confidence interval can give more information to readers than that from a grouped exposure analysis.

### 2.2.7 Confounding

Above we have considered the functional form for the exposure of interest within our model. We now consider the role of confounders and why we should control for their effects in our model.

A confounder is a variable that is correlated with both the outcome and the exposure of interest but does not lie on the causal pathway between them. Confounding can cause the observed effect size to be larger or smaller than the true effect size and can even make the observed effect size go in the opposite direction. For example, alcohol may be found to be associated with the risk of lung cancer but this may be due to confounding caused by exposure to tobacco smoke which is associated with alcohol assumption and lung cancer (especially before the ban on smoking in public places) [125].

It is therefore important to take account of confounding, especially when the exposure effect is relatively small compared with that of the confounder(s) we are controlling for. In matched case-control studies we can match individuals based on the levels of confounding variables. In clinical trials we can control for confounding by randomisation. In unmatched case control studies and cohort analyses we adjust for confounding by including the confounders in our regression models. After adjusting for known confounders we may be left with residual confounding—confounding relating to unknown, unmeasured variables that we have not adjusted our analysis for.

There is not always agreement on what set of confounders should be adjusted for in analyses. This may be due to insufficient evidence showing that a potential confounder is associated with the disease or because there may be a debate as to whether a potential confounder is independently associated with the disease rather than on the causal pathway. It is often the case that sensitivity analyses are produced to show the effect of different levels of adjustment for confounding.

All of the methods described in this chapter for including exposures in the disease model e.g. groups, splines, fractional polynomials can be applied to confounding variables. Typically in practice, however, the most widely used approaches for allowing for confounding variables are the use of linear terms and groups of the confounding variable, even if more complex terms (such as polynomials or splines) are used for the exposure of interest. Inappropriate modelling

of confounders can lead to biased parameter estimates for the exposure of interest.

### 2.3 Conclusion

In this chapter we have considered the Cox model for modelling time to event data and possible forms for the linear predictor when we have a continuous exposure and a non-linear exposure–disease relationship. We can discretise the continuous exposure and use indicators of group membership, polynomial functions, fractional polynomials or spline functions in our linear predictor.

The grouped exposure method does not provide a realistic model for the exposure–disease relationship and as we have shown introduces parameter bias when the true exposure–disease relationship is non-linear. It can also be sensitive to the choice and number of cutpoints. Polynomial functions can be useful but can have undesirable features such as turning points at the end of the exposure range.

Fractional polynomials use polynomial functions chosen from a limited set of powers which can provide a wide range of plausible exposure–disease relationships.

Spline functions are piecewise polynomial functions that come in many different flavours. The most commonly used spline functions are restricted cubic splines that use a small number of knots and are constrained to have continuous first and second derivatives; they can however be sensitive to the number and placement of knots. P-splines use a greater number of knots than restricted cubic splines and the smoothness of the fitted function is achieved by penalising the likelihood function.

Confounding occurs when we have a variable that is correlated with the outcome and the exposure of interest, but does not lie on the causal pathway between them. We need to adjust for the effects of confounding so that the observed effect size for our exposure is independent of the other variables in our model.

We introduce exposure measurement error and consider its effects on a range of epidemiologically plausible shapes for the exposure–disease relationship in chapter 3 before considering how we may correct for the effects of exposure measurement error in later chapters.

# Chapter 3

## The effects of exposure measurement error

In this chapter we introduce measurement error and consider models for the measurement error process. We then consider the effects of classical measurement error on the shape of the observed exposure–disease relationship, and our power to detect non-linearity, using simulation. We simulate true exposure values from a range of epidemiologically plausible shapes for the exposure–disease relationship (linear, threshold, U-shaped, J-shaped, increasing-quadratic, asymptotic, non-linear threshold). We use our true exposure to and simulate our observed exposure by adding random measurement error. We model the observed exposure–disease relationship using three methods that were described in chapter 2: grouped exposure analyses, fractional polynomials, and P-splines.

### 3.1 Measurement error and misclassification

Throughout this dissertation we will be interested in mismeasurement [51]: measurement error in a quantitative variable or misclassification in a categorical variable. Measurement error will be our primary focus, sometimes also referred to as ‘errors-in-variables’ in older discourse.

Some exposures of interest may be assumed to be free of measurement error, for example age, height and weight, whilst others we may be able to measure less precisely, such as long term blood pressure, energy consumption or alcohol intake. Measurement error is the term we use to describe the difference between our observed exposure measurements and their true values. Measurement error may occur due to our inability to accurately measure the true exposure of interest or within-person variation.

The sources of measurement error can be complex. For example in nutritional epidemiology the food frequency questionnaire (FFQ) is a commonly used method for assessing individuals’ dietary intake. The FFQ asks individuals to report frequency of consumption, and often portion

### 3. The effects of exposure measurement error

---

size, of a range of different foods over a period of time e.g. the last three months. From this information an individual's typical nutrient intake is calculated based on the responses they have given. This has many potential sources of measurement error including:

- Individuals tend to report exposure values closer to social norms. People are likely to over report their exposure of healthy foods such as fruit and vegetables whilst under report their consumption of unhealthy foods such as crisps and sweets;
- Individuals may report no consumption of a food when they do actually consume it—this is particularly the case for episodically consumed foods such as fish;
- Individuals may respond differently depending on how the question is presented;
- Individuals may report different consumption levels for the same period on different days;
- Individuals may find it difficult to remember their consumption over a long time frame;
- Certain foods might not be contained within the survey—e.g. foods consumed by different ethnic groups;
- It is difficult to translate reported intake into accurate measures of nutrient intake as this will depend for example on brand and portion size of the actual food consumed.

These different sources of error may have systematic and random components. The observed exposure measurements will typically be subject to a number of sources of error and within-person variation. We can not expect to be able to identify nor quantify each aspect of it; in fact we usually do not need to.

Misclassification is the analogue of measurement error when we are dealing with a categorical variable. Categorical variables can arise in two ways. Firstly a variable can take a set of discrete values such as an individual being a current smoker or non-smoker, or secondly, we can create a categorical variable from a continuous one by saying that an exposure value belongs to the  $i$ th category if the exposure lies in some interval, where the intervals are non-overlapping and cover the whole exposure range. These two different types of categorical variable tend to have different misclassification properties. Under the first we often assume that all individuals within each group are equally likely to be misclassified. In the second situation it is much more plausible that where the true exposure measurement lies within the category will affect its probability of being misclassified i.e. those close to the category boundary are more likely to be misclassified than those in the centre. In the first case we can use a probability matrix to describe the probability of misclassification, however this cannot be done in the second.

Sometimes our exposure of interest may be subject to misclassification and measurement error. Take for example the reporting of alcohol consumption—there will be misclassification as to whether an individual consumes alcohol or not, and then there will be measurement error

---

amongst those who do consume alcohol. These type of exposures need specific models that take into account the non-exposed category.

It may be possible to reduce the error in our exposure measurements, for example by using more accurate equipment or a different assay procedure. However, in most situations there is little we can do to reduce the measurement error of our observed exposure. In clinical trials we may be able to reduce the impact of measurement error by increasing the variance of our true exposure. We cannot, however, eliminate within-person variability.

There are other sources of error which can lead us to have an inaccurate measure of true exposure, these include:

- only being able to measure an exposure to a given degree of accuracy,
- misrecording of the observed measurement,
- missing data.

We shall not be considering any of these situations in this dissertation.

### **3.1.1 What is the ‘truth’?**

True exposure needs to be carefully defined according to the context; this will often be something that would be very difficult or impossible to actually observe. Sometimes ‘true’ exposure will be usual level of exposure, where usual exposure is defined as long-term average of (possible) observations. What is meant by long-term will depend on the context and usually is not specifically defined. An advantage of using usual exposure is that the measurement error has mean zero by definition. Usual exposure is used a lot in epidemiological studies where exposures such as blood pressure or fasting blood glucose can be subject to substantial within-person variation.

Sometimes there is debate as to what the exposure of interest should be; whether peaks in exposure are more important than long-term average exposure, or in the case of blood pressure whether the difference between diastolic and systolic pressure levels [126] is more important than either of the two individually.

In some situations we may be able to measure true exposure, this is known as a gold standard measurement. We may not be able to obtain gold standard measurements for all patients in a study and have to use some imperfect measure of exposure instead. This situation may arise because:

- of the cost and/or time involved in obtaining the gold standard measure,
- it may not be ethical to obtain a gold standard measure for a large set of individuals.

Sometimes a gold standard measurement may not be available, but an alloyed gold standard

### 3. The effects of exposure measurement error

---

[127] instead. An alloyed gold standard is an exposure measurement that is subject to error, but substantially less than the commonly used method of exposure measurement. The main biological alloyed gold standards used in nutritional epidemiology are:

- urinary potassium — measurement of potassium intake,
- doubly labelled water — measurement of energy expenditure,
- urinary nitrogen — measurement of protein intake.

Alloyed gold standard measurements are normally treated as if they are gold standard measurements in practice.

If we are interested in usual exposure, then continuous exposure measurement would allow us to ascertain usual exposure. In practice continual exposure measurement is rarely feasible, however repeat measurements on an individual's exposure can allow us to better estimate true usual exposure. The average of repeat measurements for an individual will contain considerably less measurement error, than each measurement individually.

When we are able to observe true exposure, or repeat observations, for some individuals we can investigate the relationship between the true and observed exposures. We shall consider this further in chapter 4.

#### 3.1.2 Error models

In epidemiology we are interested in ascertaining an estimate of the true exposure–disease relationship, however we are only able to observe a mismeasured version of the true exposure. We need to be able to model the relationship between the observed exposure and the truth in terms of its components: the true exposure and measurement error. Various models have been proposed which we now discuss. Firstly we shall introduce some notation that we shall use throughout this dissertation. We use the notation of Carroll *et al.* [33]. They note that there is no common notation used in the literature:

- $W$  — observed exposure that is subject to measurement error,
- $X$  — true exposure,
- $U$  — measurement error,
- $Z$  — confounding variable (assumed to be free of measurement error unless specifically stated).

---

## Classical model

The classical measurement error model is the simplest and most widely used model of the measurement error process. The classical measurement error model is given by

$$W = X + U$$

where the measurement error is unbiased i.e.  $E(U|X) = 0$ . Often the measurement error is additionally assumed to follow a normal distribution, since if  $X$  is also normally distributed then  $W$  will be too.

## Berkson Model

An alternative to the Classical model is that of Berkson [128], also sometimes referred to as the control knob model [129]. The Berkson measurement error model assumes

$$X = W + U$$

where  $E(U|W) = 0$ .

Examples of when Berkson error exists are:

- In a clinical trial when a subject is given a pre-specified dose of a drug. The true amount that the subject receives will be the pre-specified amount plus any error in this amount, which could come, for example, from inaccuracies in measuring the dosage.
- In radiation epidemiology where the amount of radiation an individual is exposed to is calculated using an equation based on a number of covariates. The true amount will equal the calculated amount plus equation error.
- In occupational epidemiology when a job exposure matrix [130] is used. These matrices assign a single exposure value to all individuals with the same characteristics. Berkson error is present since the true level of exposure for each individual will vary around the exposure value assigned to the individual from the matrix [131].

It is possible that an exposure can be subject to both Berkson and classical measurement error. For example in Table 1 of Heid *et al.* [29] they list the different sources of measurement error that may have occurred in measuring residential radon exposure and classify them as classical or Berkson. Classical and Berkson error have different effects on the exposure–disease relationship so it is important to try and distinguish between the two sources of error. For example, Berkson error has no effect on the magnitude of the parameter estimates in a linear model when the observed exposure–disease relationship is linear whereas, as we shall later in this chapter, the effect of classical measurement error is to attenuate the relationship.

### 3. The effects of exposure measurement error

---

#### Multiplicative measurement error

Instead of having an additive effect, measurement error may instead have a multiplicative effect. In this case our measurement error model is

$$W = XU$$

where  $U$  is independent of  $X$ , and has mean 1 which gives it the property of being unbiased. Note how this can also be seen as an additive model where the measurement error variance is proportional to  $X^2$ . Multiplicative measurement error can lead to observed values that are very much larger than the true values and hence influential observations in regression analyses. Often in regression analyses the natural logarithm of the exposure values is taken to reduce the effect of influential observations. Under multiplicative measurement error taking the logarithm of the observed data also has the effect of converting the multiplicative errors into additive ones i.e.

$$\log W = \log X + \log U.$$

On the log-scale the measurement error term will be biased since  $E(\log U) < 0$  however in situations where we have no measurement of true exposure available this is of little consequence.

Multiplicative measurement error model has been found to be appropriate for:

- models of radiation exposure e.g. radon epidemiology [132],
- vitamin A in the Nurses Health Study [133],
- anonymising financial data about companies, since it can be effective at disguising particularly large true measurements [134] and retains structural zeroes in the data.

#### Other error models

The classical measurement error model places some assumptions on the relationship between the true and observed exposure which may be violated in a number of ways:

- The observed measurement may not be an unbiased measure of the true exposure. This could be in the form of an additive or multiplicative bias.
- The measurement error component of the observed measurements may also be dependent on the value of covariates  $Z$ .
- The variance of the measurement error may differ between individuals.

---

By allowing for the possibility that each of these violations may occur leads us to consider a more general form of measurement error model [33],

$$W = \alpha_0 + \alpha_1 X + \boldsymbol{\alpha}_z^T \mathbf{Z} + U$$

where  $E(U|X, \mathbf{Z}) = 0$  and  $U$  is uncorrelated with both  $X$  and  $\mathbf{Z}$ . This model is much more plausible for modelling many real world exposures where the relationship between true and observed exposure, and the role of confounders is unknown.

Another model discussed by Carroll *et al.* [33] is that the observed exposure  $W$  has multiple sources of error: some that have an additive effect and some that have a multiplicative effect. This leads us to the measurement error model

$$W = XU_1 + U_2$$

where  $E(U_1|X) = 0$  and  $E(U_2|X) = 0$ . The exact choice of measurement error model and the features it should possess to provide a good fit to the data will depend on the problem at hand.

### 3.1.3 The effects of measurement error

As mentioned in chapter 1 the three main effects of exposure measurement error are: bias in parameter estimation in models; loss of power in detecting relationships between variables; and masking of features of the data [33]. We shall see these effects in action in section 3.4.3. Because of this ‘*triple whammy*’ of effects it is essential to quantify and attempt to correct for the effects of exposure-measurement error if we are trying to ascertain the true exposure–disease relationship, however it is often overlooked [135].

In general, the effect of exposure measurement error is to bias the observed exposure–disease relationship towards the null. Evidence of a non-clinically relevant or a null relationship between the observed exposure and disease may not mean that there is no clinically relevant relationship between the true exposure and disease. The null relationship may result from bias and/or a lack of power to detect the relationship due to the effects of measurement error. Also, confidence intervals for parameter estimates from regressions that use observed exposure measurements will not have the correct coverage of the parameter for the true exposure–disease relationship. Measurement error may make it difficult to compare the results from studies that have measured the same exposure if the amount of measurement error in the exposure measurements varies greatly between studies, and the studies have not been corrected for its effects.

We shall talk about methods for quantifying the measurement error in the observed exposure and methods for correcting for it in chapter 5. However, here we define a common measure of the amount of measurement error present in the observed exposure assuming that we know the true value of the variance of our true and observed exposures. The regression dilution ratio

### 3. The effects of exposure measurement error

---

(RDR),  $\lambda$ , is defined as the ratio of the variance of the true exposure to that of the observed exposure

$$\lambda := \frac{\text{Var}(X)}{\text{Var}(W)} = \frac{\text{Var}(X)}{\text{Var}(X) + \text{Var}(U)}.$$

Under the classical measurement error model the variance of the observed exposure will always be greater than or equal to that of the true exposure so that the RDR will always lie in the range  $(0, 1]$ . An RDR of 1 corresponds to no measurement error in the exposure and the RDR decreases towards zero as the proportion of measurement error in the observed exposure increases. An estimated value of the RDR is often reported in epidemiological studies that quantify and/or correct for measurement error.

The effects of exposure measurement error in the linear model where we have a linear exposure–disease relationship are well known. The effect of measurement error is to reduce the slope, or attenuate, the regression. Assume that we have a simple linear model relating our true exposure  $X$  to some outcome  $Y$ ,

$$Y = \alpha + \beta X + \epsilon. \quad (3.1)$$

If instead of fitting the model we are interested in, the model of equation 3.1, we decided to proceed using our observed exposure  $W$  in place of the true exposure  $X$  and fit

$$Y = \alpha^* + \beta^* W + \epsilon^* \quad (3.2)$$

then under the model of equation 3.1  $\beta = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$ , whilst under the model of equation 3.2  $\beta^* = \frac{\text{Cov}(W, Y)}{\text{Var}(W)}$ . If the measurement error is independent of the true exposure, and of the outcome, then  $\text{Cov}(W, Y) = \text{Cov}(X, Y)$  and the quantity of interest  $\beta$  is related to the observed value  $\beta^*$  via the relationship

$$\beta^* = \frac{\text{Cov}(W, Y)}{\text{Var}(W)} = \frac{\text{Cov}(X, Y)}{\text{Var}(W)} = \frac{\text{Var}(X)}{\text{Var}(W)} \beta = \lambda \beta. \quad (3.3)$$

Therefore the slope of the regression using the observed exposure  $\beta^*$ , is the slope from the regression using the true exposure  $\beta$ , multiplied by the RDR. Clearly this hints at a way of correcting for the effects of exposure measurement error in the linear model if we can estimate  $\lambda$ ; we shall consider this in detail in chapter 4.

The RDR was shown by Rosner, Spiegelman, and Willett [28] to approximately correspond to the degree of attenuation in the logistic model when the odds ratio is below 3. Hughes [136] showed that under the Cox proportional hazards model the RDR approximately corresponds to the true level of attenuation of a linear exposure–disease relationship under the rare disease assumption (i.e. if the number of observed events is relatively low) and if the risk gradient is relatively low.

---

Although the effect of random exposure measurement error on linear exposure–disease relationships is well understood, the effect on non-linear relationships is less well known. Measurement error is known in general to attenuate non-linear relationships towards a null exposure–disease relationship although this is not necessarily always the case, as we shall see later in this chapter. The fact that we can easily see the attenuation caused by exposure measurement error when the exposure–disease relationship is linear and can show simply that the RDR corresponds (approximately) to the degree of this attenuation makes it easy to understand. However, when the exposure–disease relationship is non-linear it is usually difficult to obtain concise expressions for the precise effect of measurement error on the exposure–disease relationship.

One exception to this is given by Kuha and Temple [137] who studied the effects of exposure measurement error on quadratic relationships. They show that if the true exposure–disease relationship is quadratic then, under classical measurement error, the observed exposure–disease relationship will also be quadratic. They also note that the true turning point of a quadratic relationship will lie between the observed turning point and the mean of the true exposure distribution.

As concise expressions for the effects of measurement error on non-linear relationships are in general difficult to obtain, we investigate the effect of classical measurement error on a range of exposure–disease relationships using simulation. In the next section we consider shapes of the exposure–disease relationship that are commonly observed in epidemiological studies to inform our choices for the simulation study of section 3.4.

## **3.2 Shape of the exposure–disease relationship**

There are many difficulties in accurately ascertaining the shape of exposure–disease relationships including model selection and lack of power to detect clinically important features of the relationship. For example, exposure distributions tend to be skewed with few people at the extreme ends of the exposure range which can make it harder to accurately estimate the relationship in the extremes, however it is often the case that these are the areas in which we are most interested and where the highest risk is to be found.

In this dissertation we mainly focus on non-linear exposure–disease relationships. However, many relationships are linear such as the relationship between cigarette smoking and the risk of lung cancer, and between radiation and cancer [25].

The most commonly observed shape of non-linear relationship is the J- or U-shaped relationship. A J-shape or U-shape relationship has been found in the relationship of alcohol consumption and all cause mortality [138–140]. A J-shaped relationship has also been observed between DBP and risk of cardiovascular and all-cause death, and death from cardiovascular disease amongst those with hypertension [141].

### 3. The effects of exposure measurement error

---

Relationships where the risk rises sharply at low exposure levels and levels off at high levels of exposure have been observed by Cates [142] who describes a convex exponential relationship between unprotected sex with an infected partner and the risk of contracting an STD. Pope *et al.* in their investigation of the relationship between cardiovascular mortality and exposure to airborne fine particulate matter found a similarly shaped relationship. Polesel *et al.* [143] found an S-shape relationship between ethanol intake and risk of cancer of the upper digestive tract.

Relationships that increase in steepness with increasing exposure have been observed in the relationships of: maternal age and risk of Down's syndrome, osteoporosis and the risk of fracture, and blood pressure and the risk of cardiovascular disease [25].

Rose's Theory of Preventive Medicine [25] gives threshold relationships between intraocular pressure and risk of glaucoma and anaemia and the display of symptoms of anaemia. A study of the relationship between systolic blood pressure (SBP) and risk of cardiovascular and all-cause death in the Framingham Heart Study by Port *et al.* [144] suggested a threshold of SBP below which the relationship with risk of death is flat and above an increased risk with increasing SBP. A true threshold relationship is usually epidemiologically implausible and the use of the term 'threshold' is more often used to mean that there is a level of exposure below/above which the risk gradient is minimal or flat. What is deemed as 'minimal' will depend on the context.

We shall therefore include in our simulation a linear relationship (for comparison); two threshold relationships — one that is flat below the threshold and linear above and a second where the relationship is flat below a threshold, is then quadratic up to a second threshold, after which the relationship increases linearly; increasing quadratic; J-shaped; U-shaped and asymptotic relationships as these are some of the most commonly observed shapes for the exposure–disease relationship.

The labels given to shapes of exposure–disease relationships are not consistent. For example J- and U-shaped are often used to describe the same shape of relationship, since it is usual for those termed U-shaped relationships to exhibit at least some asymmetry. Also J-shape and threshold are often used to describe the same relationship especially if the upturn at low levels of exposure is small and there is much uncertainty about it, and because a sharp threshold is usually biologically implausible.

The shape of relationship observed may also be partially dependent on the modelling technique used. As discussed in chapter 2, polynomial functions can display implausible turning points at the extremes of the exposure range. Ideally we would prefer to know the shape of the exposure–disease relationship *a priori*. Although we may be able to use biological plausibility to give us some intuition about features of the exposure–disease relationship, such as monotonicity, we usually have to resort to data driven model selection methods. It is much better to try and elicit the shape of the exposure–disease relationship from the data than impose a poorly fitting model.

Now that we have discussed the shapes of relationship that are observed in practice and have identified seven shapes for the exposure–disease relationship to include in our simulation, we look at how we generate data from a Cox proportional hazards model and how we can make our simulations comparable across the different shapes by controlling the degree of non-linearity.

### 3.3 Methods for simulation

#### 3.3.1 Generating survival data

In the simulation study that follows we shall simulate data from Cox proportional hazards models, which were described in detail in chapter 2. Here we show how we may generate values from a Cox model.

The survival function,  $S$ , for the Cox proportional hazard model with a linear predictor  $f(\mathbf{X}; \boldsymbol{\beta})$ , is given by

$$S(t|\mathbf{X}) = \exp\{-H(t|\mathbf{X})\} = \exp\{-H_0(t) \exp(f(\mathbf{X}; \boldsymbol{\beta}))\}.$$

So, the distribution function is

$$\begin{aligned} F(t|\mathbf{X}) &= 1 - S(t|\mathbf{X}) \\ &= 1 - \exp\{-H_0(t) \exp(f(\mathbf{X}; \boldsymbol{\beta}))\}. \end{aligned}$$

The distribution function of any random variable is distributed uniformly on the interval  $[0, 1]$ . Hence  $F(T|\mathbf{X}) \sim U(0, 1)$  and by symmetry we also have that  $1 - F(T|\mathbf{X}) \sim U(0, 1)$ , so [145],

$$U := \exp(-H_0(T) \exp(f(\mathbf{X}; \boldsymbol{\beta}))) \sim U(0, 1).$$

By inverting the equation above so that  $T$  is the subject of the formula

$$T = H_0^{-1}\{-\log(U) \exp(-f(\mathbf{X}; \boldsymbol{\beta}))\}.$$

The baseline hazard function needs to be specified with the exponential, Weibull and Gompertz distributions being possible choices. In our simulation study we shall be using the exponential baseline hazard function which implies that the hazard is constant over time.

By choosing an exponential baseline hazard,  $h_0(t) = \rho$ , then the integrated hazard  $H_0(t) = \rho t$  and  $H_0^{-1}(t) = \frac{t}{\rho}$ . Hence we obtain the following formula for the survival time  $T$ ,

$$T = -\frac{\log U}{\rho \exp(f(\mathbf{X}; \boldsymbol{\beta}))}.$$

By choosing suitable values for  $\rho$  and  $\boldsymbol{\beta}$  and by generating random uniform numbers  $U$ , we can generate observations from the desired Cox model.

### 3. The effects of exposure measurement error

---

#### 3.3.2 Determining the degree of non-linearity

In order that we are able to make fair comparisons between different shapes for the exposure–disease relationship investigated in the simulation study we shall keep the degree of non-linearity constant across shapes. Here we describe how we achieve this by minimising the squared difference between the shape of the relationship and the best fitting line to that relationship averaged over the distribution of the true exposure.

Suppose that we have an exposure  $X$  with density function  $\phi(X)$ , and that the model of interest is a Cox model of the form

$$\log h(t|X) = \log h_0(t) + \beta(f(X) - \overline{f(X)}), \quad (3.4)$$

where  $\overline{f(X)}$  is the mean of  $f(X)$  over the true exposure distribution

$$\overline{f(x)} = \int_{-\infty}^{\infty} f(x)\phi(x)dx.$$

We want to find the best fitting model with a linear relationship between exposure and outcome

$$\log h(t|X) = \log h_0^*(t) + \gamma(X - \mu_x). \quad (3.5)$$

For given  $\beta$  we want to find the value of  $\gamma$  that minimises  $S$ , the squared difference between the linear predictors in equations 3.4 and 3.5,

$$S = \int_{-\infty}^{\infty} (\beta(f(x) - \overline{f(x)}) - \gamma(x - \mu_x))^2 \phi(x)dx.$$

The partial differential of  $S$  with respect to  $\gamma$  is

$$\frac{\partial S}{\partial \gamma} = -2 \int_{-\infty}^{\infty} (x - \mu_x) \left( \beta(f(x) - \overline{f(x)}) - \gamma(x - \mu_x) \right) \phi(x)dx$$

By setting  $\frac{\partial S}{\partial \gamma} = 0$  and solving for  $\hat{\gamma}$ ,

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} (x - \mu_x) (\beta(f(x) - \overline{f(x)}) - \hat{\gamma}(x - \mu_x)) \phi(x)dx \\ \hat{\gamma} \int_{-\infty}^{\infty} (x - \mu_x)^2 \phi(x)dx &= \beta \int_{-\infty}^{\infty} \left( (f(x) - \overline{f(x)})(x - \mu_x) \right) \phi(x)dx \\ \hat{\gamma} &= \beta \frac{\int_{-\infty}^{\infty} (f(x) - \overline{f(x)})(x - \mu_x) \phi(x)dx}{\int_{-\infty}^{\infty} (x - \mu_x)^2 \phi(x)dx}. \end{aligned}$$

We can obtain values of  $\beta$  for a set of shapes so that they have the same non-linearity  $S$  by:

- 
1. Fixing the value of  $\beta = \beta_1$  for one shape
  2. Calculating  $\hat{\gamma} = \hat{\gamma}_1$  for this shape
  3. Calculating  $S = S_1$  using  $\beta_1$  and  $\hat{\gamma}_1$
  4. For each of the other shapes calculate  $\hat{\gamma}$ , and search for the value of  $\beta$  that gives the same value for  $S = S_1$ .

## 3.4 Simulation Study

We conducted a simulation study to explore the shape of the observed exposure–disease relationship when the exposure is subject to measurement error using a grouped exposure analysis, P-splines and fractional polynomials, which were described in chapter 2.

Although it is known that non-linear exposure–disease relationships are generally attenuated by exposure measurement error, what has not been previously shown is the precise effect of measurement error on the most commonly observed shapes of exposure–disease relationships.

### 3.4.1 Data generation

Data for the true exposure for individual  $i$ ,  $X_i$ , were generated from a normal distribution with mean 10 and variance 1. The exposure for each individual was assumed to be subject to classical measurement error such that the observed exposure  $W_i$ , was given by  $W_i = X_i + U_i$  where  $U_i$  was normally distributed with mean 0 and variance  $\sigma_u^2$ . The  $U_i$  were generated such that they were independent of each other, and of the true level of exposure,  $X_i$ . In order to investigate how the shape of the exposure–disease relationship changes with increasing measurement error under each of the modelling approaches we let the measurement error variance  $\sigma_u^2$  take the values (0, 0.25, 0.5, 1) corresponding to RDRs of 1, 4/5, 2/3, 1/2 respectively. Note that the case when  $\sigma_u^2 = 0$  corresponds to no exposure measurement error and is included to illustrate the performance of the modelling methods in the absence of measurement error and for comparison. An RDR of 0.5 corresponds to when the measurement error variance is as large as the variance of the true exposure. RDRs in this range are commonly seen in epidemiological investigations. Lewington *et al.* [146] found RDRs of 0.51 and 0.56 for systolic blood pressure in the Glostrup and Framingham studies, respectively; 0.52 and 0.54 for diastolic blood pressure; and 0.68 and 0.63 for total cholesterol. Whilst the Emerging Risk Factors Collaboration [11] found regression dilution ratios of log lipoprotein(a) and log triglycerides levels (adjusted for age and sex) of 0.87 and 0.72 respectively. All data generation and analyses for this simulation study were carried out in R.

We considered the seven models for the shape of the true exposure–disease relationship on the logarithmic scale; linear, linear threshold, U-shaped, J-shaped, increasing quadratic, asymptotic and non-linear threshold. Equations for the form of the log-hazard ratios for each of these

### 3. The effects of exposure measurement error

Shape	Form of $\log(h(t X))$
Linear	$\log(h_0(t)) + \beta_1 X$
Linear threshold *	$\log(h_0(t)) + \beta_2 I_{\{X > 10\}}(X - 10)$
U-shaped	$\log(h_0(t)) + \beta_3 (X - 10)^2$
J-shaped	$\log(h_0(t)) + \beta_4 (X - 9)^2$
Increasing quadratic	$\log(h_0(t)) + \beta_5 (X - 7)^2$
Asymptotic	$\log(h_0(t)) + \beta_6 / (4 - X)^2$
Non-linear threshold *	$\log(h_0(t)) + \beta_7 (I_{\{11.75 > X > 10.75\}}(X - 10.75)^2 + I_{\{X > 11.75\}}(2X - 22.5))$

\*  $I_{\{X > a\}}$  is an indicator function that equals 1 when the inequality  $X > a$  is satisfied and 0 otherwise.

Table 3.1: Models considered for the relationship between true exposure  $X$  and log-hazard in the simulation study.

models are given in table 3.1.

Under the linear threshold relationship the threshold was placed at the mean of the exposure distribution so that the threshold effect could be clearly identified. Often in practice the threshold value will lie in the tail of data and will be more difficult to identify. The nadir of the J-shaped relationship was chosen at 9 which is one standard deviation from the mean so that the upturn at the lowest levels of exposure could be picked up. The increasing quadratic relationship was chosen such that the nadir did not fall too far outside the range of the exposure distribution so that the relationship showed sufficient curvature.

$\beta_1 = 0.3$  was chosen for the linear model which gives an approximate doubling of the risk between the bottom and top fifths of true exposure.  $\beta_i, i = 2, \dots, 7$  were chosen so that the degree of non-linearity in each of the functions was the same where non-linearity was quantified as the squared difference, averaged over the distribution of  $X$ , between the true relationship and that which would be observed if a linear exposure–disease model as described in the previous section. (Note that we truncated the integral at  $x = 4$  for the asymptotic relationship to avoid the asymptote.) We set  $\beta_6 = 10$  and calculated  $\beta_i, i = 2, 3, 4, 5, 7$  to have the same degree of non-linearity for all other models.

For each observation, survival times were generated for each of the seven shapes for the exposure–disease relationship from a model with constant baseline hazard,  $h_0(t) = \rho$ , and observations were censored at time 10 if an event had not occurred. This corresponds to a ten year follow up which is a realistic follow up time for an epidemiological study. Admittedly the censoring scheme is likely to be vastly more complicated in practice however we do not believe that our results will be extremely sensitive to the choice of censoring pattern, however they are likely to be sensitive to the degree of censoring.  $\rho = 0.01$  was chosen for the linear model so that approximately 10% of individuals would suffer an event during the follow up period. This value was chosen because low event rates are typical in epidemiological studies. In order that our simulations were comparable between shapes for the exposure–disease relationship we

---

Shape	$\beta$	$\rho$
Linear	0.300	0.010
Threshold	0.282	0.009
U-shaped	0.060	0.010
J-shaped	0.060	0.009
Increasing quadratic	0.060	0.005
Asymptotic	10.000	0.056
Non-linear threshold	0.258	0.010

---

Table 3.2: Parameter values for the true exposure–disease relationships used in the simulation study.

chose  $\rho$ , via simulation of extremely large datasets, such that it gave approximately the same number of events as in the linear case since for fixed  $\rho$  we found that the number of events varied greatly between shapes. Parameter values for  $\beta$  and  $\rho$  used in the simulation study are given in table 3.2.

For each value of  $\sigma_u^2$  and each shape for the exposure–disease relationship we generated 1,000 datasets containing  $[W_i, t_i, e_i]$ , for  $i = 1, \dots, 15,000$  where  $t_i$  is the length of the observation period for individual  $i$  and  $e_i$  is an indicator that takes the value one if the individual suffered an event during the observation period and zero otherwise.

A sample size of 15,000 was chosen to give approximately 80% power to reject the null hypothesis of linearity of the exposure–disease relationship at the 5% level when there is no measurement error under the increasing quadratic relationship using a fractional polynomial analysis. This was determined by performing a simulation substudy prior to the main study where we used the same data generating mechanism as in the main study.

### 3.4.2 Statistical methods evaluated

We considered three methods of modelling the exposure–disease relationship in our simulation study: grouped exposure analysis, fractional polynomials and P-splines. The best fitting model with four degrees of freedom was chosen in each simulated data set in order to be able to compare the three modelling methods. This was achieved by splitting the observed exposure values into five groups under the grouped exposure analysis, selecting the best fitting FP2 model for fractional polynomials, and by choosing the smoothing parameter in the P-spline model to give approximately 4 effective degrees of freedom.

- Grouped exposure analysis — For each simulation we grouped the simulated values according to quintiles of the observed exposure values to obtain five equally sized groups. We then fit the model

$$\log h(t|X) = \log h_0(t) + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \beta_3 X^{(4)} + \beta_5 X^{(5)} \quad (3.6)$$

### 3. The effects of exposure measurement error

---

where  $X^{(k)}$  is an indicator of group membership as defined in section 2.2.1. The third group has been chosen as the reference. This choice of reference makes visual comparisons easier with the other two methods since the mean of the observations within the third group will approximately equal the overall mean which is a common choice as the reference for fractional polynomial and P-spline analyses.

- Fractional polynomials — for each simulation we chose the FP2 model with lowest deviance amongst all possible FP2 models

$$\log h(t|X) = \log h_0(t) + \begin{cases} \beta_1 X^{(p_1)} + \beta_2 X^{(p_2)} & \text{if } p_1 \neq p_2 \\ \beta_1 X^{(p)} + \beta_2 X^{(p)} \log X & \text{if } p_1 = p_2 = p \end{cases} \quad (3.7)$$

where the powers were selected from the restricted set  $\wp$  discussed in section 2.2.3.

- P-Splines — for each simulation we fit a P-spline model on 10 equally sized intervals (using 13 B-spline basis functions and 17 knots) and chose the smoothing parameter such that the effective degrees of freedom of the fitted model equals approximately 4,

$$\log h(t|X) = \log h_0(t) + \sum_{j=1}^{12} \beta_j b_{j+1}(X). \quad (3.8)$$

We also considered the power to detect non-linearity of the three methods at the 5% level of significance using sample sizes of 15,000, 10,000, 5,000 and 1,000. This was achieved by performing the following tests:

- For groups we take twice the difference in the log-likelihood between the model

$$\log h(t|X_G) = \log h_0(t) + \beta_1 X_G$$

where  $X_G$  is formed by assigning each individual the mean exposure value of the quintile in which the exposure lies, and the model of equation 3.6, testing against the Chi-square distribution with three degrees of freedom.

- For fractional polynomials we test twice the difference between the log-likelihoods of the linear Cox model

$$\log h(t|X) = \log h_0(t) + \beta_1 X \quad (3.9)$$

and the best fitting FP2 model against the Chi-square distribution with three degrees of freedom.

- Similarly for P-splines we compare twice the difference in the (unpenalised) log-likelihood between the linear model (equation 3.9) and the P-spline model with 4 effective degrees of freedom (equation 3.8) and test against the Chi-square distribution with three degrees

---

of freedom.

Note that this is different to the standard test for non-linearity that is produced by the `pspline` function in R. The symmetry of the spline basis functions means that an approximate test for linearity can be achieved by performing a generalised least squares regression of the spline coefficients from our fitted model on the centre of the spline basis functions using the variance-covariance matrix as the known variance matrix for the coefficients [79]. By finding the difference in the two Wald statistics of these models we can perform an approximate test of non-linearity against a Chi-square distribution with 3 degrees of freedom. This approach does not require fitting of an additional Cox model which is computationally expensive compared with the generalised least squares fit. With modern computing power an exact test against the Cox model with linear predictor seems more appropriate in practice.

### 3.4.3 Results

We display the results of the simulation graphically. Overall, the P-spline and fractional polynomial models gave very similar results for the shape of the exposure–disease relationships except for the threshold relationships where there was a significant difference. Hence, we include the results for fractional polynomials for all shapes (figure 3.2) and only the threshold relationships for the P-spline model (figure 3.3). Under each simulation we may pick a different fractional polynomial model hence for each simulation we produced fitted values of  $\log \hat{h}(t|X) - \log \hat{h}(t|10)$  over a fine grid of points and averaged them to produce the curves of figure 3.2. For the P-spline and grouped exposure analyses we saved the parameter values for each model and averaged them to obtain the models shown in the graphs of figures 3.1 and 3.3. The approach used for the fractional polynomial analysis would have provided the same results but requires more storage.

#### **Performance of methods when there is no exposure measurement error**

As shown in section 2.2.6 a grouped exposure analysis does not provide unbiased estimates for the mean exposure within each group unless the true exposure–disease relationship is linear. We can see, however, in figure 3.1 that the effect of this on the observed exposure–disease relationship is small when there is no measurement error, although we can see it more clearly when the relationship shows greater curvature such as under the J- and U-shaped relationships. The grouped exposure analysis does not perform too badly at picking out the threshold relationship although the central group is a composite of individuals above and below the threshold and therefore makes the threshold appear less sharp. This is an example of groups being unable to pick out local features of the data. The non-linear threshold also picks up the threshold, but again the threshold appears less sharp with the measurement error is larger. Under the true J-shaped relationship the hazard ratio of the first group only just lies above that for the sec-

### 3. The effects of exposure measurement error

---

ond and the nadir is essentially lost; the relationship could easily be mislabeled as showing a threshold relationship. Viewers of this graph may conclude incorrectly that the lowest hazard occurs at an exposure value of around 9.5 when it actually is located at 9.

Fractional polynomials and P-splines do very well in recreating the true exposure–disease relationship when there is no measurement error except for the threshold relationships. Fractional polynomials suffer a lack of ability to pick up the threshold shapes which is to be expected since sudden changes in gradient are not features of polynomial models (figure 3.2). Instead, the observed exposure–disease relationship is J-shape in appearance. P-splines, which have the ability to fit local features of the data, perform much better in terms of picking up the threshold (figure 3.3). Since P-splines are composed of piecewise polynomial functions it still has some difficulties around the threshold value as we would expect. We investigated using P-splines with more degrees of freedom but this did not seem to significantly improve the fit (not shown).

Table 3.3 shows the estimated power to reject the null hypothesis of a linear exposure–disease relationship under each of the three analysis methods, seven shapes for the exposure–disease relationship, and different degrees of measurement error for sample sizes of 15,000, 10,000, 5,000 and 1,000. We can see that both fractional polynomials and P-splines have much greater ability to detect non-linearity of the exposure–disease relationship than under the grouped exposure method except for under the threshold model where fractional polynomials perform slightly worse than groups. As the sample size decreases, P-splines continue to outperform the other methods. We can see from table 3.3 that the type I error under the linear relationships is much less than the nominal 5% level for the fractional polynomial analyses. This is because, as noted in section 2.2.3, the powers are selected from the limited set  $\emptyset$ , meaning that the true number of degrees of freedom in testing the null hypothesis of a linear relationship is less than the 3 that are used in the test [96].

#### **Effect of exposure measurement error on the shape of the observed exposure–disease relationship**

Measurement error has the effect of increasing the range of the observed exposure, since the observed exposure will have a larger variance than the true exposure, and it reduces the range of the hazard across the range by ‘mixing’ up those with low and high true hazard. We can see this in figure 3.1 as the group means become more dispersed with increasing measurement error, and the range of the hazard reduced. Under the grouped exposure analysis we saw that the threshold shaped relationships were relatively sharp when there was no measurement error. With increasing measurement error we rapidly lose the true threshold shape with the relationship appearing increasingly J-shaped. This is what we might expect since the effect of measurement error will be to mix individuals from above and below the threshold. Under the true J-shaped relationship, using the grouped exposure analysis, the nadir and upturn in risk for

---

those with the lowest exposure levels are lost after the addition of even the smallest amount of measurement error considered, which is much less than is often observed in practice. Both the increasing quadratic and asymptotic relationships appear increasingly linear as the degree of measurement error increases.

Under the fractional polynomial analysis (figure 3.2) we see that both of the threshold relationships appear J-shaped for all levels of measurement error except for the most extreme where we observe an increasing quadratic relationship. The observed shape for the exposure–disease relationship stems from the inability of fractional polynomials to fit the threshold function well in the absence of exposure measurement error. P-splines (figure 3.3) perform much better retaining some of the feature of the threshold until the most extreme level of measurement error when the relationship appears to be more of an increasing quadratic. The increasing quadratic and asymptotic shapes for the exposure–disease relationship appear increasingly linear as the degree of measurement error increases as was the case with the grouped exposure analysis. In figure 3.2 we see that under the J-shaped relationship we do not lose the nadir and upturn in risk as we did under the group based analysis. The nadir can be seen to move away from the mean value. It is of particular interest to note that both under the J-shaped and threshold relationships there are ranges of the exposure outcome relationship where the relationship is biased *away* from and not towards the null under each of the analysis methods.

As we have seen in figures 3.2 and 3.3 the effect of random measurement error is to make the exposure–disease relationship appear more linear. This results in an associated loss of power to detect non-linearity. P-splines continue to have a greater power to detect non-linearity, over fractional polynomials and grouped exposure analyses when the exposure is subject to measurement error. When the sample size is small we have low power to detect non-linearity even if the true exposure had been observed, and in this case the effect of measurement error on our power to detect non-linearity is small.

From the analytical results of Kuha and Temple [137] we would expect the nadir for the J-shaped relationship to occur at 9.00, 8.75, 8.50, 8.00 for RDRs of 1, 4/5, 2/3, 1/2 respectively, and to remain at 10 under the U-shaped relationship. Under the J-shaped relationship not only does the location of the nadir move, the hazard at the nadir is closer to the null. Often there is much interest in accurately estimating the nadir, such as in the J-shaped relationship between alcohol consumption and all cause mortality [32, 147] where the consumption of small quantities of alcohol appears to reduce the risk of mortality. The nadir will tell us the ‘optimal’ level of alcohol consumption, and the degree of protection against disease it offers which could be used to shape public health policy. The effects of exposure measurement error create considerable problems in estimating the location of the nadir, and the degree of protection offered at this minimum. In the case of the relationship between alcohol consumption and all-cause mortality the assumption of classical measurement error is violated and so the effect of exposure measurement error on this relationship may differ from that observed in our simulation study.

### 3. The effects of exposure measurement error

---

Measurement error appeared to preserve the shape of the exposure–disease relationship albeit in a much diluted form, except for in the case of the threshold relationships where the exposure measurement error distorted the threshold. However, in individual simulations the observed shape of the exposure–disease relationship may have been different from the truth. It is not always easy to distinguish the true shape of the exposure–disease relationship from the observed exposure–disease relationship. Some correction techniques, however, as we shall see in the next chapter, rely on being able to choose the correct form for the exposure–disease relationship from the observed relationship. It may also be difficult to detect non-linearity even with a relatively large sample size of 15,000 individuals and 1,500 events. Some intuition can be gained as to what the effect of classical measurement error may be on the shape of an exposure–disease relationship by imagining the relationship as a piece of string that is simultaneously being pulled smoothly away from the mean in the horizontal direction and towards the  $x$ -axis in the vertical direction.

Here we have concerned ourselves purely with a single exposure subject to classical measurement error. The results obtained in the presence of confounders, or under different measurement error structures may be quite different especially if there is correlation of errors between variables. In most cases, however, the effect of exposure measurement error will be to attenuate the exposure–outcome relationship and the results above may still provide some indication of the effects on the observed shape.

## 3.5 Conclusion

Exposure measurement error can lead to a triple whammy of effects: parameter bias, loss of power to detect interesting relationships between variables, and loss of features of the data. It is well known that exposure measurement error attenuates the exposure–disease relationship when it is linear but the precise effect is less well known for non-linear relationships. We conducted a simulation study to look more closely at the effects of exposure measurement error on commonly observed exposure–disease relationship shapes (linear, threshold, J-shaped, U-shaped, increasing quadratic, asymptotic and non-linear threshold) under the Cox model. Fractional polynomials were particularly poor at picking up the threshold relationships even when there was no measurement error, and all methods performed badly in the presence of measurement error. In the case of the J-shaped relationship we saw the nadir moving away from the mean and being lost from the grouped exposure analysis. Under the J-shaped and threshold relationships we also saw ranges in which the shape of the relationship was biased *away* not towards the null. Under the increasing quadratic and asymptotic relationships measurement error made the relationship appear increasingly linear. Fractional polynomials and P-splines generally increased our ability to detect non-linearity over the grouped exposure method. Measurement error, however, decreases our ability to detect non-linear exposure–disease relationships. Now that we have considered the effects of exposure measurement error on non-linear

Method	RDR	Shape of exposure–disease relationship							
		Linear	Threshold	U-shaped	J-shaped	Increasing quadratic	Asymptotic	Non-linear threshold	
Grouped Exposure	15000	1	0.04	0.63	0.47	0.45	0.52	0.23	0.28
		4/5	0.05	0.40	0.32	0.31	0.34	0.16	0.22
		2/3	0.04	0.26	0.23	0.24	0.26	0.12	0.18
		1/2	0.04	0.14	0.14	0.15	0.14	0.10	0.11
	10000	1	0.05	0.46	0.32	0.30	0.34	0.17	0.19
		4/5	0.05	0.26	0.21	0.21	0.23	0.13	0.15
		2/3	0.05	0.18	0.16	0.16	0.15	0.10	0.14
		1/2	0.05	0.11	0.11	0.10	0.12	0.08	0.10
	5000	1	0.04	0.20	0.14	0.14	0.16	0.09	0.10
		4/5	0.04	0.12	0.09	0.10	0.10	0.08	0.08
		2/3	0.04	0.09	0.08	0.09	0.08	0.06	0.07
		1/2	0.04	0.07	0.07	0.07	0.06	0.06	0.06
	1000	1	0.06	0.08	0.07	0.07	0.07	0.07	0.06
		4/5	0.06	0.07	0.06	0.07	0.06	0.06	0.05
		2/3	0.06	0.06	0.05	0.06	0.07	0.06	0.05
		1/2	0.06	0.06	0.05	0.05	0.06	0.06	0.06
Fractional polynomial	15000	1	0.01	0.64	0.76	0.77	0.76	0.41	0.70
		4/5	0.01	0.42	0.54	0.52	0.51	0.24	0.43
		2/3	0.01	0.26	0.34	0.34	0.29	0.15	0.25
		1/2	0.01	0.10	0.14	0.16	0.12	0.10	0.10
	10000	1	0.01	0.41	0.54	0.53	0.52	0.26	0.46
		4/5	0.01	0.25	0.33	0.33	0.30	0.15	0.27
		2/3	0.01	0.17	0.24	0.21	0.20	0.10	0.18
		1/2	0.02	0.09	0.11	0.12	0.11	0.06	0.09
	5000	1	0.01	0.18	0.22	0.22	0.21	0.09	0.18
		4/5	0.02	0.11	0.14	0.13	0.14	0.06	0.11
		2/3	0.02	0.09	0.09	0.09	0.10	0.05	0.08
		1/2	0.02	0.05	0.06	0.05	0.05	0.02	0.05
	1000	1	0.02	0.03	0.04	0.03	0.03	0.02	0.02
		4/5	0.01	0.02	0.03	0.02	0.02	0.03	0.02
		2/3	0.01	0.02	0.03	0.03	0.02	0.02	0.02
		1/2	0.02	0.02	0.02	0.02	0.02	0.02	0.02
P-spline	15000	1	0.08	0.84	0.86	0.87	0.87	0.62	0.91
		4/5	0.08	0.64	0.69	0.68	0.68	0.43	0.76
		2/3	0.08	0.47	0.53	0.53	0.48	0.33	0.52
		1/2	0.08	0.27	0.30	0.33	0.26	0.22	0.28
	10000	1	0.07	0.65	0.70	0.69	0.68	0.44	0.77
		4/5	0.08	0.44	0.50	0.48	0.48	0.31	0.53
		2/3	0.07	0.33	0.38	0.38	0.36	0.22	0.39
		1/2	0.07	0.21	0.24	0.25	0.23	0.15	0.24
	5000	1	0.07	0.38	0.40	0.38	0.38	0.23	0.42
		4/5	0.08	0.24	0.27	0.28	0.27	0.17	0.28
		2/3	0.08	0.20	0.21	0.21	0.21	0.14	0.20
		1/2	0.08	0.14	0.15	0.15	0.15	0.11	0.16
	1000	1	0.10	0.14	0.15	0.15	0.14	0.12	0.13
		4/5	0.09	0.11	0.12	0.12	0.10	0.10	0.11
		2/3	0.09	0.11	0.12	0.11	0.09	0.11	0.10
		1/2	0.09	0.10	0.09	0.10	0.08	0.10	0.09

Table 3.3: Estimated power to detect non-linearity under each non-linear exposure–disease relationship shape and estimated type I error in a test of the null hypothesis of linearity when the true association is linear, using grouped exposure, fractional polynomial, and P-spline analyses.

### 3. The effects of exposure measurement error

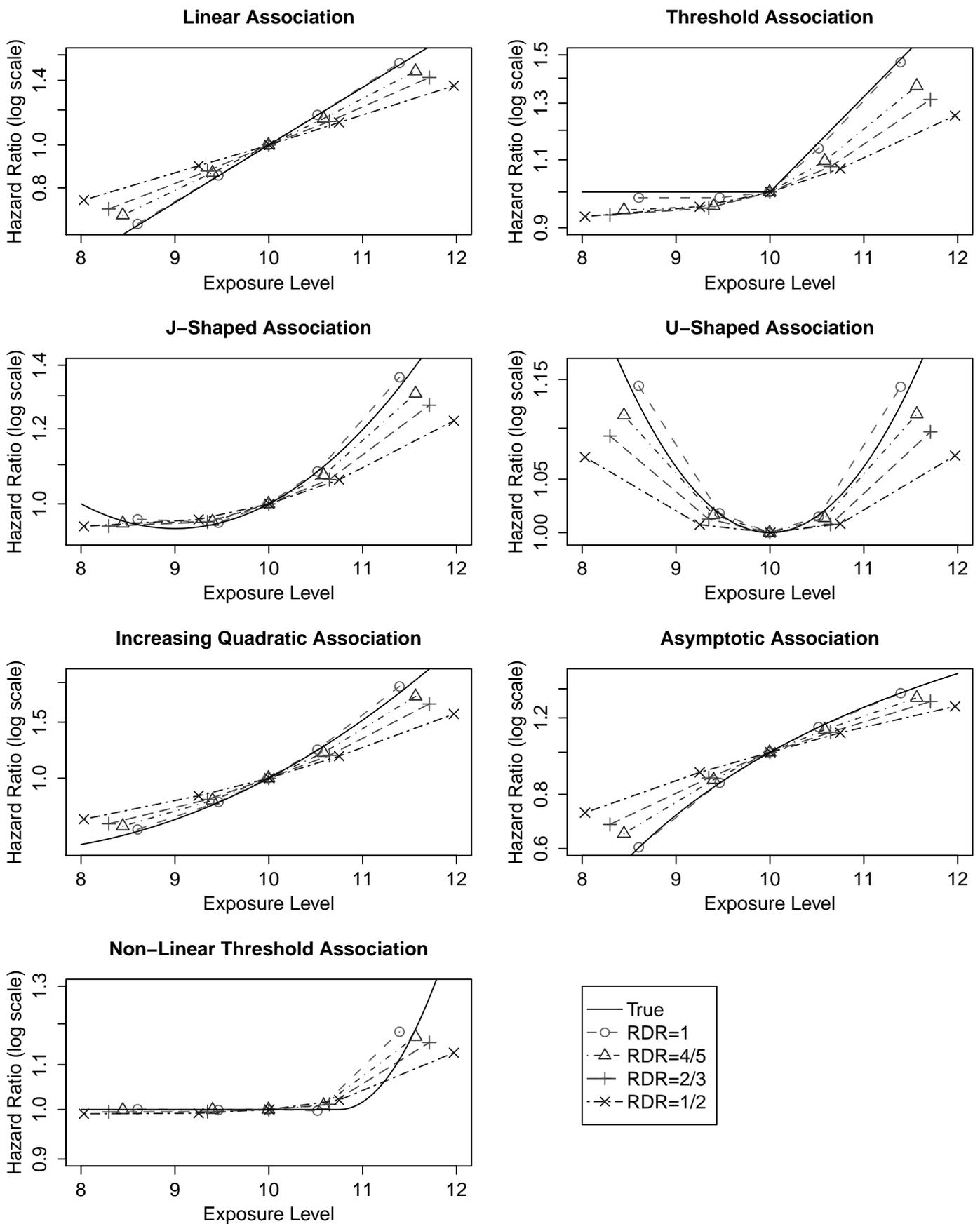


Figure 3.1: Grouped exposure analysis showing the effect of random measurement error on the observed exposure–disease relationship.

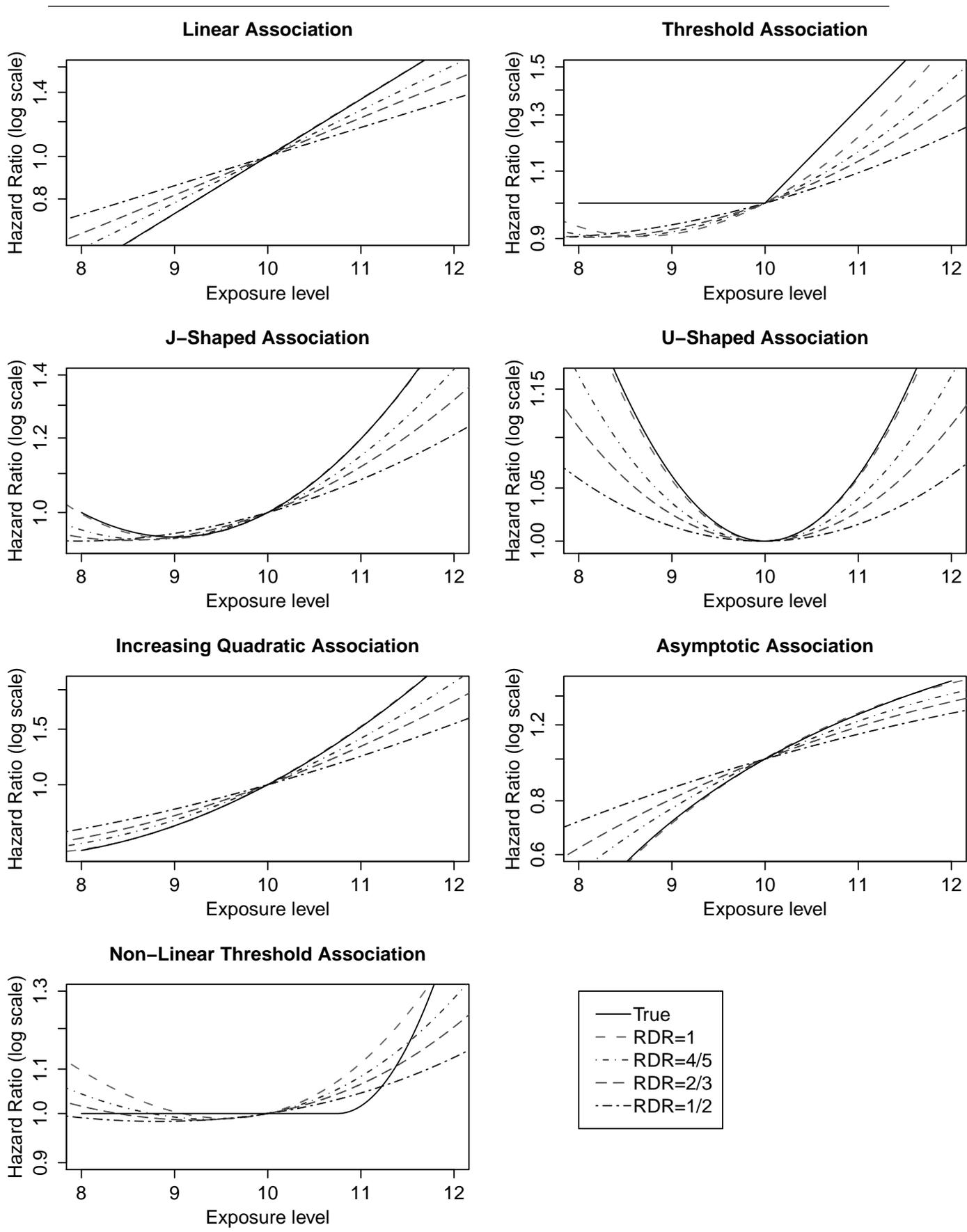


Figure 3.2: Fractional polynomial analysis showing the effect of random measurement error on the observed exposure–disease relationship.

### 3. The effects of exposure measurement error

exposure–disease relationships, in chapter 4 we look at a range of possible correction methods.

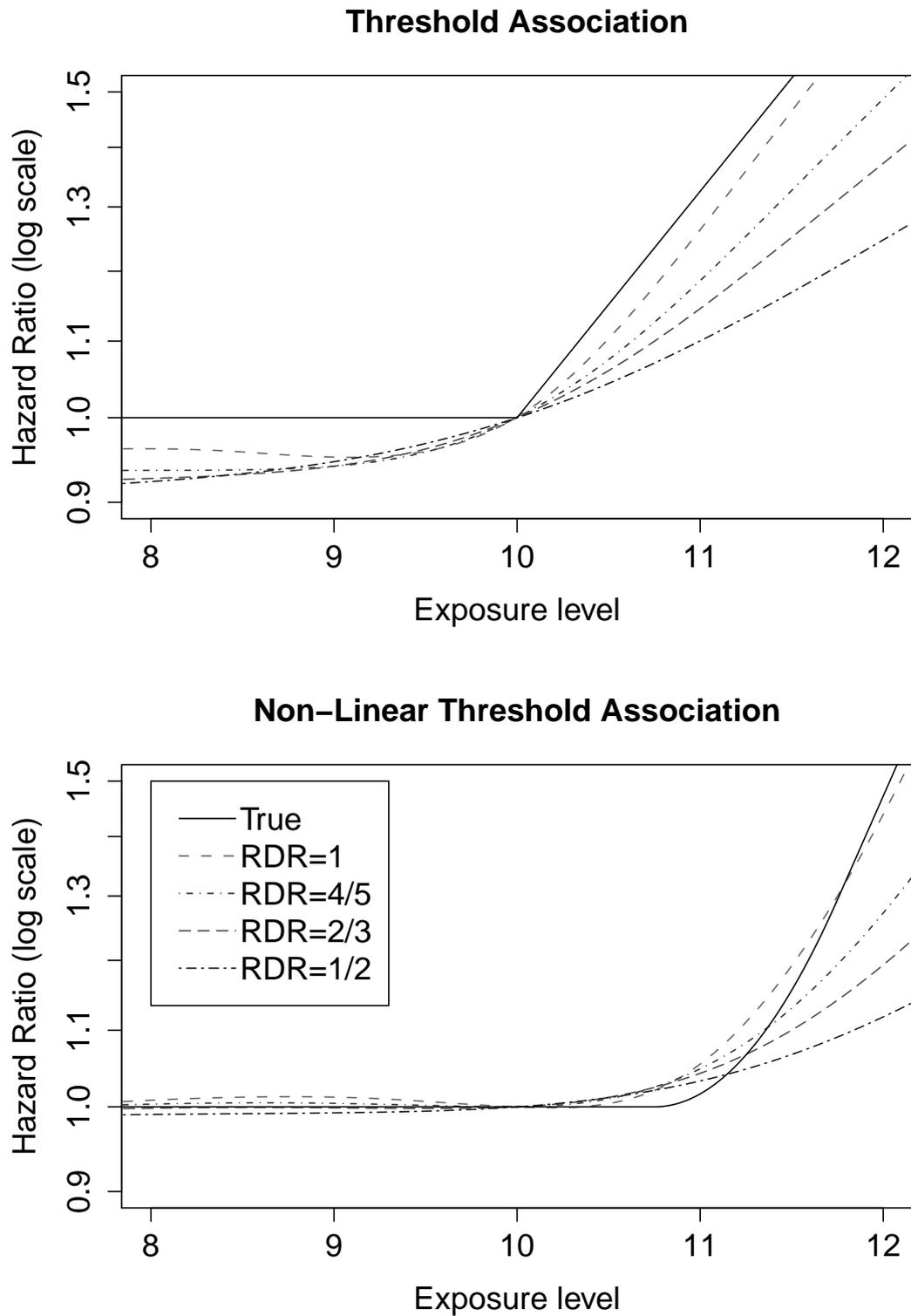


Figure 3.3: P-spline analyses showing the effect of random measurement error on the observed threshold and non-linear threshold shaped exposure–disease relationships.

# Chapter 4

## Methods of correcting for exposure measurement error

In this chapter we consider methods that have been proposed for correcting for the effects of exposure measurement error. We also develop two new correction methods: structural fractional polynomials and group-SIMEX. Much of the literature on measurement error correction focuses on linear models, and linear exposure–disease relationships. However, in this dissertation we are interested in methods for correcting for measurement error in Cox models when the exposure–disease relationship is non-linear.

In chapter 3 we saw that the effect of measurement error is to bias non-linear exposure–disease relationships, usually towards a null relationship between exposure and disease. If we wish to predict the future risk of an individual experiencing an event based on a single mismeasured exposure measurement then a prediction obtained from the model using the observed exposure will be unbiased, although measurement error will introduce greater uncertainty. If, however, our aim is to characterise the true relationship between exposure and disease then it is essential that we correct for the bias that measurement error introduces.

This chapter does not aim to be exhaustive given the large number of methods that have been proposed, but to outline some of the methods available, especially those most widely used in practice. For further reading we point the reader to the books highlighted in chapter 1, particularly Carroll *et al.* [33] who give a comprehensive description of methods for correcting for exposure measurement error, with chapters devoted to regression calibration, SIMEX, score function methods as well as a chapter dedicated to the Cox model. Schneeweiß and Augustin [148] gave a review of recent advances in the measurement error literature in 2006; Augustin and Schwarz [149] give a review of correction methods for Cox models; and Guolo [150] provides a review of robust measurement error correction methods.

### 4.1 An extra piece of information

Although Fuller [50] §1.6.1 gives an example where the effects of measurement error can be corrected for using the observed exposure alone, we ordinarily require some extra data, additional to the observed exposure, that allows us to identify the parameters of the measurement error model. There are three sources of this additional data: (alloyed) gold standard exposure measurements, replicate exposure measurements, and instrumental variables.

(Alloyed) Gold standard measures were discussed in section 3.1.1 and are where we observe the true level of exposure, or an exposure with significantly less measurement error than the observed exposure, in addition to the observed exposure for a subset of individuals.

Replicate measures are additional exposure measurements that are obtained for all, or a subset, of individuals. Replicate measures may be taken at baseline, or on one or more occasions post baseline. If the repeat measurements are true replicates, then the expected value of the repeat exposure, conditional on the observed exposure, is an unbiased measure of true exposure. Replicate measurements cannot inform us as to whether our mismeasured exposure measurements are subject to systematic measurement error. The timing in obtaining replicate exposure measurements may be important; for example sufficient time may need to elapse between baseline and repeat exposure measurements to ensure minimal correlation of measurement errors. If the time between baseline and repeat exposure measurements is large, however, then the level of true exposure may have changed. The measurement error in replicate biological measurements may often be assumed to be uncorrelated. Care does have to be taken, because in many practical situations this assumption may not hold. For example, it has been found that there is a high correlation in the errors of replicate FFQs [151].

Sometimes we can measure an instrumental variable for all or a subset of individuals—a variable which is correlated with the true exposure, uncorrelated with the measurement error, and uncorrelated with the outcome given the true exposure and confounders. A second mismeasured exposure measurement obtained from an independent measuring technique could be used as an instrumental variable. For further discussion of the role of instrumental variables in correcting for the effects of exposure measurement error see Carroll *et al.* [33].

All three sources of additional data can come from data internal to the study or from an external source. Clearly data collected from within the same study are preferable. If data are obtained from an external source then we have to consider the transportability of any parameter estimates that we estimate from the external data but use in modelling the exposure–disease relationship of interest. External sources may include previous validation studies of exposure–disease relationships, or reliability studies. It may be appropriate to make the assumption of transportability if the current and previous study are similar e.g. they have used the same assay method. It should be acknowledged that the transported data are subject to extra uncertainty, and sensitivity analyses can be used to assess the effect of deviations from the transported

---

values. The motivating dataset for this thesis, the ERFC, has replicate measurements (often multiple) available for the exposures of interest within subsets of participants. Therefore, we shall focus principally on methods that are applicable when replicate measurements are available.

## 4.2 Classification of methods

Before we consider methods for correcting for the effects of measurement error, we discuss two important concepts: the difference between so called differential and non-differential error, and the classification of correction methods into structural and functional methods.

Non-differential error is when the distribution of the outcome given  $(X, W, Z)$  depends only on  $(X, Z)$  i.e. the observed exposure  $W$ , tells us nothing more about the outcome, than the true exposure and any confounders. In most situations we can assume that the effect of measurement error is non-differential. An important example of differential measurement error is recall bias in a case-control study. Another case where differential measurement error is observed is when a continuous exposure subject to measurement error is split into groups, such as in the grouped exposure analyses described in section 2.2.1. The grouping introduces differential error except when the hazard is constant within groups [152]. This is because those individuals closer to the boundary between groups are more likely to be misclassified than those in the centre of the group.

Distinguishing between structural and functional methods is merely a form of classifying the different methods. Structural correction methods place, implicitly or explicitly, distributional assumptions on the true exposure. Functional correction methods make no assumption about the distribution of the true exposure—the true exposure could be deterministic or random and hence these methods are more general.

## 4.3 Correction methods for continuous exposures

### 4.3.1 Regression Calibration

Regression calibration is one of the most widely used tools for correcting for the effects of classical measurement error. The idea is simple: take the expectation of the true exposure, given the observed. Firstly we consider regression calibration in the linear model, and then we consider how the idea can be applied to the Cox model when certain assumptions are valid. We go on to consider how this principle can be extended to non-linear exposure–disease relationships and apply it to P-splines and fractional polynomials.

## 4. Methods of correcting for exposure measurement error

---

### Regression calibration in the linear model

In its most general form where the true linear model is  $E(Y|X) = f(X; \beta)$ , regression calibration can be expressed as

$$E(Y|W) = E(f(X; \beta)|W).$$

For example, in the case where the true exposure–disease relationship is linear,

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{4.1}$$

but we observe mismeasured exposure values  $W$ , we can apply regression calibration by taking the expected value of the outcome  $Y$  in equation 4.1, conditional on our observed exposure

$$\begin{aligned} E(Y|W) &= E(E(Y|W, X)|W) \\ &= E(E(Y|X)|W) \\ &= E(\beta_0 + \beta_1 X|W) \\ &= \beta_0 + \beta_1 E(X|W) \end{aligned} \tag{4.2}$$

where the second line follows from the assumption of non-differential measurement error. Because of its simplicity, regression calibration has become a popular method for correcting for the effects of measurement error in models where the exposure–disease relationship is linear [43, 44]. Linear regression calibration makes no assumption about the distribution of the true exposure and is therefore a functional approach.

Regression calibration provides us with a two-stage approach for correcting for the effects of measurement error: firstly we estimate  $E(X|W)$  and then we plug our estimate  $\hat{E}(X|W)$  in place of  $X$  in the regression of  $Y$  on  $X$ .

We need a method for estimating  $E(X|W)$ . Usually a linear relationship between  $X$  and  $W$  is assumed, and this assumption gives unbiased estimates for linear regression even if it does not hold [33]. If we have gold standard exposure measurements then we can estimate  $E(X|W)$  by regressing  $X$  on  $W$  for the subset of individuals for which we have gold standard measurements, i.e. we fit the model

$$X = \alpha_0 + \alpha_1 W + \epsilon \tag{4.3}$$

for those individuals for which we have observed  $X$  and  $W$ . Alternatively, in the case that for each individual (within a subset) we have a baseline exposure measurement  $W_1$  and a replicate exposure measurement  $W_2$ , we can fit the model

$$W_2 = \alpha_0 + \alpha_1 W_1 + \epsilon^*. \tag{4.4}$$

This will give the same parameter estimates because  $\text{Cov}(W_2, W_1) = \text{Var}(X)$ , since the measurement error in the replicate measurement is independent of the measurement error in the baseline measurement.

We can check whether a linear relationship between the true and observed exposure is appropriate using regression diagnostics. A linear relationship gives unbiased estimates in linear regression even if this assumption does not hold.

Once we have estimated  $E(X|W)$  using either of these methods we replace  $X$  with this estimate in a second-stage regression

$$Y = \beta_0 + \beta_1 \hat{E}(X|W) + \epsilon^*.$$

A more efficient approach, when we have observed the true exposure, is to use efficient regression calibration which takes a weighted average of the parameters from the second-stage regression above, and the parameters from the outcome model fitted to only those individuals for which the true exposure was observed [153].

In chapter 3 we introduced the regression dilution ratio which we defined as  $\lambda := \frac{\text{Var}(X)}{\text{Var}(W)}$ .  $\alpha_1$  in the model of equations 4.3 or 4.4 can be seen to be the regression dilution ratio. In equation 4.3

$$\alpha_1 = \frac{\text{Cov}(X, W)}{\text{Var}(W)} = \frac{\text{Var}(X)}{\text{Var}(W)} = \lambda$$

and in equation 4.4

$$\alpha_1 = \frac{\text{Cov}(W_1, W_2)}{\text{Var}(W_2)} = \frac{\text{Var}(X)}{\text{Var}(W_2)} = \lambda.$$

The parameter of interest  $\beta_1$  in equation 4.1 is given by  $\frac{\text{Cov}(Y, X)}{\text{Var}(X)}$ . Note that this is equal to

$$\frac{\text{Cov}(Y, W)}{\text{Var}(W)} \frac{\text{Var}(W)}{\text{Var}(X)} = \frac{\beta_1^*}{\lambda}$$

where  $\beta_1^*$  is the slope parameter of the naïve regression of  $Y$  on  $W$ . So an alternative approach to the application of regression calibration is to firstly regress  $X$  on  $W$  (as in equation 4.3), or  $W_2$  on  $W_1$  (as in equation 4.4), to obtain an estimate of  $\lambda$ , then regress  $Y$  on  $W$  to obtain an estimate of  $\beta_1^*$  and finally calculate the corrected estimate  $\hat{\beta}_1 = \frac{1}{\lambda} \hat{\beta}_1^*$ .

Regression calibration is just one method through which an estimate of  $\lambda$  can be obtained in the case of a single exposure measured with error in the absence of confounders, and Thompson *et al.* [154] compare alternative methods. Although Thompson *et al.* show that other methods perform better in terms of variance, regression calibration extends easily to both the case where we have multiple exposures measured with error, and when we have confounders.

When we have a set of confounders  $Z$  in our disease model, we need to additionally include

#### 4. Methods of correcting for exposure measurement error

---

them in the regression calibration models of equations 4.3 and 4.4. In this case the regression dilution ratio becomes [33],

$$\lambda_{X|Z} = \frac{\text{Var}(X|Z)}{\text{Var}(W|Z)} = \frac{\text{Var}(X|Z)}{\text{Var}(X|Z) + \text{Var}(U)}.$$

Note that  $\lambda_{X|Z}$  will be smaller than the RDR unadjusted for confounders,  $\lambda$ , since  $\text{Var}(X|Z) \leq \text{Var}(X)$ , with equality if, and only if,  $X$  is independent of  $Z$  i.e. the effect of measurement error is greater in the presence of confounders that are correlated with the true exposure of interest.

In many situations it is not just the exposure of interest that is subject to measurement error, but also confounders. The effect of measurement error in confounders is similarly to bias the confounder-disease relationship, usually towards no relationship between confounder and disease. Hence, when we adjust for the effect of confounders that are measured with error the adjustment will not be complete and we may be left with residual confounding. If a confounder is measured with error, and the measurement error component of the confounder is correlated with that of the exposure of interest, then the effect of measurement error can be to bias the observed relationship in either direction [155].

When multiple exposures are measured with error we can use a multivariate version of regression calibration to obtain measurement error corrected parameters

$$\beta = \Lambda^{-1}\beta^*$$

where  $\beta$  is the vector of coefficients obtained from the regression using the observed exposure and

$$\Lambda = \Sigma_W^{-1}\Sigma_X$$

with  $\Sigma_X$  and  $\Sigma_W$  the variance-covariance matrices of the true and observed exposures respectively. As above, if we additionally have a set of perfectly measured confounders  $Z$  then the RDR,  $\Lambda$ , should be calculated using conditional variances. In chapters 6 and 7 we shall look at methods of assessing the measurement error structure including whether the measurement error is correlated with confounders.

#### Regression calibration in the Cox model

Thus far we have considered regression calibration in the linear model. In a Cox regression where the exposure–disease relationship is linear the log hazard is

$$\log h(t|X) = \log h_0(t) + \beta X.$$

Under the assumption of non-differentiability [42],

$$\begin{aligned}
h(t|W) &= \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} P(\{T < t + \epsilon\} | \{T \geq t\}, W) \\
&= \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} E(P(\{T < t + \epsilon\} | X, \{T \geq t\}, W) | \{T \geq t\}, W) \\
&= \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} E(P(\{T < t + \epsilon\} | X, \{T \geq t\}) | \{T \geq t\}, W) \\
&= E(h(t|X) | \{T \geq t\}, W)
\end{aligned} \tag{4.5}$$

hence,

$$h(t|W) = h_0(t) E(\exp(\beta^T X) | W, \{T \geq t\})$$

which depends on the past history of the process through  $\{T \geq t\}$ . However, the time dependence of the expectation can be expected to be small if the cumulative failure intensity is small, which is typical in epidemiological studies. This assumption allows us to use the approximation

$$h(t|W) \approx h_0(t) E(\exp(\beta^T X) | W).$$

If the distribution of  $X|W$  is normally distributed with constant variance, then

$$\begin{aligned}
h(t|W) &= h_0(t) \exp(\beta^T E(X|W) + \frac{1}{2} \beta^T \text{Var}(X|W) \beta) \\
&= h_0^*(t) \exp(\beta^T E(X|W)).
\end{aligned}$$

This is very useful if we are able to calculate  $E(X|W)$  (e.g. using the methods described earlier in this section) since it allows us to use standard Cox proportional hazards regression software but where we use  $E(X|W)$  in place of  $W$ . Clayton [156] uses regression calibration within risk sets to dispense with the rare-disease assumption. Regression calibration has also been shown to be approximately true for logistic regression under similar assumptions [28].

### Regression calibration for non-linear exposure–disease relationships

If a non-linear exposure–disease relationship is modelled by some function  $f(X)$  of our true exposure we cannot simply replace  $f(X)$  with  $f(E(X|W))$  in our model since  $f(E(X|W)) \neq E(f(X)|W)$ , unless  $f(\cdot)$  is affine. Therefore, for non-linear exposure–disease relationships we need to be able to calculate  $E(f(X)|W)$  which can be difficult. In the special case of a quadratic Cox model, when  $X|W$  is normally distributed with constant homoscedastic variance, we may just replace  $X^2$  with  $E(X^2|W)$  since

$$\begin{aligned}
\log h(t) &= \log h_0(t) + \beta_1 E(X|W) + \beta_2 E(X^2|W) \\
&= \log h_0(t) + \beta_1 E(X|W) + \beta_2 ((E(X|W))^2 + \text{Var}(X|W)) \\
&= \log h_0^*(t) + \beta_1 E(X|W) + \beta_2 E(X|W)^2.
\end{aligned}$$

## 4. Methods of correcting for exposure measurement error

---

Since  $\text{Var}(X|W)$  is a constant, the term  $\beta_2 \text{Var}(X|W)$  is subsumed by the baseline hazard.

If we assume that the distribution of  $X|W$  is normal then we can estimate the mean and variance parameters of the distribution from our regression calibration model; we discuss this further in chapter 6. We can then calculate  $E(f(X)|W)$  analytically or using numerical methods. We shall now consider how regression calibration can be applied to P-spline and fractional polynomial models.

### Structural P-splines

Carroll *et al.* [157] proposed a P-spline based method for correcting for exposure measurement error using regression calibration. In the situation where we observe the true exposure our P-spline model, with power spline basis, is given by

$$\log h(t|X) = \log h_0(t) + \sum_{j=0}^p \beta_j X^j + \sum_{j=1}^l \beta_{p+j} (X - t_j)_+^p. \quad (4.6)$$

When we observe mismeasured exposure measurements instead, we can apply regression calibration to obtain an estimate of the relationship we would have observed under equation 4.6,

$$\log h(t|W) \approx \log h_0(t) + \sum_{j=0}^p \beta_j E(X^j|W) + \sum_{j=1}^l \beta_{p+j} E((X - t_j)_+^p|W).$$

The terms of the model that involve expectations are dependent on the choice of distribution for  $X|W$ . As discussed above the normal distribution could be chosen. Carroll *et al.* also suggest, for purposes of robustness, that a flexible parametric family that includes the normal distribution, such as a mixture of normals, could be used. The simulation study Carroll *et al.* conducted showed that structural P-splines exhibited much smaller bias and mean squared error compared to the SIMEX method which shall be discussed below. Carroll *et al.* suggest that the expectations can be found numerically, using Gaussian-quadrature [158] for example. In appendix A.1 we show that the expectations can actually be calculated analytically when the distribution of  $X|W$  is normally distributed, and we give R code in appendix F to calculate them. Carroll *et al.* called this method structural P-splines because they specified explicitly the distribution of the true exposure distribution.

### 4.3.2 Structural fractional polynomials

We propose applying regression calibration to fractional polynomial models, in the same way that Carroll *et al.* applied it to P-spline models. We call this new model a structural fractional

polynomial model. A structural fractional polynomial of degree  $m$  is given by

$$\log h(t|W) \approx \log h_0(t) + \sum_{j=1}^m \beta_j E(H_j(X)|W).$$

As with the structural P-spline approach we need to specify the distribution of  $X|W$ . If we assume that the  $X|W$  is normally distributed then the expectation of  $X^p|W$  cannot be worked out analytically for negative, and fractional  $p$ . In fact, for fractional  $p$ ,  $X^p|W$  is not well defined since the normal distribution is defined over the entire real line. We propose shifting  $X|W$  so that there is only a negligible probability of  $X|W$  being negative, and truncating the distribution at zero. We do not prove theoretically the properties of adopting this approach, but it appears to work well in practice. We discuss how the data can be shifted appropriately further in chapter 6. In the next section we provide exact results when  $X|W$  is log-normally distributed. We shall refer to the linear model in this section as we will then obtain correction factors that would need to be applied to the intercept term, which is not present in the Cox model.

### Exact results for log-normally distributed data

If the distribution of  $X|W$  is log-normally distributed then we are able to calculate analytically  $E(H_j(X)|W)$  for structural fractional polynomial models. Suppose  $\log X$  has mean  $\mu_x$  and variance  $\sigma_x^2$ , and that  $\log W|X \sim N(\log X, \sigma_u^2)$  then  $\log X|W$  is distributed normally

$$\log X|W \sim N(\lambda \log W + (1 - \lambda)\mu_x, (1 - \lambda)\sigma_x^2) \quad (4.7)$$

where  $\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$  i.e. the RDR relating  $\log W$  to  $\log X$ . Note that the expectation of the distribution is linear in  $W$  with constant variance.

Firstly we shall consider a mismeasured exposure in the absence of confounders. Suppose that the model for the observed data is a fractional polynomial of degree 2, since this is the highest order fractional polynomial usually considered

$$Y = \alpha^* + \beta_1^* W^{(p_1)} + \beta_2^* W^{(p_2)}. \quad (4.8)$$

There are essentially four different sets of powers that need to be considered when applying regression calibration to the model in equation 4.8; by considering these scenarios we will also be working out the required functions for fractional polynomials of degree 1:

1.  $p_1 \neq p_2$  and  $p_1, p_2 \neq 0$
2.  $p_1 = 0$  and  $p_2 \neq 0$  or  $p_1 \neq 0$  and  $p_2 = 0$
3.  $p_1 = p_2 \neq 0$
4.  $p_1 = p_2 = 0$

#### 4. Methods of correcting for exposure measurement error

---

There are four expectations that we need to calculate for structural fractional polynomial models:

$$1. E(X^p|W) = k(p)W^{\lambda p} \quad (4.9)$$

$$2. E(\log X|W) = \log(W^\lambda) + (1 - \lambda)\mu_x \quad (4.10)$$

$$3. E(X^p \log X|W) = k(p) [W^{\lambda p} \log(W^\lambda) + (1 - \lambda)(\mu_x + p\sigma_x^2)W^{\lambda p}] \quad (4.11)$$

$$4. E((\log X)^2|W) = (1 - \lambda) [\sigma_x^2 + (1 - \lambda)\mu_x^2] + \log(W^\lambda) [\log(W^\lambda) + 2(1 - \lambda)\mu_x] \quad (4.12)$$

where  $k(p) = \exp\left((1 - \lambda)\left(\frac{\sigma_x^2 p^2}{2} + p\mu_x\right)\right)$ . Equation 4.9 is the  $p$ th moment, equation 4.10 the mean, and equation 4.12 the variance minus the squared mean, of the log-normal distribution 4.7. The derivation of equation 4.11 is given in appendix A.2.

These are all functions in  $W^\lambda$  so we shall consider fractional polynomial models fit using  $W^\lambda$  i.e.

$$E(Y|W^\lambda) = \alpha' + \beta'_1(W^\lambda)^{(p_1)} + \beta'_2(W^\lambda)^{(p_2)}. \quad (4.13)$$

Then we obtain the following relating the coefficients of the model in equation 4.13 to those using the true exposure

$$E(Y|X) = \alpha + \beta_1 X^{(p_1)} + \beta_2 X^{(p_2)}. \quad (4.14)$$

For each of the four cases described above

1.

$$E(Y|W) = \alpha + \beta_1 k(p_1)W^{\lambda p_1} + \beta_2 k(p_2)W^{\lambda p_2} \quad (4.15)$$

$$\text{So } \alpha = \alpha', \beta_1 = \frac{\beta'_1}{k(p_1)}, \beta_2 = \frac{\beta'_2}{k(p_2)}$$

2. Let  $p_1 = 0, p_2 \neq 0$  (The case that  $p_1 \neq 0, p_2 = 0$  is analogous)

$$\begin{aligned} E(Y|W) &= \alpha + \beta_1(\log W^\lambda + (1 - \lambda)\mu_x) + \beta_2 k(p_2)W^{\lambda p_2} \\ &= (\alpha + \beta_1(1 - \lambda)\mu_x) + \beta_1 \log W^\lambda + \beta_2 k(p_2)W^{\lambda p_2} \end{aligned} \quad (4.16)$$

$$\text{So } \alpha = \alpha' - \beta'_1(1 - \lambda)\mu_x, \beta_1 = \beta'_1, \beta_2 = \frac{\beta'_2}{k(p_2)}$$

3. Let  $p := p_1 = p_2$

$$\begin{aligned}
E(Y|W) &= \alpha + \beta_1 k(p) W^{\lambda p} + \beta_2 k(p) [W^{\lambda p} \log(W^\lambda) + (1 - \lambda)(\mu_x + p\sigma_x^2) W^{\lambda p}] \\
&= \alpha + k(p) (\beta_1 + ((1 - \lambda)\mu_x + p\sigma_x^2)) + \beta_2 k(p) W^{\lambda p} \log(W^\lambda) \quad (4.17)
\end{aligned}$$

So  $\alpha = \alpha'$ ,  $\beta_1 = \frac{\beta'_1}{k(p)} - ((1 - \lambda)\mu_x + p\sigma_x^2 \beta'_2)$  and  $\beta_2 = \frac{\beta'_2}{k(p)}$

4.

$$\begin{aligned}
E(Y|W) &= \alpha + \beta_1 (\log(W^\lambda) + (1 - \lambda)\mu_x) + \\
&\quad \beta_2 \{ (1 - \lambda) [\sigma_x^2 + (1 - \lambda)\mu_x^2] + \log(W^\lambda) [\log(W^\lambda) + 2(1 - \lambda)\mu_x] \} \\
&= [\alpha + \beta_1 (1 - \lambda)\mu_x + \beta_2 (1 - \lambda)(\sigma_x^2 + (1 - \lambda)\mu_x^2)] + \beta_1 (1 + 2(1 - \lambda)\mu_x) \log(W^\lambda) \\
&\quad + \beta_2 (\log(W^\lambda))^2 \quad (4.18)
\end{aligned}$$

So,  $\alpha = \alpha' - (1 - \lambda) \left[ \beta'_1 \frac{\mu_x}{1+2(1-\lambda)\mu_x} - \beta'_2 (\sigma_x^2 + (1 - \lambda)^2 \mu_x^2) \right]$ ,  $\beta_1 = \beta'_1 \frac{1}{1+2(1-\lambda)\mu_x}$  and  $\beta_2 = \beta'_2$ .

An alternative to calculating adjusted coefficients is to calculate the appropriate expectations, and substitute them into the fractional polynomial models in a similar way to the P-spline, and linear regression calibration models.

When we additionally have confounders we can assume that

$$\begin{pmatrix} \log X \\ \log W \end{pmatrix} | \mathbf{Z} \sim N \left( \begin{pmatrix} \mu_{x|z} \\ \mu_{x|z} \end{pmatrix}, \begin{pmatrix} \sigma_{x|z}^2 & \sigma_{x|z}^2 \\ \sigma_{x|z}^2 & \sigma_{w|z}^2 \end{pmatrix} \right)$$

so that then we have

$$\log X | \log W, \mathbf{Z} \sim N (\lambda \log W + (1 - \lambda_{x|z}) \mu_{x|z}, (1 - \lambda_{x|z}) \sigma_{x|z}^2). \quad (4.19)$$

which we note is the same as equation 4.7 except that the quantities  $\mu_{x|z}$ ,  $\lambda_{x|z}$  and  $\sigma_{x|z}^2$  are now conditional on the the confounders  $\mathbf{Z}$ . Therefore, we can use equations 4.9-4.12 to calculate the necessary expectations (conditional on  $\mathbf{Z}$ ) to use in our regression analysis. If we have baseline and repeat measurements  $W_1, W_2$  then we can estimate  $\mu_{x|z}$ ,  $\lambda_{x|z}$  and  $\sigma_{x|z}^2$  using maximum likelihood as we shall describe in section 5.2.1.

### 4.3.3 Corrected score function

The score function,  $s_X(Y, X; \beta)$ , for our disease model using the true exposure, and the true value  $\beta$  is such that

$$E(s_X(Y, X; \beta)) = 0.$$

#### 4. Methods of correcting for exposure measurement error

---

The equality is generally not true when  $X$  is replaced with the observed mismeasured exposure,  $W$ , in the equation above. The aim of the approach is to find an adjusted score function  $s_W(Y, W; \beta)$  such that

$$E(s_W(Y, W; \beta)) = 0.$$

One way to do this is to find a function,  $s_W(Y, W; \beta)$ , such that

$$E(s_W(Y, W; \beta)|X, Y) = s_X(Y, X; \beta)$$

since then

$$E(s_W(Y, W; \beta)) = E(E(s_W(Y, W; \beta)|X, Y)) = E(s_X(Y, X; \beta)) = 0$$

for the true value of  $\beta$ .

Chan and Mak [159] give a corrected score function for least squares regression and this was expanded upon by Cheng and Schneeweiss [160]. Stefanski [161] prove the general condition that for a corrected score function to exist the underlying estimating function has to be an entire function in the complex plane (i.e. it has to have no singularities). The score equation derived from the partial likelihood of the Cox model does not satisfy this condition, however Nakamura [162] gave approximately corrected score functions that were correct to first and second order.

Augustin [163] shows that if one uses the Breslow likelihood, where all censoring times in the interval  $[\tau_{j-1}, \tau_j)$  are shifted to  $\tau_{j-1}$ , and piecewise constant hazard is assumed

$$\log \ell^{Br} = \sum_{j=1}^k \left( d_j \log \lambda_j + \sum_{i \in D(\tau_j)} \beta^T \mathbf{X}_i - \lambda_j (\tau_j - \tau_{j-1}) \sum_{i \in R(\tau_j)} e^{\beta^T \mathbf{X}_i} \right)$$

then a corrected score function exists because it does not contain any singularities in the complex plane. Augustin also showed that Nakamura's corrected score function is exact under the Breslow likelihood. Augustin notes that further theoretic development is required to extend the corrected score method to the case where mismeasured covariates have a non-linear effect [164].

When there is no known corrected score function an alternative approach is to use Monte Carlo corrected scores [165]. This approach shares many similarities with SIMEX (see section 4.3.4) and relies on the result that for any suitably smooth function  $E(f(W + i\sigma_u Z))$  is unbiased for  $f(X)$ , where  $i = \sqrt{-1}$  and  $Z$  is a standard normal random variate. For example, under the classical measurement error model the naïve estimate is biased by the term  $\sigma_u^2$

$$E(W^2) = E((X + U)^2) = E(X^2) + 2E(XU) + E(U^2) = E(X^2) + \sigma_u^2.$$

However [166],

$$\begin{aligned}
 E((W + i\sigma_u Z)^2|W) &= E(W^2 - \sigma_u^2 Z^2 + 2i\sigma_u WZ|W) \\
 &= E(W^2) - \sigma_u^2 \\
 &= E(X^2).
 \end{aligned}$$

Since any terms that involve a multiple of  $i$  correspond to odd powers of  $Z^p$ , which have expectation zero, the real part of  $E(f(W + i\sigma Z)|W)$  may be taken as our estimator. The Monte-Carlo corrected score procedure involves: calculating the score equations for a large number of generated datasets,  $\tilde{\mathbf{W}} = \mathbf{W} + i\mathbf{Z}$ , and then solving the average of these score equations to obtain an estimate of  $\beta$ .

We do not consider this approach further in this dissertation, although corrected score functions for Cox models when the exposure–disease relationship is non-linear may be an interesting direction for future research.

#### 4.3.4 SIMulation EXtrapolation—SIMEX

The idea behind Simulation Extrapolation (SIMEX), proposed by Cook and Stefanski [167], is simple—we add increasing amounts of additional measurement error to the data to observe how the estimates of the model parameters vary. We then extrapolate this trend back to the case where we have no measurement error to obtain parameter estimates for the true exposure–disease relationship. SIMEX is a functional measurement error model because although we make assumptions about the measurement error distribution, we make none about the distribution of the true exposure.

We first describe the method in detail, then consider the extrapolation step, and conclude our overview of SIMEX by looking at some of the many variations of SIMEX that have been proposed.

##### Method

The method for conducting a SIMEX analysis of a Cox model with linear predictor  $f(\mathbf{W}; \beta)$  is:

1. Fit the outcome model using the observed data

$$\log h(t|W) = \log h_0(t) + f(\mathbf{W}, \beta_0)$$

to obtain a set of parameter estimates  $\hat{\beta}_0 = (\hat{\beta}_0^{(1)}, \dots, \hat{\beta}_0^{(p)})$ , where  $p$  is the number of parameters in our model.

2. Obtain an estimate of the measurement error variance,  $\sigma_u^2$ . This could be obtained, for ex-

#### 4. Methods of correcting for exposure measurement error

---

ample, from a regression calibration model, method of moments estimator, or an external reliability study.

- Let  $\zeta = \{\zeta_1, \dots, \zeta_K\}$  be a set of strictly positive scale factors,  $\zeta \in \{0.5, 1, 1.5, 2\}$  is usually chosen. Let  $B$  be the number of pseudo datasets to be generated for each value of  $\zeta$ .  $B$  should be large, typically  $B = 100 - 200$  is chosen. Then,

for  $k = 1, \dots, K$  {

for  $b = 1, \dots, B$  {

Generate  $\mathbf{W}_{k,b}^*(\zeta_k) = \mathbf{W} + \sqrt{\zeta_k} \mathbf{Z}_{k,b}$ , where  $\mathbf{Z}_{k,b}$  is a vector of randomly generated standard normal variates.

Then fit the disease model  $\log h(t|\mathbf{W}_{k,b}^*) = \log h_0(t) + f(\mathbf{W}_{k,b}^*(\zeta_k), \boldsymbol{\beta}_{k,b})$  to obtain a set of parameter estimates  $\hat{\boldsymbol{\beta}}_b(\zeta_k) = (\hat{\beta}_{k,b}^{(1)}, \dots, \hat{\beta}_{k,b}^{(p)})$  }

Calculate the mean parameter vector  $\hat{\boldsymbol{\beta}}(\zeta_k) = \frac{1}{B} \sum_{b=1}^B \hat{\boldsymbol{\beta}}_b(\zeta_k)$

}

- For each regression coefficient  $\beta^{(m)}$  in the model, we consider  $\beta^{(m)}(\zeta)$  as a function,  $f(\zeta, \beta^{(m)}(\zeta))$ , of the scale factor  $\zeta_k$  for  $k = 0, \dots, K$ , where  $\zeta_0 = 0$  corresponds to the observed exposure–disease relationship. We must specify a function to estimate the relationship, and this choice is discussed below. We fit the chosen function to  $\beta_k^{(m)}, k = 0, \dots, K$  and extrapolate back to  $\zeta = -1$  to obtain final estimates  $\hat{\beta}^{(m)} = f(\zeta = -1, \beta(\zeta))$ .

Extrapolating back to  $\zeta = -1$  corresponds to no measurement error because  $W^*(-1) = W + \sqrt{-1}\sigma_u Z = W + i\sigma_u Z$ , and taking the variance of this we obtain  $\text{Var}(W^*(-1)) = \text{Var}(W) - \sigma_u^2 = \text{Var}(X)$ .

SIMEX is intuitively appealing for many reasons. Firstly, although computationally intensive it is easy to implement in standard software and has been implemented in R by Lederer and Küchenhoff [168] and in STATA by Hardin and Schmiediche [169]. Secondly, it is a method that can be easily explained to non-statisticians and can provide graphs that help us to visualise the effects of measurement error. Figure 4.1 shows a simulated example where the true exposure–disease relationship (solid line) is U-shaped, and the variance of the observed exposure is one and a half times that of the true (i.e.  $\text{RDR}=2/3$ ). We can see that as we increase the value of  $\zeta$  the relationship is gradually attenuated. In this example, we also see that the SIMEX estimate, whereby we have extrapolated each of the sets of parameter estimates back to  $\zeta = -1$ , does very well in recreating the true exposure–disease relationship.

The theory for SIMEX was originally developed for linear models. Crainiceanu, Ruppert, and Corresh [170] establish the SIMEX procedures application to Cox models.

### Illustration of SIMEX method

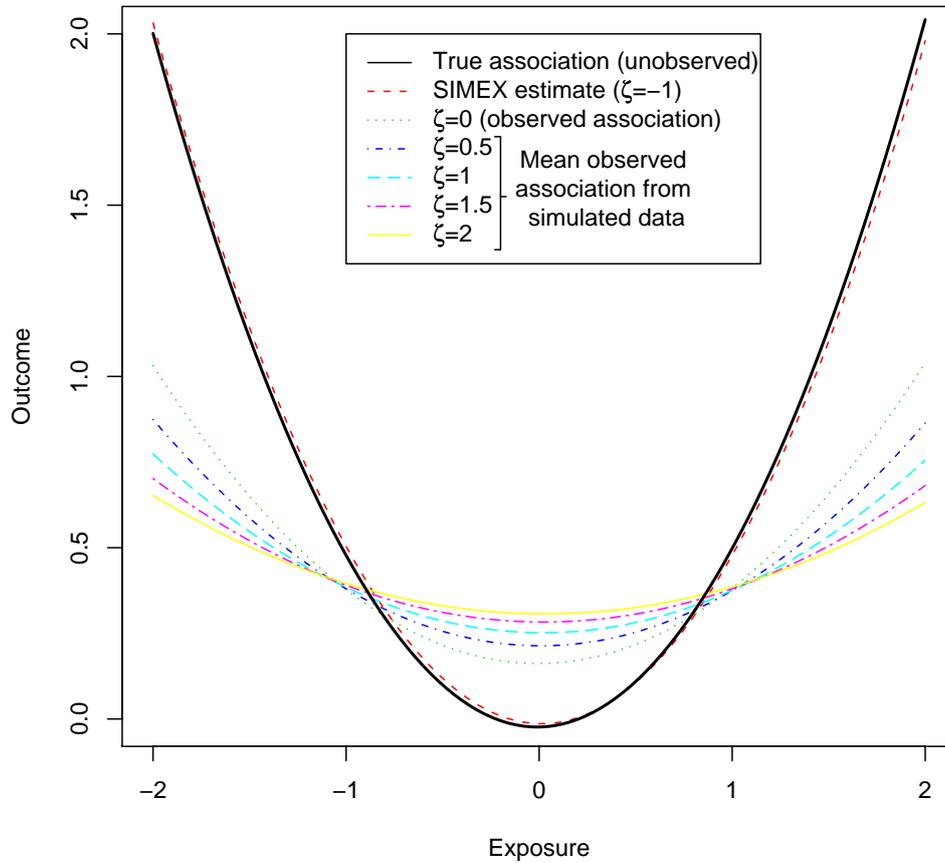


Figure 4.1: [Illustration of the SIMEX method.] Illustration of the SIMEX method for a true U-shaped relationship given by  $y = \frac{x^2}{2}$ . One thousand standard normal values were generated for the true exposure, normally distributed measurement error with variance 0.5 was added to form the observed exposure. Quadratic relationship between outcome and exposure was assumed in the SIMEX model.  $B = 200$  and rational linear extrapolant were used.

### Choice of extrapolation function

The choice of extrapolation function is key in obtaining the correct degree of correction for the effects of measurement error. Three extrapolation models are commonly used:

- Based on Taylor expansions of  $\beta(\zeta)$  about  $\zeta$ 
  - Linear —  $\beta(\zeta) = a + b\zeta$
  - Quadratic —  $\beta(\zeta) = a + b\zeta + c\zeta^2$
- Linear Rational —  $\theta(\zeta) = \gamma_1 + \frac{\gamma_2}{\gamma_3 + \zeta}$

#### 4. Methods of correcting for exposure measurement error

---

The linear rational extrapolant is exact for the linear model with a single mismeasured covariate subject to classical measurement error since

$$\begin{aligned}\beta(\zeta) &= \frac{\text{Cov}(W + \sigma_u \sqrt{\zeta} Z, Y)}{\text{Var}(W + \sigma_u \sqrt{\zeta} Z)} \\ &= \frac{\text{Cov}(X, Y)}{\sigma_x^2 + (1 + \zeta)\sigma_u^2} \\ &= \frac{\frac{\text{Cov}(X, Y)}{\sigma_u^2}}{(\frac{\sigma_x^2}{\sigma_u^2} + 1) + \zeta}.\end{aligned}$$

The linear rational extrapolant is more difficult to fit than the linear and quadratic extrapolants as it requires a non-linear least squares fitting routine. To obtain suitable starting values for the fit Carroll *et al.* [33] suggest fitting a quadratic model to  $(\zeta, \hat{\theta}(\zeta))$ , taking the fitted values from the quadratic model for  $\delta = (0, \frac{\zeta_{max}}{2}, \zeta_{max})$ , and then calculating initial estimates  $(\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)$  for  $(\gamma_1, \gamma_2, \gamma_3)$ :

$$\begin{aligned}\hat{\gamma}_3 &= \frac{d_{12}\delta_2(\delta_1 - \delta_0) - d_{01}(\delta_2 - \delta_1)}{d_{01}(\delta_2 - \delta_1) - d_{12}(\delta_1 - \delta_0)} \\ \hat{\gamma}_2 &= \frac{d_{12}(\hat{\gamma}_3 + \delta_1)(\hat{\gamma}_3 + \delta_2)}{\delta_2 - \delta_1} \\ \hat{\gamma}_1 &= a_0 - \frac{\hat{\gamma}_2}{\hat{\gamma}_3 + \delta_0}\end{aligned}$$

where  $d_{ij} = |\delta_i - \delta_j|$ . The rational linear extrapolant has a singularity in the extrapolation region if  $\gamma_3 \in (0, 1)$ . If  $\gamma_3$  were estimated to lie in this range it would usually indicate that the extrapolation function was a poor fit. Also, if the parameter relating to the quadratic term used in determining the initial estimates of  $(\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)$  is numerically zero, and we use the values of  $\delta$  suggested by Carroll and Stefanski then the equations for  $(\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)$  are singular.

Misspecification of the extrapolation function can lead to badly biased results. Under the linear rational extrapolant small negative values of  $\gamma_3$  will result in  $\hat{\beta}(-1)$  being very different to  $\hat{\beta}(0)$  as we are close to the asymptote. This could give a very biased parameter estimate if the true extrapolation function is not the rational linear.

It is essential to check plots of  $(\zeta, \hat{\beta}(\zeta))$  with the fitted function plotted for reasonableness. Since the coefficients for the quadratic extrapolation function are calculated to obtain initial values for the parameters of the non-linear least squares fit, they can be saved easily and used for comparison.

The choice of extrapolation function to use is usually pragmatic with the quadratic extrapolation function normally used in practice because of the rational linear extrapolant's numerical instability, and because the quadratic model tends to give a conservative correction for measurement error [169]. The lack of a theoretical selection of the extrapolation function, and the lack of indication as to when it may give poor results, is unsatisfactory. We can expect

quadratic extrapolation to perform best when the measurement error variance is small relative to the variance of the underlying exposure. Different extrapolation models may give good fits to the estimates  $(\zeta, \hat{\beta}(\zeta))$  but the extrapolated estimates of  $\beta(-1)$  may vary greatly between them.

The parameters within the disease model are correlated. In order to reduce this we can centre covariates, which can help to remove turning points in the extrapolation function. We also investigated allowing for the correlation by fitting a multivariate model to our extrapolation data, however we found that it did not appear to give any appreciable improvement.

### **SIMEX on the RDR scale**

The SIMEX procedure investigates how the parameter estimates for the disease model change as the measurement error variance is increased; we suggest that an alternative approach may be to perform SIMEX on the scale of the RDR, using  $\lambda$  instead of  $\sigma_u^2$  as our input into the SIMEX procedure. Under this approach we additionally have to estimate  $\text{Var}(W)$  from our dataset to give absolute values since the RDR is a ratio. We want to create datasets  $W(\lambda^*)$  such that

$$\text{Var}(W(\lambda^*)) = \frac{1}{\lambda^*} \text{Var}(X) = \frac{1}{\lambda} \text{Var}(X) + \left( \frac{1}{\lambda^*} - \frac{1}{\lambda} \right) \text{Var}(X).$$

This leads us to a simple way to create our mismeasured datasets

$$W^*(\lambda^*) = W + \sqrt{\frac{\lambda - \lambda^*}{\lambda^*}} \sigma_w^2 Z$$

for a set of values  $\lambda^* = \{\lambda_1^*, \dots, \lambda_K^*\}$ . We would then perform our extrapolation on  $(\lambda^*, W(\lambda^*))$  and extrapolate back to  $\lambda^* = 1$ . The advantage of SIMEX using our proposed scale is that for polynomial functions of  $X$  the extrapolation is exact in the corresponding polynomial in  $\lambda^*$ . Hence, we do not need to use the unstable rational linear extrapolant.

### **Application of SIMEX to spline and fractional polynomial models**

We can easily use SIMEX with spline based models, for example Crainiceanu *et al.* used splines to model the relationship between the estimated glomerular filtration rate, a measure of kidney function, and chronic kidney disease [170].

With a fractional polynomial analysis we would typically obtain models of different functional form for each dataset. One approach to using SIMEX with fractional polynomial models would be to use fractional polynomials to choose a model for the observed exposure–disease relationship and then apply SIMEX to that model. An alternative approach is to add increasing amounts of measurement error and fit the best fitting fractional polynomial model (of fixed degree) to each data set. We could then find fitted values for each model over a grid of exposure values, take the average fitted value at each point, and extrapolate back pointwise to obtain the SIMEX

#### 4. Methods of correcting for exposure measurement error

---

corrected exposure–disease relationship. This approach is very computationally intensive because for each of the simulated datasets many fractional polynomial models need to be fit. The resulting relationship would also not be of parametric form and hence could only be displayed graphically, and standard errors could not be obtained easily.

##### Variance estimation

Two methods have been proposed to estimate the variance of the parameters obtained from SIMEX—jackknife estimation [166] and asymptotic estimation [171]. We discuss only the former as the asymptotic estimation method does not apply to the Cox model. The variance of the SIMEX estimate,  $\hat{\beta}_{SIMEX}$ , can be expressed as

$$\text{Var}(\hat{\beta}_{SIMEX}) \approx \text{Var}(\hat{\beta}_{true}) + \text{Var}(\hat{\beta}_{SIMEX} - \hat{\beta}_{true}) \quad (4.20)$$

where  $\hat{\beta}_{true}$  is the hypothetical estimator calculated using true exposure. The first component of the right hand side of equation 4.20 relates to the sampling variability, whilst the second relates to measurement error variability. An estimate of  $\text{Var}(\hat{\beta}_{true})$  can be obtained by calculating for each value of  $\zeta$ ,

$$\widehat{\text{Var}}(\hat{\beta}_{true}) = \frac{\sum_{b=1}^B \text{Var}(\hat{\beta}_b(\zeta))}{B}.$$

This can then be extrapolated back to  $\zeta = -1$  in the same way as we did to find parameter estimates of  $\hat{\beta}$ .

For the second term an estimate of  $\text{Var}(\hat{\beta}_{SIMEX} - \hat{\beta}_{true})$  can be calculated as

$$\widehat{\text{Var}}(\hat{\beta}_{SIMEX} - \hat{\beta}_{true}) = -\frac{1}{B-1} \sum_{b=1}^B (\hat{\beta}_b(\zeta) - \hat{\beta}(\zeta))(\hat{\beta}_b(\zeta) - \hat{\beta}(\zeta))^T.$$

This can then be extrapolated back to the case  $\zeta = -1$  analogously. Note that we require the minus sign since the estimate will be positive for  $\zeta > 0$ , zero when  $\zeta = 0$ , and *negative* when extrapolating back to  $\zeta = -1$ . We can see this by considering the components of the variance

$$\text{Var}(\hat{\beta}_b(\zeta) - \hat{\beta}(\zeta)) = \text{Var}(\hat{\beta}_b) + \text{Var}(\hat{\beta}) - 2 \text{Cov}(\hat{\beta}_b(\zeta), \hat{\beta}(\zeta)).$$

$\text{Cov}(\hat{\beta}_b(\zeta), \hat{\beta}(\zeta)) = \text{Var}(\hat{\beta}(\zeta))$  since  $\hat{\beta}(\zeta) = E(\hat{\beta}(\zeta)|X)$  and we show in appendix A.3 that  $\lim_{\zeta \rightarrow -1} \text{Var}(\hat{\beta}_b(\zeta)) = 0$ . Therefore,

$$\lim_{\zeta \rightarrow -1} \text{Var}(\hat{\beta}_b(\zeta) - \hat{\beta}(\zeta)) = \lim_{\zeta \rightarrow -1} -\text{Var}(\hat{\beta}(\zeta)) = -\text{Var}(\hat{\beta}(-1)) = -\text{Var}(\hat{\beta}_{SIMEX}).$$

In practice, rather than extrapolating each of the two components separately, it is easier to

---

calculate

$$\widehat{\text{Var}}(\hat{\beta}_{SIMEX}(\zeta)) \approx \frac{\sum_{b=1}^B \text{Var}(\hat{\beta}(\zeta))}{B} - \frac{1}{B-1} \sum_{b=1}^B (\hat{\beta}_b(\zeta) - \hat{\beta}(\zeta))(\hat{\beta}_b(\zeta) - \hat{\beta}(\zeta))^T$$

and then extrapolate this back to  $\zeta = -1$  [33].

The SIMEX procedure has been used in a number of applications [172–174]. Although SIMEX has been used for linear exposure–disease relationships we believe that better, less computationally intensive methods, such as regression calibration should be used in this case. In principle SIMEX can be used with any distribution for the measurement error, however a normal distribution is almost always assumed. A drawback of SIMEX is that there exists no model selection procedure so one needs to make the correct selection of naïve model. As we saw in chapter 3, the observed exposure–disease relationship may appear to have a different functional form to the true relationship due to measurement error.

### SIMEX variants

Many variants of SIMEX have been proposed. Empirical SIMEX is a variant of SIMEX that does not assume that the measurement error is normally distributed or homoscedastic [175]. Instead of forming pseudo datasets by adding on measurement error, pseudo data are created by taking linear combinations of replicate measurements. This means that replicate measurements must be available for all individuals. It is uncommon in practice for all individuals to have at least one replicate measurement, even if the study is designed so that all individuals have replicate measurements taken, since at least some individuals will typically be lost to follow-up. This reduces the practical usefulness of this method.

Nolte [176] has developed a variant of SIMEX for multiplicative measurement error and Ronning and Rosemann [177] have developed a method they call generalised SIMEX (GSIMEX) for when the measurement error in the response is correlated with that in the predictor. Staudenmayer and Ruppert [178] suggest how SIMEX can be applied to local polynomial regression. Alpizar-Jara *et al.* [179] give a method for dealing with systematic measurement error when applying SIMEX.

### 4.3.5 Multiple imputation

Cole, Chu and Greenland [180] proposed that multiple imputation, which is used to impute missing data, could also be used for measurement error problems when we have gold standard measurements within a validation substudy. We discuss in chapter 5 how this method can be applied when we have repeat measurements. We can view the true exposure as being missing for those individuals not in the validation study. Multiple imputation can be used to correct for differential measurement error in our exposure as well as non-differential error, which methods

## 4. Methods of correcting for exposure measurement error

---

such as regression calibration and SIMEX cannot. We can create imputed datasets

$$X_{IM}(W, Y) = E(X|W, Y) + e \quad (4.21)$$

where  $E(X|W, Y)$  can be estimated using a regression calibration model e.g.

$$X = \alpha_0 + \alpha_1 W + \alpha_2 Y + \epsilon$$

and  $e$  is a random draw from the residuals,  $\epsilon$ , of this regression. The regression calibration model may be more complex and contain for example polynomial functions of  $W$ , or an interaction between  $W$  and  $Y$ . If we make the assumption of non-differential error then there will be no  $WY$  interaction/

We can form  $K$  datasets using equation 4.21,  $X_{IM}^{(k)}(W, Y)$ ,  $k = 1, \dots, K$  and fit the outcome model to each of these imputed datasets to obtain parameter estimates  $\hat{\beta}_{IM}^{(k)}$ . These parameter estimates can then be combined using Rubin's rules [181] to find estimates of the regression parameters for the true exposure–disease relationship

$$\hat{\beta} = \frac{1}{K} \sum_{k=1}^K \hat{\beta}_{IM}^{(k)}$$

with variance

$$\widehat{\text{Var}}(\beta) = \frac{1}{K} \sum_{k=1}^K \widehat{\text{Var}}(\hat{\beta}_{IM}^{(k)}) + \frac{K+1}{K(K-1)} \sum_{k=1}^K (\hat{\beta}_{IM}^{(k)} - \hat{\beta}_{IM})^2.$$

Multiple imputation can easily be performed using the `mice` [182] package in R.

What is described above is suitable for a linear outcome model. In the Cox model instead of a simple outcome  $Y$ , our outcome consists of an event indicator and the length of the observation period. In this situation White and Royston [183] propose that the event indicator and the Nelson-Aalen estimator of the cumulative hazard function are used in the imputation model.

### 4.3.6 Moment reconstruction

Moment reconstruction [184] aims to recreate the first two joints moments of the true exposure distribution with  $Y$  by generating

$$X_{MR}(W, Y) = E(X|Y) + G(W - E(W|Y))$$

where

$$G = \sqrt{\frac{\text{Var}(X|Y)}{\text{Var}(W|Y)}}.$$

Moment reconstruction allows for differential measurement error and provides an exact, rather than an approximate, correction for the effects of exposure measurement error in a logistic regression when the exposure–disease relationship is linear. Remember that it was noted earlier that regression calibration was only approximate for this model.

Thomas, Stefanski, and Davidian [55] have recently proposed a method, Mean Adjusted Imputation, that extends the principle of moment reconstruction and aims to find  $\hat{X}$  such that  $\hat{X}$  shares its first  $r$  moments with  $X$ , and also the cross-products of  $\hat{X}^r Y$  and  $\hat{X}^r Z$ . This method relies on the fact that if  $W \sim N(X, \sigma_u^2)$  then  $E(\sigma_u^r H_r(\frac{W}{\sigma_u})) = X^r$  where  $H_r(x)$  is the  $r$ th order Hermite polynomial

$$H_{n+1}(x) = xH_n(x) - (n-1)H_{n-1}(x)$$

with  $H_0(x) = 1$  and  $H_1(x) = x$ . Their method minimises the squared distance between the observed exposure and the estimated true exposure, subject to the constraint that the moments and cross products are equal to their corresponding estimates  $\hat{n}_{rk}$ . This method recreates the corrected score estimators given by Cheng and Schneeweiss [160] for polynomial functions in linear models.

Moment reconstruction and Mean Adjusted Imputation both use the outcome  $Y$  in obtaining their imputed values, however in Cox models the outcome consists of two components, event and censoring indicators. Thomas, Stefanski and Davidian [55] considered three possible approaches to including both the event and censoring indicators:

1. Include the censoring indicator as a variable in the moment matching.
2. Perform the moment matching within each level of the censoring indicator.
3. Perform the moment matching for each risk set.

They found that there was little difference between the results using the three methods and therefore recommend the first approach because it is the simplest to implement. The results of their simulations were that the corrected score method is preferable for linear models, over their proposed method and regression calibration. However, the corrected score method cannot be used for non-linear exposure–disease relationships and the estimation of  $E(X^p|W)$  requires assumptions to be made about the distribution of  $X|W$ . This method makes no assumptions about the distribution of  $X|W$  and can be computed quickly using the code given by the authors.

### 4.3.7 Density estimation of the true exposure

Structural correction methods require us to make distributional assumptions. One way to avoid making these assumptions is to try and recreate the distribution of the true exposure  $X$  using non-parametric density estimation. In some situations the distribution of the true exposure may be of interest in itself. Non-parametric density estimation in the presence of measurement error

## 4. Methods of correcting for exposure measurement error

---

is difficult because we observe the convolution of densities

$$f_W = f_X * f_U.$$

In order to get  $f_X$  it is necessary to find the deconvolution of the two densities; this is a difficult problem. Fan and Truong [185] discuss non-parametric kernel regression when the exposure is measured with error. Delaigle and Hall [186] propose a SIMEX procedure for ascertaining an optimal bandwidth since the standard methods are suboptimal when measurement error is present. For further discussion of deconvolution based methods see Carroll *et al.* [33].

### 4.3.8 Bayesian approaches

Although we explained in chapter 2 that we would not be considering Bayesian approaches in this dissertation, we give some methods for completeness. Berry, Carroll and Ruppert [187] suggested Bayesian spline modelling approaches. The first method they suggested, which they call iterative conditional modes, is only partially Bayesian and maximises each component of the likelihood componentwise. The second method is a fully Bayesian model that utilises the results from the first method to provide starting values for the Markov chain Monte Carlo (MCMC). This method performed better than the structural splines described above in their simulation.

Cheng and Crainiceanu [188] have developed a Bayesian approach for Cox models. They use a piecewise constant function to fit the log baseline hazard and a low-rank spline to fit the log-hazard for the exposure of interest using the Bayesian equivalent of the difference penalty for P-splines, random walk priors. To fit their spline based model to an example consisting of 15,792 patients took 59.1 hours to obtain 10,000 samples for one MCMC chain on a modern PC. The dataset we shall be considering in later chapters will be over fifteen times this size and hence exemplifies why we will not be considering Bayesian approaches.

## 4.4 Correction methods for grouped data

In chapter 2 we discussed the problems associated with grouped exposure analyses and in section 4.2 we noted that categorising a continuous mismeasured exposure typically induces differential error. Despite these problems, grouped exposure analyses are frequently used in practice and therefore we consider some of the methods that can be used to correct them for the effects of exposure measurement error. When the exposure is grouped, measurement error manifests itself as misclassification between groups.

---

#### 4.4.1 MacMahon’s method

MacMahon *et al.* [35] used a simple method for correcting for the effects of non-differential exposure measurement error under a grouped exposure analysis where replicate measurements of exposure were available. We shall refer to this approach as MacMahon’s method, as used by the Fibrinogen Studies Collaboration [189]. This approach has also been referred to as the MacMahon-Peto method [13, 190] or Peto’s method (although this is potentially confusing because Peto’s method also refers to a method Peto developed for meta-analysis). MacMahon’s method has been used by many studies [13, 36, 191].

The idea behind MacMahon’s method is that instead of plotting the hazard ratios obtained from a grouped exposure analysis against the mean value of the observed exposure values within each group, we plot the hazard ratio for each group against the mean of replicate exposure measurements maintaining the baseline groups. The effect of exposure measurement error is to ‘mix up’ the groups we would have observed using the true exposure—the highest and lowest groups contain disproportionately many individuals whose observed baseline exposure measurements happened to be higher/lower than their ‘true’ exposure level. MacMahon’s method obtains an unbiased estimate of the mean true exposure of the observations in each group. We shall investigate the performance of MacMahon’s method for recreating the shape of the exposure–disease relationship in chapter 5.

#### 4.4.2 Misclassification SIMEX—MCSIMEX

Proposed by Küchenhoff, Mwalili, and Lesaffre [192] the misclassification-SIMEX (MCSIMEX) is the extension of the SIMEX procedure of section 4.3.4 to the case where we have discrete exposure groups. Let  $G = \{G_i, i = 1, \dots, g\}$  be the levels that our discrete covariate can take and define the misclassification matrix,  $\Pi = \{\pi_{ij}\}$  by

$$\pi_{ij} = P(W \in G_i | X \in G_j).$$

$\Pi$  is an input into the MCSIMEX method and must be estimated, for example using a gold standard measure.

The model for the observed exposure–disease relationship is

$$\log h(t) = \log h_0(t) + \sum_{i=2}^g \beta_i I(W \in G_i).$$

In the SIMEX procedure we scaled the measurement error variance by a scale factor  $\zeta$ . In MCSIMEX we increase the degree of misclassification using  $\Pi^\zeta := E\Lambda^\zeta E^{-1}$ , with  $\Lambda$  being the diagonal matrix of eigenvalues of  $\Pi$  and  $E$  the corresponding matrix of eigenvectors.

## 4. Methods of correcting for exposure measurement error

---

We can generate misclassified data as  $W(\zeta) := \text{MC}(\Lambda^\zeta)(W)$  where  $\text{MC}(\Lambda^\zeta)$  is the misclassification operation. For  $W \in G_i$  the misclassification operation works by sampling from  $G$  where  $G_j$  is chosen with probability  $\pi_{ij}$ . The SIMEX procedure can be applied using this approach for creating misclassified data for increasing values of  $\zeta$ . It is suggested that the linear or quadratic extrapolation functions are used, or  $\beta(\zeta) = \exp(a + b\zeta)$  which is the correct extrapolation function for a logistic regression with misclassified response. When the misclassification probabilities are small  $\Lambda^\zeta$  may not exist for small values of  $\zeta$ . Küchenhoff recommends an approximation procedure [193] that gives a matrix  $\Lambda^*$  for which  $\Lambda^{*\zeta}$  exists when  $\Lambda^\zeta$  does not.

This method does not allow for differential measurement error because it assumes the probability of misclassification is the same for all individuals within each group. When estimating the shape of an exposure–disease relationship the exposure will typically be grouped into five or more groups, this would require the estimation of a large misclassification matrix as an input into the MCSIMEX procedure.

### 4.4.3 Group-SIMEX

We propose a SIMEX based approach which we call group-SIMEX for when we have observed mismeasured continuous exposure measurements but wish to perform the analysis on groups of the exposure. The aim is to get around the problems caused by differential error, by adding random measurement error to the continuous exposure before applying the grouping. The group-SIMEX procedure is:

- Group the continuous data  $W$  into groups  $G = \{G_1, \dots, G_g\}$
- Fit the naïve model to  $G$  to obtain  $\beta_0 = \{\beta_{G_2}, \dots, \beta_{G_g}\}$  (where  $G_1$  is the reference group)
- Obtain an estimate of the measurement error variance,  $\sigma_u^2$
- Generate normally distributed random pseudo errors:  $\sqrt{\zeta}\sigma_u Z$
- Add the pseudo errors to the observed exposure and group the data according to the cutpoints used to define  $G$  to obtain  $G_\zeta$ .
- Fit the model to  $G_\zeta$  for each  $\zeta$
- For each parameter  $\beta_{G_i}$  extrapolate back to the case of no measurement error ( $\zeta = -1$ ).

We considered applying the quadratic and rational linear extrapolation functions for extrapolating in group-SIMEX:

1. Quadratic—The quadratic extrapolant is unable to sufficiently capture the curvature in the extrapolation data, causing it to extrapolate back to parameter estimates that undercorrect for the effects of measurement error.

---

2. Rational linear extrapolant—The rational linear extrapolant can fit the  $\hat{\beta}(\zeta)$  very well however:

- the non-linear least squares routine required to fit the extrapolation model often does not converge;
- if the parameter value does not change much with the addition of increasing amounts of measurement error, this can cause the fitted function to have an asymptote in the range of extrapolation;
- the fit is very sensitive and small changes in the simulated values can lead to wild changes in the extrapolated value;
- when simulating simple examples we found that the extrapolation often would often cause groups to switch their ordering.

We considered using a cubic extrapolant, and we also considered weighting the extrapolation points, neither of which worked well. The cubic extrapolant did tend to pick up the curvature better than the quadratic, but problems were experienced with turning points in the extrapolation region.

We found that it may be optimal to down weight or even exclude  $\zeta = 0$  (the naïve regression) and instead introduce a ‘small’ value for  $\zeta$  to simulate for;  $\zeta = 0$  is based on only one observation whereas  $\zeta > 0$  is based on  $B$  replicates. The parameters of an individual dataset can be very sensitive to one individual moving between groups. This is more of a problem with the Cox model because when individuals with events associated with them move between exposure groups (as a result of measurement error) it can cause large changes in parameter values. For each  $b$ ,  $\beta_b(\zeta)$  is not a continuous function and is instead a step function since  $\zeta_b(\zeta)$  will remain constant in some range  $(\zeta, \zeta + \epsilon]$  until an observation crosses the boundary between groups. For large sample sizes where the number of steps will typically be large, and since  $\hat{\beta}(\zeta) = \frac{1}{B} \sum_{b=1}^B \beta_b(\zeta)$ , we expect that  $\hat{\beta}(\zeta)$  will behave the same as if it were continuous. We, therefore, suspect that group-SIMEX will work best when the data contain many events within each group. This is most likely to be the case when we have few groups, and a high proportion of events, or a large sample size.

#### 4.4.4 Natarajan’s method

Natarajan [194] proposed the following regression calibration based procedure for a dichotomous predictor where we have observed the continuous mismeasured exposure,  $W$ , and true exposure  $X$  in a validation substudy:

1. Fit a calibration model for  $E(X|W, Z)$  for  $X$ ,  $W$  and  $Z$  in the validation substudy. Using the regression coefficients obtained from the calibration model, calculate the predicted value  $\hat{X}_i = \hat{E}(X_i|W_i, Z_i)$  for each subject  $i$  who is not in the validation study and let

## 4. Methods of correcting for exposure measurement error

---

$\hat{X}_i = X$  for those individuals in the validation study.

2. Calculate  $X_{rcb} = I(\hat{X} > c)$  for chosen cutpoint  $c$ .
3. Use  $X_{rcb}$  in the exposure–disease model.

This method is similar in principle to the group-SIMEX procedure discussed above whereby we perform our measurement error correction on the continuous exposure before forming groups. Although Natarajan dichotomised the exposure the method can be used to split the exposure into multiple exposure groups. When we only have repeat measurements then we can still use this method but  $\hat{X}_i$  will be imputed from the regression calibration model for all individuals.

### 4.5 Calculation of confidence intervals

In the description of each of the methods above, we have not described how to take account of the additional uncertainty from our measurement error model, in the disease model. In almost all cases we can take this uncertainty into account using the nonparametric bootstrap, the easiest method being resampling with replacement from the original data [195, 196], although other methods for censored data have been considered [197]. Bootstrap estimation of confidence intervals can be computationally intensive, especially when the dataset is large. This is not the only way to take account of the uncertainty, for example sandwich estimators can sometimes be found [28, 33], and Wood [198] proposed a multiple imputation based approach.

When using regression calibration there typically will be much more uncertainty in the second stage Cox model than the regression calibration. Hence, many authors choose to make no correction. Rosner [28] considered how the number of replicate measurements per individual, and the number of individuals with replicates, affected the size of the confidence intervals for parameter estimates within their logistic regression models and found that when the number of individuals with replicate measurements was small then increased numbers of replicate measurements per individual gave much greater precision. However, as the number of individuals with replicates increased, the number of replicates per individual made little difference. Beyond 100 individuals with replicates, the increase in the size of the confidence intervals for the parameters of the disease model, due to the uncertainty in the measurement error model, was small. This suggests that little is lost by not correcting the confidence intervals when the measurement error model can be well estimated. It also suggests that validation studies should aim to obtain a single repeat on as many individuals as possible, rather than obtaining many replicates per individual, unless the number of individuals in the validation study is to be small.

### 4.6 Conclusion

Usually we require additional information in order to estimate the measurement error model parameters. This additional information can be from an (alloyed) gold standard, replicate mea-

---

surements or an instrumental variable. Regression calibration is arguably the most commonly used approach for correcting for the effects of exposure measurement error whereby we take the expectation of the true exposure given the observed. When regression calibration is applied to non-linear exposure–disease relationships structural assumptions usually have to be made. Regression calibration has been applied to P-splines. We have shown how it can be applied to fractional polynomials and given analytical correction formulae for when the distribution of  $X|W$  is log-normal. SIMEX is an intuitive approach that makes no assumptions about the distribution of the true exposure. SIMEX involves two steps: SIMulation, whereby we create multiple mismeasured datasets and fit our disease model, and EXtrapolation, whereby we extrapolate the average parameter estimates from the first stage back to the case where we have no error. The choice of extrapolation function is usually pragmatic and SIMEX tends to work best when the measurement error variance is small. Many other methods have been proposed including, moment reconstruction, multiple imputation, Bayesian approaches, and methods that aim to recreate the true exposure distribution via deconvolution.

It is common for continuous exposures to be grouped; this introduces differential measurement error. Measurement error manifests itself as misclassification when continuous exposures are grouped. MacMahon’s method is a simple method that uses repeat measurements to find unbiased estimates of the group means. MCSIMEX is an extension of SIMEX to categorical variables but it assumes non-differential misclassification. We therefore proposed group-SIMEX which adds measurement error to the continuous exposure in the simulation step before the exposure is grouped. Natarajan applied regression calibration to the continuous exposure before it was grouped, again avoiding the need to correct for differential error. Confidence intervals for the analysis model should be corrected for the uncertainty in the measurement error model. This can be difficult or computationally intensive and the impact may be small if the measurement error model can be well estimated. In chapter 5 we investigate the performance of several of the methods described in this chapter.

#### **4. Methods of correcting for exposure measurement error**

---

# Chapter 5

## Performance of correction methods

This chapter considers the performance of some of the methods discussed in chapter 4 through the use of simulation studies.

Firstly, we look at methods of correcting for exposure measurement error in grouped exposure analyses. In simulation study 1 we extend the simulation study of chapter 3 to consider MacMahon's method. We then investigate some of its theoretical properties. In simulation study 2 we compare the methods of Natarajan, regression calibration, group-SIMEX, moment reconstruction, and multiple imputation for correcting a grouped exposure analysis for the effects of both random and systematic measurement error, when the underlying exposure is continuous. In simulation study 3 we consider the performance of continuous models for the exposure–disease relationship that are corrected for measurement error, using structural fractional polynomial and P-spline models, again by extending the simulation study of chapter 3.

### 5.1 Simulation study 1: Evaluation of MacMahon's method

#### 5.1.1 MacMahon's method

Under MacMahon's method we plot the hazard ratio for each group in a grouped exposure analysis against an unbiased estimate of the mean true exposure within each group maintaining the baseline groups. These unbiased estimates are usually obtained from replicate exposure measurements.

Although MacMahon's method has been widely used, its ability to correct for the effects of exposure measurement error in a grouped exposure analysis, where the true exposure–disease relationship is non-linear, has not been previously investigated. We choose to consider this method alone, and not as part of the next section where we consider a number of methods for correcting grouped exposure analyses, because MacMahon's method does not supply us with corrected estimates of the hazard ratio for each group. MacMahon's method merely is a way of

## 5. Performance of correction methods

---

visually displaying a measurement error corrected shape for the exposure–disease relationship.

### 5.1.2 Simulation procedure

The data generation mechanism and the scenarios considered in this simulation are the same as in chapter 3. We additionally generated a replicate exposure measurement for each individual  $W_{i2} = X_i + U_{i2}$ , where  $U_{i2}$  is normally distributed with mean zero and with the same variance as  $U_i$ , and was generated independently of both  $X$  and  $U_i$ . We performed the same grouped exposure analysis as in chapter 3, but we additionally calculated the mean of the replicate exposure values according to the baseline exposure groups within each group.

### 5.1.3 Results

In figure 5.1 we plot the mean of the replicate exposure values according to the baseline exposure groups within each group, against the average estimated hazard ratio from the 1,000 grouped exposure analyses. We can see that MacMahon’s method performs well for all shapes for the exposure–disease relationship except for the threshold relationships. There does appear to be some deviation from the true shape of the exposure–disease relationship when the shape of the exposure–disease relationship displays greater curvature. MacMahon’s method overcorrects for the exposure groups most distant from the mean under the J-shaped, U-shaped, and increasing quadratic shapes for the exposure–disease relationship. MacMahon’s method performs particularly badly against the true shape of the threshold relationships this is because exposure measurement error means that individuals above and below the threshold are mixed in the baseline groups, which occludes the threshold.

### 5.1.4 Discussion

The baseline groups of observed exposure will typically be composites of individuals whose true exposure values belong to several groups. The corrected group means we obtain from MacMahon’s method are not the means of the groups of the true exposure (which is what we are usually interested in), but are the means of the true exposure for individuals who formed the baseline groups. Hence, the group means obtained from MacMahon’s method lie closer to the mean of the exposure distribution, than the means of the corresponding groups of the true exposure. We therefore view the exposure–disease relationship over a reduced range. As we can see in figure 5.1, this reduction is considerable when the measurement error variance is large. As the group means under MacMahon’s method are not the means of the true exposure groups it severely hinders the interpretation of the plots obtained using this approach.

We can show theoretically that MacMahon’s method does not provide us with the shape of the true exposure–disease relationship unless the relationship is linear. Let  $W_1^{(k)} = 1$  if the baseline exposure measurement is in the  $k$ th category and zero otherwise. An unbiased estimate

---

of true exposure for individuals whose observed baseline exposure is in the  $k$ th category is  $E(W_2|W_1^{(k)} = 1)$ . If the true exposure–disease model is  $\log(h(t|X)) = \log h_0(t) + \beta g(X)$ , for some function  $g(X)$ , then the log-hazard for an individual with exposure value  $E(W_2|W_1^{(k)} = 1)$  is  $\beta g(E(W_2|W_1^{(k)} = 1))$ . However, under MacMahon’s method we plot  $\beta E(g(X)|W_1^{(k)} = 1)$ . These two quantities are only the same if  $g(X)$  is linear or if  $\text{Var}(W_2|W_1^{(k)}) = 0$ . How far the log-hazard ratios obtained from MacMahon’s method are from the true exposure–disease relationship will depend on the  $\text{Var}(W_2|W_1^{(k)} = 1)$ . In general this bias will be smaller when more groups are used. What MacMahon’s method does provide us with is hazard ratios that are corrected for exposure measurement error based on groups of the observed baseline exposure, although this is of little practical interest.

In the simulation we considered quantile groups of the observed data, however groups are often chosen with respect to clinically relevant cutpoints. In this case, the application of MacMahon’s method will typically give results where the group means lie to one side of the category, or may not even lie within the defined baseline category at all. This can severely hinder interpretation.

MacMahon’s method is a simple, quick, and easy to implement method of measurement error correction. We have discussed how MacMahon’s method will provide us with a plot of the true exposure–disease relationship when it is linear. Although if the relationship is linear, we can easily fit a model with a linear term in the predictor, and correct for measurement error using an estimate of the RDR as described in chapter 4. Under this approach we do not view the relationship over a reduced range. When the true exposure–disease relationship is non-linear we have seen that MacMahon’s method will generally provide us with a graph for the exposure–disease relationship that is not too dissimilar from the true exposure–disease relationship, unless the relationship exhibits a threshold.

We have, however, shown that MacMahon’s method does not provide us with the *true* shape for the exposure–disease relationship, unless the relationship is linear. We therefore recommend that MacMahon’s method may be used to provide a ‘quick and dirty’ approximation to the true exposure–disease relationship, but that other correction methods should be considered if our aim is to accurately characterise the shape of the true exposure–disease relationship.

## 5. Performance of correction methods

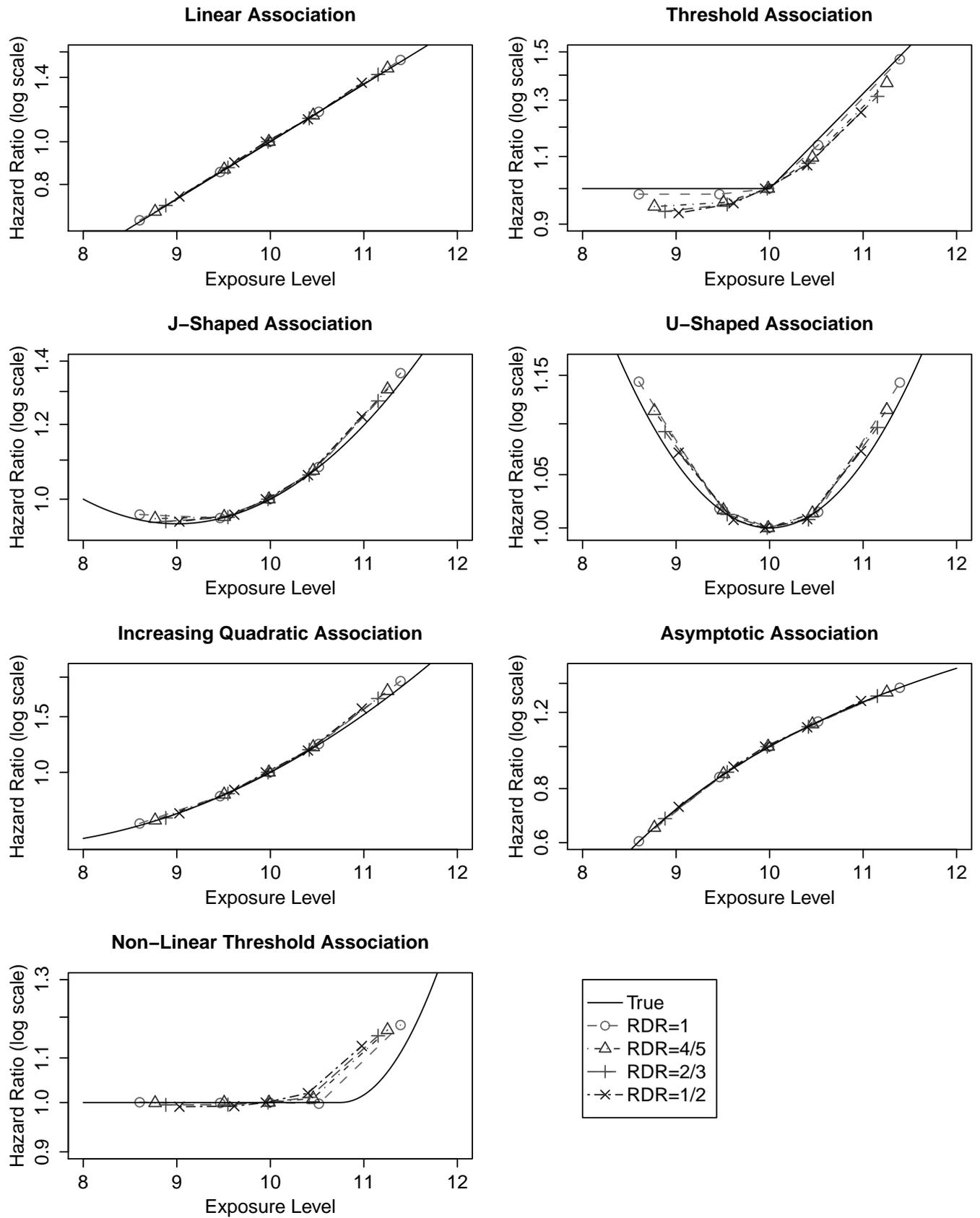


Figure 5.1: Grouped exposure analysis of the exposure–disease relationship using MacMahon’s method for correcting for the effects of exposure measurement error.

---

## 5.2 Simulation Study 2: Correction methods for categorised continuous exposures

As we saw in chapter 3 there are a number of methods for correcting for the effects of measurement error in a continuous exposure on an exposure-disease relationship. There are also methods for correcting for misclassification of fundamentally categorical exposures, that is categorical exposures which are not derived from an underlying continuous measure, and are based on estimated misclassification probabilities and focus on a binary outcome [199–201]. However, methods for correcting for the effects of measurement error in continuous exposures on the relationships found in categorised exposure analyses have received very little attention.

This simulation study was initially motivated by the paper of Natarajan [194], whose proposed method for correcting a grouped exposure analysis of a continuous exposure subject to measurement error was described in section 4.4. As we shall see later in this section, Natarajan’s approach is flawed.

We therefore decided to compare Natarajan’s method for correcting for the effects of misclassification when a continuous exposure subject to measurement error is categorised, with methods introduced in chapter 4. We consider an alternative regression calibration based approach, moment reconstruction and multiple imputation of the continuous exposure followed by categorisation, and group-SIMEX. In this section we restrict our attention to when the true exposure–disease relationship is linear. We use logistic regression in our simulation studies as this is the situation considered by Natarjan in her study, and as mentioned in chapter 1 logistic regression is often used to model survival data. We provide results in appendix B for linear regression and discuss how these models can be applied to survival data in the discussion.

### Observed and naïve estimates

Suppose that we are interested in the relationship between a true continuous exposure,  $X$ , and a dichotomous outcome,  $Y$ . We focus on a dichotomised version of the underlying continuous exposure,  $X_C = I(X > C)$ , where  $C$  is a pre-defined cutpoint. In a grouped exposure analysis of the exposure–disease relationship using the dichotomised exposure  $X_C$  we fit the model

$$g(E(Y|X_C)) = \beta_0 + \beta_1 X_C \quad (5.1)$$

where  $g(\cdot)$  is the link function e.g. the identity function for linear regression or the logit function for logistic regression. We are interested in estimating the parameter  $\beta_1$ , which is the regression coefficient in a linear regression, or the log-odds ratio in a logistic regression.

The observed continuous exposure,  $W_1$ , is assumed to be given by the measurement error model

$$W_1 = \alpha_0 + \alpha_1 X + U_1. \quad (5.2)$$

## 5. Performance of correction methods

---

where  $U_1$  has mean 0 and variance  $\sigma_{U_1}^2$ , and is independent of  $X$  and  $Y$ . This implies that the measurement error is non-differential. The model encompasses both classical and systematic measurement error models.

The naïve approach to estimating  $\beta_1$  is to perform a grouped exposure analysis on the dichotomised observed exposure,  $W_{1C} = I(W_1 > C)$ :

$$g(E(Y|W_{1C})) = \beta_0^* + \beta_1^* W_{1C}. \quad (5.3)$$

The estimate of  $\beta_1^*$  obtained from the naïve model is not an unbiased estimate of  $\beta_1$  because of the misclassification in  $W_{1C}$  caused by measurement error in the observed continuous exposure  $W$ . Therefore, we need to correct for the misclassification in  $W_{1C}$  in order to obtain an unbiased estimate of the parameter  $\beta_1$ .

### Scenarios

As discussed in chapter 4, we require additional exposure measurements to be able to estimate the parameters of the measurement error model 5.2. In this section we consider the following four scenarios:

- The observed exposure is subject to *classical* measurement error i.e.  $\alpha_0 = 0, \alpha_1 = 1$ :

Scenario (1a) We assume that we have a validation substudy where we have observed the true exposure  $X$  for a random subset of individuals. An example of this is when anthropometric variables such as weight and height are self-reported by all study participants, and are measured by a nurse for a random subsample of individuals.

Scenario (1b) We assume that we have replicate measurements available for a subset of individuals. In this scenario we have  $W_j = X + U_j, j = 1, 2$  where  $U_1, U_2$  are independent of each other,  $X$ , and  $Y$ . An example of this situation is where repeated blood pressure measurements are available only for a subset of individuals and the exposure of interest is usual level.

- The observed exposure is subject to *systematic* measurement error i.e. such that  $\alpha_1 \neq 1$ :

Scenario (2a) Similarly to scenario (1a) we assume that we have a validation substudy where we have observed the true exposure  $X$  for a random subset of individuals.

Scenario (2b) When our observed exposure is subject to systematic measurement error, we require that a different measure of exposure is available for a subset of individuals, which is an unbiased measure of the true exposure. For most correction methods we require two additional measurements of this different exposure to be available in order to estimate the parameters of the measurement error model of equation 5.2. We therefore assume that we have two additional exposure measurements  $W_j = X + U_j, j = 2, 3$  where the errors  $U_2, U_3$  are independent of each other,  $X, Y$  and  $U_1$ , and have the

same variance  $\sigma_{U_2}^2$ . An example of this is in nutritional epidemiology where we may have obtained FFQs from all individuals, but where food record measurements are available for a subset of individuals.

We do not consider the situation that  $\alpha_0 \neq 0$  in this chapter. If  $\alpha_0 \neq 0$  we can use the methods described for scenarios (1a) or (1b) if the cutpoint  $C$  is a quantile of the exposure distribution, or the methods of scenarios (2a) or (2b) if the cutpoint is fixed.

## 5.2.1 Methods

In this section we start by describing how we can fit measurement error models using method of moments and maximum likelihood approaches, we then give a brief description of each of the methods to be investigated in the simulation study.

### Estimating the parameters of the measurement error model

In chapter 4 we described methods of correcting for measurement error, however we did not detail methods for estimating the parameters of the measurement error model. When we have observed the true exposure for a subset of individuals, such as under scenarios (1a) and (2a), then these parameters are easily estimated. For example, to estimate  $E(X|W_1)$  and  $\text{Var}(X|W_1)$  we can regress  $X$  on  $W_1$  to obtain an estimate of  $E(X|W_1)$  and the variance of the residuals from this model gives an estimate of  $\text{Var}(X|W_1)$ .

When we only have repeat exposure measurements available, such as under scenarios (1b) and (2b), then estimation of the measurement error parameters is more complicated. If  $W_1, W_2$  are replicates subject to classical error, with measurement error variance  $\text{Var}(U)$ , then we can still regress  $W_2$  on  $W_1$  to obtain an estimate of  $E(X|W_1)$ . However, the variance of the residuals from this model is  $\text{Var}(X|W_1) + \text{Var}(U)$ . One way to obtain  $\text{Var}(X|W_1)$  is to subtract an estimate of the measurement error variance from the variance of the residuals from the regression of  $W_2$  on  $W_1$ . A method of moments estimate of  $\text{Var}(U)$  is given by  $\widehat{\text{Var}}(U) = \frac{\widehat{\text{Var}}(W_1 - W_2)}{2}$ .

The method of moments approach to estimating  $\text{Var}(X|W_1)$  can sometimes give negative variance estimates in practice. An alternative approach to fitting the measurement error model, which can be parameterised to ensure non-negative variance estimates, is to assume that  $(W_1, W_2)|M$  and  $(X, W_1)|M$  are jointly normally distributed and use maximum likelihood to obtain parameter estimates, where  $M$  is a set of variables we wish to condition upon,

$$W_1, W_2|M \sim N \left( \begin{pmatrix} E(X|M) \\ E(X|M) \end{pmatrix}, \begin{pmatrix} \text{Var}(X|M) + \text{Var}(U) & \text{Var}(X|M) \\ \text{Var}(X|M) & \text{Var}(X|M) + \text{Var}(U) \end{pmatrix} \right) \quad (5.4)$$

$$X, W_1|M \sim N \left( \begin{pmatrix} E(X|M) \\ E(X|M) \end{pmatrix}, \begin{pmatrix} \text{Var}(X|M) & \text{Var}(X|M) \\ \text{Var}(X|M) & \text{Var}(X|M) + \text{Var}(U) \end{pmatrix} \right). \quad (5.5)$$

## 5. Performance of correction methods

---

It can be shown that  $X|(W_1, M)$  is then normal with mean and variance

$$E(X|W_1, M) = \frac{W_1 \text{Var}(X|M) + E(X|M) \text{Var}(U)}{\text{Var}(X|M) + \text{Var}(U)} \quad (5.6)$$

$$\text{Var}(X|W_1, M) = \frac{\text{Var}(X|M) \text{Var}(U)}{\text{Var}(X|M) + \text{Var}(U)}. \quad (5.7)$$

$E(X|M)$ ,  $\text{Var}(X|M)$ , and  $\text{Var}(U)$  can be estimated by maximum likelihood using 5.4, where we let  $E(X|M) = \gamma_0 + \gamma_M^T M$ . To ensure that  $\text{Var}(X|M)$ , and  $\text{Var}(U)$  are positive we can use the parameterisation [184],

$$\text{Var}(X|M) + \text{Var}(U) = \exp(\theta_1) \quad (5.8)$$

$$\text{Var}(X|M) = \exp(\theta_1) \frac{\exp(\theta_2)}{\exp(\theta_2) + 1} \quad (5.9)$$

which ensures that the variance parameters are positive. Maximum likelihood procedures naturally extend to the situation where we have confounders  $\mathbf{Z}$ , by conditioning on them appropriately. Method of moments and maximum likelihood approaches will typically give similar results.

### Natarajan's method (RC1)

Under Natarajan's method [194] we fit the regression calibration model

$$X = \alpha_0 + \alpha_1 W_1 \quad \text{or} \quad W_2 = \alpha_0 + \alpha_1 W_1 \quad (5.10)$$

depending on whether we have true exposure measurements or repeat observations for a subset of individuals. We can use this regression calibration model to form  $\tilde{X} = \alpha_0 + \alpha_1 W_1$ . The values  $\tilde{X}$  are dichotomised according to the cutpoint  $C$  to obtain the binary variable  $\tilde{X}_C = I(\tilde{X} > C)$  which is used in the disease model

$$g(E(Y|\tilde{X}_C)) = \beta_0^{rc1} + \beta_1^{rc1} \tilde{X}_C. \quad (5.11)$$

Under scenario (2b) we only require one of  $W_2$  or  $W_3$  to have been observed, unlike for the methods that follow. In the simulation that follows we fit the model by regressing the mean of  $W_2$  and  $W_3$  on  $W_1$ .

### Regression calibration based method (RC2)

As discussed in chapter 4 the assumption of non-differential error is not valid when a continuous exposure subject to classical measurement error is dichotomised. However, we shall consider what happens if we do make the non-differential error assumption. In this case we

have

$$g(E(Y|W_{1C})) = \beta_0^{rc2} + \beta_1^{rc2} E(X_C|W_{1C}).$$

To fit this model we need to estimate  $E(X_C|W_{1C})$ . When we have observed the true exposure  $X$  in a validation substudy we can obtain  $E(X_C|W_{1C})$  directly by regressing  $X_{1C}$  on  $W_{1C}$  using logistic regression. Otherwise, if we assume that  $X$  and  $W_1$  are jointly normally distributed then

$$E(X_C|W_{1C} = 1) = P(X > C|W_1 > C) \quad (5.12)$$

$$= \frac{P(\alpha_0 + \alpha_1 W_1 + e > C, W_1 > C)}{P(W_1 > C)} \quad (5.13)$$

$$= \frac{\int_C^\infty \left( 1 - \Phi \left( \frac{C - E(X|W_1)}{\sqrt{\text{Var}(X|W_1)}} \right) \right) \phi \left( \frac{w - E(W)}{\sqrt{\text{Var}(W)}} \right) dw}{1 - \Phi \left( \frac{C - E(W)}{\sqrt{\text{Var}(W)}} \right)} \quad (5.14)$$

where  $\phi(\cdot)$  and  $\Phi(\cdot)$  are the probability and cumulative density functions of the standard normal distribution respectively. Under scenario (1b) we estimate  $E(X|W_1)$  and  $\text{Var}(X|W_1)$  via maximum likelihood using the methods described above. Under scenario (2b) we assume that  $W_2, W_3|W_1$  is normally distributed and proceed similarly via maximum likelihood.

### Moment Reconstruction (MR)

MR aims to create  $X^{MR}(W, Y)$  that shares its first two joint moments with the first two joint moments of  $X$  with  $Y$ . It has been shown that [184, 202],

$$X^{MR}(W, Y) = E(X|Y) + G(W - E(W|Y))$$

where  $G = \sqrt{\frac{\text{Var}(X|Y)}{\text{Var}(W_1|Y)}}$ . We propose that the values  $X^{MR}(W, Y)$  are dichotomised according to the cutpoint  $C$  to obtain the binary variable  $X_C^{MR} = I(X^{MR}(W, Y) > C)$ . We then use  $X_C^{MR}$  in our disease model

$$g(E(Y|X_C^{MR})) = \beta_0^{MR} + \beta_1^{MR} X_C^{MR}.$$

Under scenarios (1a) and (2a) we can estimate  $E(X|Y)$  and  $\text{Var}(X|Y)$  by regressing  $X$  on  $Y$  and under scenarios (1b) and (2b) we use maximum likelihood.

### Multiple Imputation (MI)

When the true exposure has been observed in the validation substudy we can replace  $X$  in the disease model by imputed values from the model

$$X = \gamma_0 + \gamma_1 W_1 + \gamma_2 Y + \epsilon.$$

## 5. Performance of correction methods

---

Imputed measurements for  $X$  are given by

$$X^{MI}(W_1, Y) = E(X|W_1, Y) + e$$

where  $e$  is a random draw from the residuals  $\epsilon$ . This procedure is repeated to give  $K$  multiply imputed datasets. We propose forming  $X_C^{MI(k)}$  by dichotomising  $X^{MI(k)}$  for  $k = 1 \dots K$  and fitting the disease model

$$g(Y|X^{MI(k)}) = \beta_0^{(k)} + \beta_1^{(k)} X_C^{MI(k)}$$

for each of the  $K$  imputed datasets. The parameter estimates obtained from the  $K$  models can be combined using Rubin's rules [181] which were described in section 4.3.5.

The imputation model can be fitted directly under scenarios (1a) and (2a). Under scenario (1b) we estimate  $E(X|W_1, Y)$  and  $\text{Var}(X|W_1, Y)$  via maximum likelihood for those individuals who were not in the validation substudy; for those in the validation substudy we allow for the fact that we have an additional mismeasured exposure measurement by imputing from the model where we have additionally conditioned on the observed  $W_2$ . Under scenario (2b) we proceed similarly via maximum likelihood assuming that  $W_2, W_3|W_1, Y$  is normally distributed.

### Group-SIMEX (gSIMEX)

For each of several values of  $\zeta$  we generate  $B$  pseudo datasets

$$W_b^*(\zeta) = W_1 + \sqrt{\zeta} \sigma_u Z_b, \quad b = 1, \dots, B$$

where  $Z_b$  is a random standard normal variate.  $B$  is typically chosen to be between 100 and 200. We then fit the disease model

$$g(E(Y|W_{C,b}^*(\zeta))) = \beta_0^{gS(b)} + \beta_1^{gS(b)} W_{C,b}^*(\zeta)$$

where  $W_{C,b}^*(\zeta) = I(W_b^*(\zeta) > C)$ , to each pseudo dataset noting that when  $\zeta = 0$  we have the naïve model of equation 5.3. The average parameter estimates obtained from the set of values for  $\zeta$  are then extrapolated back to  $\zeta = -1$ , which corresponds to no measurement error. Suitable extrapolants include the rational linear and quadratic.

When  $W_1$  is subject to systematic measurement error (scenarios (2a), (2b)) we use the approach of Alpizar *et al.* [179]. They suggest performing an initial calibration step to create a new exposure variable,  $W_1^*$ , which is free from systematic error, where

$$W_1^* = \frac{W_1 - \alpha_0}{\alpha_1}. \quad (5.15)$$

gSIMEX is then applied using  $W_1^*$  as our 'observed' exposure. The gSIMEX method requires

an estimate of  $\sigma_u^2$  as an input into the model, however, when the error is systematic we have to adjust the measurement error variance to take account of the transformation of the observed values in equation 5.15, so  $\sigma_{u^*}^2 = \frac{\sigma_u^2}{\alpha_1^2}$  is used in the gSIMEX procedure. Under scenarios (1a), (1b), (2a) we can estimate  $\sigma_u^2$  by maximum likelihood. Under scenario (2b) we can either estimate  $\sigma_u^2$  by assuming a trivariate normal distribution for  $(W_1, W_2, W_3)$ , or use a method of moments estimate.

In appendix F we provide R code for fitting MI, MR, and gSIMEX models. We also include code for plotting the SIMEX extrapolation function.

## 5.2.2 Simulation procedure

For a sample of 5,000 individuals the true continuous exposure,  $X$ , was generated from a normal distribution with mean 0, and variance 1. The observed exposure,  $W_1$  was generated using equation 5.2 with  $(\alpha_0 = 0, \alpha_1 = 1)$  for scenarios (1a) and (1b) and  $(\alpha_0 = 0, \alpha_1 = 0.5)$  for scenarios (2a) and (2b).  $U_1$  was generated from a normal distribution with mean 0 and variance  $\sigma_u^2$ , for values  $\sigma_u^2 = 0.25, 1$ .

For scenarios (1a) and (2a) we assumed that the true exposure  $X$  was observed in a random subset of 10% or 50% of individuals. Under scenarios (1b) and (2b) we assume that  $X$  is completely unobserved. We generated two additional exposures  $W_2$  and  $W_3$  from the classical measurement error model where the errors  $U_2, U_3$  are independently distributed normal with means 0, variance  $\sigma_{U_3}^2 = \sigma_{U_2}^2 = \sigma_{U_1}^2$ , and were generated independent of  $U_1, X$ , and  $Y$ . The additional exposure measurements were assumed to be observed in a random subset of 10% or 50% of individuals. Under scenario (1b) only  $W_2$  was used in the correction methods, under scenario (2b) we used both  $W_2$  and  $W_3$ .

Dichotomised true and observed exposures,  $X_C$  and  $W_C$ , were formed using the fixed cutpoints  $C = 0, 1$ . The outcome for each individual,  $Y$ , was generated according to the logistic model  $P(Y = 1) = (1 + e^{-(\beta_0 + \beta_1 X)})^{-1}$ , for values  $\beta_1 = \log(1.5), \log(2)$  and  $\beta_0 = -2.5$ . This results in event rates ( $Y = 1$ ) of approximately 8% and 9% for  $\beta_1 = \log(1.5)$  and  $\beta_1 = \log(2)$  respectively. We conducted 1,000 simulations for each value of  $\beta_1, \sigma_u^2$  and cutpoint,  $C$ .

We applied each of the methods described in section 5.2.1 under each of the four scenarios. For the MI based method we used 90 or 50 imputed datasets for when our validation dataset contained 10% and 50% of individuals respectively. This follows the rule of thumb that the number of multiply imputed datasets should correspond to the percentage of missing data [203]. For the gSIMEX procedure we used the quadratic extrapolant and formed 100 pseudo datasets for each value of the scaling parameter  $\zeta = \{0, 0.5, 1, 1.5, 2\}$  for each simulation. For all methods, where we have observed the true exposure  $X$  for a subset of the population (scenarios (1a), (2a)) the true dichotomised exposure  $X_C$  was used in place of fitted or imputed values when they are known.

## 5. Performance of correction methods

---

The parameter of interest in this simulation is the estimate obtained from each of the modelling methods of the log-odds ratio parameter  $\beta_1$  in equation 5.1, the coefficient of the indicator that  $X > C$ .

### 5.2.3 Results

Simulation results for each scenario are given in tables 5.1–5.4. In each table we can see the expected attenuation when we perform the naïve analysis on the dichotomised observed exposure; the attenuation being more severe when  $\sigma_{u1}^2 = 1$ , and under scenarios (1b) and (2b) when the observed exposure is subject to systematic error. Method RC1 under corrects for measurement error under all scenarios. Under scenario (1b) and (2b) we observe that this method gives the naïve estimate when  $C = 0$ ; we provide an explanation for this in the discussion. Method RC2 provides relatively unbiased estimates under scenarios (1a) and (2a) when we have observed  $X$  for 10% of individuals, and shows even smaller bias when  $X$  is observed for 50% of individuals. Under scenarios (1b) and (2b) the estimates show severe upward bias.

MI and MR both perform extremely well, giving almost unbiased estimates under all scenarios; even when the size of the validation substudy is only 10%. gSIMEX gives attenuated parameter estimates, with more severe bias when the measurement error is large. This might be expected as we are extrapolating further and therefore amplifying any error in the fitted extrapolation function. gSIMEX also does not take advantage of the additional information in the repeat measurements, other than through an improved estimate of the measurement error variance. We also considered MCSIMEX (results not shown), but this also performed poorly because it makes the assumption of non-differential misclassification between exposure groups.

We repeated the simulation for a normally distributed continuous outcome  $Y$  using linear regression. The results from this simulation can be found in appendix B and are similar to those found for the logistic model.

### 5.2.4 Discussion

We have seen that MI and MR work extremely well in estimating the true parameter value  $\beta_1$ ; this is because these methods allow for non-differential error. Freedman *et al.* [202] obtained similar results in their comparison of methods when the exposure was not dichotomised. In the case where a continuous exposure subject to non-differential error is not dichotomised, regression calibration can be used. Regression calibration is already widely used in practice because of its simplicity, and is preferred because it is more efficient than MR and MI in this case [202].

Regression calibration makes the assumption of non-differential measurement error and we have seen the consequences of erroneously making this assumption in our simulation i.e. method RC2. Although our observed continuous exposure is subject to non-differential er-

ror, the dichotomised observed exposure is subject to differential measurement error. MI and MR perform well because they allow for differential error.

Natarajan's regression calibration based method did not perform well; this is because the approach is flawed. Natarajan's method only scales the exposure measurements and does not change the ordering of individuals with respect to their exposure measurements. Consider applying Natarajan's method to a linear model, when  $W$  is subject to classical measurement error,  $C = 0$ , and the true exposure  $X$  is observed in a subset of  $n_1$  individuals. We can ignore what regression calibration does when  $C = 0$  because the expected value of  $\hat{\alpha}_0$  is 0, and  $\alpha_1$  preserves the sign of  $W$  since  $\lambda \in (0, 1]$ . In this situation Natarajan's method gives

$$\begin{aligned}\beta_1^{rc1} &= \frac{\text{Cov}(\tilde{X}_C, Y)}{\text{Var}(\tilde{X}_C)} \\ &= \frac{n_1 \text{Cov}(I(X > 0), Y) + (n - n_1) \text{Cov}(I(W > 0), Y)}{\text{Var}(\tilde{X}_C)} \\ &= \frac{n_1 \text{Cov}(I(X > 0), Y)}{\text{Var}(\tilde{X}_C)} + \frac{(n - n_1) \text{Cov}(I(W > 0), Y)}{\text{Var}(\tilde{X}_C)}\end{aligned}$$

Note that  $\text{Var}(I(W > 0)) = \text{Var}(I(X > 0))$  so,

$$\begin{aligned}\beta_1^{rc1} &= \frac{n_1 \text{Cov}(I(X > 0), Y)}{n \text{Var}(I(X > 0))} + \frac{n_1 \text{Cov}(I(W > 0), Y)}{n \text{Var}(I(W > 0))} \\ &= \frac{n_1}{n} \beta_1 + \frac{n - n_1}{n} \beta_1^*\end{aligned}\tag{5.16}$$

i.e. we get a weighted average of the true and naïve estimates. When we only observe replicate measurements then  $n_1 = 0$  and we obtain the naïve estimate, as we saw in our simulations.

We initially included the cutpoint  $C = 2$  in our simulation because this cutpoint was additionally used by Natarajan. A large proportion of our simulations failed for Natarajan's method using this cutpoint because  $\max(\tilde{X}) < 2$ . When  $\sigma_u^2 = 1$ , the variance of the fitted values from the regression calibration model will be 0.5. Therefore, the probability that an imputed value from the regression model exceeding the cutpoint  $C = 2$  is only 0.002. Natarajan considered a scenario where the regression calibration model was estimated using external data. Under this scenario we impute all observations from the regression calibration model. This means that in Natarajan's simulation study, with a simulated study size of 500 individuals, there is a 31% chance of each simulated dataset not containing a single observation above the cutpoint. Natarajan, however, makes no mention of simulation failures in her paper.

Under MI we omitted the step of sampling the parameter values and instead used the point estimates of the parameters  $(\alpha_0, \alpha_1, \alpha_2, \theta_1)$  under scenarios (1b) and (2b). These parameters will be well estimated for the size of validation substudy we considered; almost all the uncertainty will be in the residuals. This approach reduced the computation involved in the imputation

## 5. Performance of correction methods

---

procedure considerably. We did perform analyses where we sampled the parameters assuming a multivariate normal distribution, using the inverse of the observed information matrix for  $(\alpha_0, \alpha_1, \alpha_2, \theta_1)$  as our estimate of the variance-covariance matrix. This approach only changed the results in the third decimal place by no more than 0.005 in either direction.

Given that we have now considered the somewhat simpler case of logistic regression, we could now potentially investigate there application to the Cox model. Under the Cox model, instead of a simple outcome  $Y$ , our outcome would consist of an event indicator  $e_i$  and a survival time  $t_i$ . The two regression calibration methods and gSIMEX extend naturally to the case where we have a survival outcome, since the outcome is not used in the measurement error model. MI and MR both include the outcome in the measurement error model. White and Royston [183] suggest that the event indicator, as well as the estimated cumulative baseline hazard,  $\hat{H}_0(t)$ , should be included in imputation models for missing data in a Cox model. This approach is probably also appropriate for MR. We leave this investigation as future work.

In this simulation we considered dichotomising the underlying continuous exposure. Each of the methods extend naturally to the situation where we categorise the underlying continuous exposure into more than two groups. We extended the simulation to consider the performance of MR and MI in the case that we are interested in estimating the parameters of a grouped exposure analysis using quintiles of the exposure. We performed the simulation using only these two methods because they were the only methods that provided unbiased parameter estimates when the exposure was dichotomised. The results of this simulation are in appendix B and show that MI and MR still provide almost unbiased estimates of the parameter estimates when we use more than two exposure groups.

Grouped exposure analyses are often used to avoid assuming a shape for the exposure–disease relationship, and to assess non-linearity. We have seen in this section that MI and MR perform well when the exposure–disease relationship is linear. These methods may therefore be good candidates for extension to the more general case where the exposure–disease relationship is non-linear, however, we leave this for future work. MR matches only the first two moments of the data and therefore may not perform well for non-linear relationships. It may also be worthwhile to consider Mean Adjusted Imputation [55] which matches higher moments of the data, and therefore allows greater flexibility.

In conclusion, we have observed that gSIMEX and regression calibration methods perform poorly at estimating a true linear exposure–disease relationship under a grouped exposure analysis. MR and MI perform well and should be preferred if continuous exposures subject to classical measurement error are to be used in grouped exposure analyses of linear exposure–disease relationships, and are candidates to be considered for when the relationship is non-linear.

$\sigma_u^2$	$c$		Using $X_C$	Naive	RC1		RC2		MI		MR		gSIMEX	
					10%	50%	10%	50%	10%	50%	10%	50%	10%	50%
$\beta = \log(1.5), (\alpha_0 = 0, \alpha_1 = 1)$														
0.25	0	Mean	0.650	0.580	0.587	0.615	0.662	0.648	0.647	0.648	0.647	0.649	0.622	0.622
		SD	0.109	0.105	0.107	0.109	0.356	0.157	0.156	0.109	0.162	0.118	0.157	0.157
1	0	Mean	0.707	0.614	0.655	0.676	0.685	0.699	0.703	0.700	0.701	0.701	0.643	0.645
		SD	0.120	0.119	0.126	0.124	0.390	0.170	0.168	0.120	0.178	0.129	0.161	0.165
1	1	Mean	0.650	0.457	0.474	0.552	0.662	0.648	0.646	0.645	0.647	0.647	0.488	0.487
		SD	0.109	0.106	0.107	0.109	0.356	0.157	0.213	0.119	0.228	0.131	0.156	0.160
1	1	Mean	0.707	0.473	0.570	0.631	0.685	0.699	0.700	0.699	0.698	0.698	0.488	0.487
		SD	0.120	0.110	0.150	0.137	0.390	0.170	0.223	0.128	0.243	0.138	0.154	0.153
$\beta = \log(2), (\alpha_0 = 0, \alpha_1 = 1)$														
0.25	0	Mean	1.106	0.982	0.994	1.043	1.133	1.105	1.100	1.103	1.100	1.104	1.054	1.054
		SD	0.112	0.107	0.110	0.111	0.363	0.160	0.155	0.112	0.165	0.120	0.160	0.161
1	0	Mean	1.169	1.019	1.080	1.118	1.167	1.165	1.163	1.163	1.160	1.163	1.070	1.071
		SD	0.110	0.107	0.112	0.112	0.336	0.152	0.153	0.110	0.161	0.117	0.150	0.154
1	1	Mean	1.106	0.766	0.798	0.930	1.133	1.105	1.098	1.098	1.096	1.100	0.819	0.818
		SD	0.112	0.105	0.107	0.111	0.363	0.160	0.205	0.120	0.226	0.132	0.157	0.159
1	1	Mean	1.169	0.785	0.941	1.044	1.167	1.165	1.162	1.161	1.155	1.161	0.816	0.815
		SD	0.110	0.102	0.131	0.120	0.336	0.152	0.204	0.116	0.219	0.127	0.147	0.146

Table 5.1: Mean and standard deviation (SD) of estimates of the log odds ratio in a logistic regression of a binary outcome  $Y$  on  $X_C$  across 1000 simulated data sets using the true exposure, the naïve method, and different correction methods when the true exposure is measured in 10% or 50% of the study population. (Scenario (1a))

$\sigma_u^2$	$c$	Using $X_C$	Naive	RC1		RC2		MI		MR		gSIMEX		
				10%	50%	10%	50%	10%	50%	10%	50%	10%	50%	
$\beta = \log(1.5), (\alpha_0 = 0, \alpha_1 = 1)$														
0.25	0	Mean	0.650	0.580	0.580	0.580	0.824	0.824	0.649	0.645	0.649	0.649	0.621	0.622
		SD	0.109	0.105	0.107	0.106	0.150	0.150	0.112	0.100	0.117	0.115	0.158	0.157
	1	Mean	0.707	0.614	0.648	0.646	0.982	0.980	0.703	0.698	0.704	0.703	0.644	0.644
		SD	0.120	0.119	0.131	0.130	0.192	0.191	0.121	0.111	0.131	0.129	0.163	0.162
1	0	Mean	0.650	0.457	0.456	0.457	0.918	0.916	0.651	0.641	0.651	0.651	0.487	0.487
		SD	0.109	0.106	0.108	0.106	0.217	0.214	0.194	0.119	0.161	0.134	0.156	0.160
	1	Mean	0.707	0.473	0.552	0.547	1.251	1.239	0.702	0.691	0.702	0.700	0.488	0.488
		SD	0.120	0.110	0.166	0.163	0.316	0.295	0.205	0.126	0.174	0.144	0.153	0.152
$\beta = \log(2), (\alpha_0 = 0, \alpha_1 = 1)$														
0.25	0	Mean	1.106	0.982	0.981	0.982	1.394	1.394	1.104	1.096	1.105	1.104	1.052	1.054
		SD	0.112	0.107	0.108	0.107	0.154	0.152	0.113	0.103	0.122	0.119	0.162	0.161
	1	Mean	1.169	1.019	1.069	1.068	1.630	1.628	1.166	1.157	1.166	1.165	1.070	1.070
		SD	0.110	0.107	0.117	0.115	0.177	0.173	0.114	0.101	0.120	0.117	0.152	0.151
1	0	Mean	1.106	0.766	0.767	0.767	1.540	1.537	1.109	1.088	1.109	1.103	0.816	0.817
		SD	0.112	0.105	0.106	0.105	0.221	0.212	0.197	0.119	0.164	0.137	0.157	0.160
	1	Mean	1.169	0.785	0.914	0.911	2.074	2.056	1.171	1.148	1.170	1.165	0.816	0.815
		SD	0.110	0.102	0.146	0.140	0.329	0.283	0.197	0.118	0.165	0.132	0.148	0.145

Table 5.2: Mean and standard deviation (SD) of estimates of the log odds ratio in a logistic regression of a binary outcome  $Y$  on  $X_C$  across 1000 simulated data sets using the true exposure, the naïve method, and different correction methods when  $W_2$  is assumed to be observed in 10% or 50% of the study population. (Scenario (1b))

$\sigma_u^2$	$c$	Using $X_C$	Naive	RC1		RC2		MI		MR		gSIMEX		
				10%	50%	10%	50%	10%	50%	10%	50%	10%	50%	
$\beta = \log(1.5), (\alpha_0 = 0, \alpha_1 = 0.5)$														
0.25	0	Mean	0.650	0.457	0.474	0.552	0.662	0.648	0.646	0.645	0.647	0.647	0.566	0.564
		SD	0.109	0.106	0.107	0.109	0.356	0.157	0.213	0.119	0.228	0.131	0.188	0.192
	1	Mean	0.707	0.548	0.570	0.631	0.685	0.699	0.700	0.699	0.698	0.698	0.615	0.613
		SD	0.120	0.164	0.150	0.137	0.390	0.170	0.223	0.128	0.243	0.138	0.221	0.223
1	0	Mean	0.650	0.287	0.321	0.464	0.662	0.648	0.647	0.644	0.645	0.645	0.319	0.317
		SD	0.109	0.104	0.105	0.108	0.356	0.157	0.256	0.128	0.296	0.143	0.167	0.166
	1	Mean	0.707	0.304	0.520	0.616	0.685	0.699	0.697	0.698	0.693	0.694	0.324	0.321
		SD	0.120	0.118	0.272	0.155	0.390	0.170	0.266	0.136	0.308	0.152	0.170	0.171
$\beta = \log(2), (\alpha_0 = 0, \alpha_1 = 0.5)$														
0.25	0	Mean	1.106	0.766	0.798	0.930	1.133	1.105	1.098	1.098	1.096	1.100	0.949	0.948
		SD	0.112	0.105	0.107	0.111	0.363	0.160	0.205	0.120	0.226	0.132	0.191	0.193
	1	Mean	1.169	0.910	0.941	1.044	1.167	1.165	1.162	1.161	1.155	1.161	1.023	1.021
		SD	0.110	0.139	0.131	0.120	0.336	0.152	0.204	0.116	0.219	0.127	0.198	0.201
1	0	Mean	1.106	0.479	0.536	0.777	1.133	1.105	1.100	1.097	1.096	1.099	0.531	0.530
		SD	0.112	0.101	0.101	0.107	0.363	0.160	0.246	0.128	0.289	0.144	0.161	0.162
	1	Mean	1.169	0.509	0.856	1.006	1.167	1.165	1.161	1.161	1.152	1.158	0.539	0.534
		SD	0.110	0.109	0.224	0.134	0.336	0.152	0.243	0.124	0.281	0.140	0.158	0.159

Table 5.3: Mean and standard deviation (SD) of estimates of the log odds ratio in a logistic regression of a binary outcome  $Y$  on  $X_C$  across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when  $X$  is assumed to have been observed in 10% or 50% of the study population. (Scenario (2a))

$\sigma_u^2$	$c$		Using $X_C$	Naive	RC1		RC2		MI		MR		gSIMEX	
					10%	50%	10%	50%	10%	50%	10%	50%	10%	50%
$\beta = \log(1.5), (\alpha_0 = 0, \alpha_1 = 0.5)$														
0.25	0	Mean	0.650	0.457	0.456	0.457	0.913	0.914	0.647	0.642	0.650	0.648	0.583	0.576
		SD	0.109	0.106	0.107	0.106	0.215	0.213	0.226	0.120	0.149	0.127	0.192	0.191
	1	Mean	0.707	0.548	0.550	0.546	1.009	1.008	0.698	0.694	0.702	0.698	0.726	0.734
		SD	0.120	0.164	0.163	0.162	0.308	0.303	0.237	0.130	0.157	0.138	0.342	0.330
1	0	Mean	0.650	0.287	0.287	0.287	0.982	0.975	0.648	0.636	0.651	0.648	0.397	0.395
		SD	0.109	0.104	0.105	0.103	0.378	0.357	0.318	0.147	0.228	0.150	0.207	0.204
	1	Mean	0.707	0.304	-0.057*	0.333	1.331	1.313	0.696	0.686	0.702	0.697	0.401	0.404
		SD	0.120	0.118	2.507*	0.963	0.564	0.517	0.332	0.156	0.240	0.160	0.206	0.208
$\beta = \log(2), (\alpha_0 = 0, \alpha_1 = 0.5)$														
0.25	0	Mean	1.106	0.766	0.766	0.767	1.533	1.533	1.098	1.092	1.104	1.101	0.974	0.970
		SD	0.112	0.105	0.106	0.105	0.218	0.211	0.225	0.121	0.152	0.131	0.195	0.194
	1	Mean	1.169	0.910	0.911	0.910	1.676	1.675	1.159	1.153	1.166	1.162	1.218	1.223
		SD	0.110	0.139	0.141	0.140	0.273	0.259	0.222	0.118	0.149	0.125	0.290	0.278
1	0	Mean	1.106	0.479	0.480	0.479	1.639	1.627	1.101	1.083	1.109	1.100	0.662	0.661
		SD	0.112	0.101	0.101	0.101	0.397	0.349	0.317	0.143	0.234	0.152	0.199	0.198
	1	Mean	1.169	0.509	0.394*	0.700	2.224	2.196	1.160	1.142	1.170	1.160	0.666	0.669
		SD	0.110	0.109	2.221*	0.385	0.585	0.484	0.315	0.142	0.231	0.148	0.194	0.193

Table 5.4: Mean and standard deviation (SD) of estimates of the log odds ratio in a logistic regression of a binary outcome  $Y$  on  $X_C$  across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when  $W_2, W_3$  are assumed to be observed in 10% or 50% of the study population. \*Based on 995 datasets since  $\max(X_C) < C$  for five datasets. (Scenario (2b))

---

## 5.3 Simulation Study 3: Comparison of structural fractional polynomials & P-splines

To compare the performance of structural fractional polynomials, which we proposed in chapter 4, versus Carroll *et al.*'s [157] structural P-splines, we again extended the simulation study of chapter 3. We start with a brief recap of the two methods and consider measures for evaluating their performance. We then describe the simulation, present the results, and finish with a brief discussion of how the two methods compare.

### 5.3.1 Methods

#### Structural fractional polynomials and structural P-splines

Structural fractional polynomial and P-spline models result from the application of regression calibration to fractional polynomial and P-spline models respectively. These models involve taking the expectation of exponents of the true exposure conditional on the observed. To estimate these expectations we have to make distributional assumptions about the conditional distribution.

A structural fractional polynomial model of degree  $m$  for the Cox model is given by

$$\log h(t|W) \approx \log h_0(t) + \sum_{j=1}^m \beta_j E(H_j(X)|W)$$

where  $H_j(X)$  is as defined in equation 2.8.

A structural P-spline model, with power spline basis, is given by

$$\log h(t|W) \approx \log h_0(t) + \sum_{j=1}^p \beta_j E(X^j|W) + \sum_{j=1}^l \beta_{p+j} E((X - t_j)_+^p|W).$$

where  $(X - t_j)_+^p$  is as defined in equation 2.10.

We assume throughout this section that  $X|W$  is normally distributed. We estimate the mean and variance of this distribution using the method of moments, as described in section 5.2.1.

#### Gauss quadrature

We can calculate all the expectations for structural P-spline models using analytical formulae (as we show in appendix A). We also showed that we can obtain explicit formulae for the expectations required to fit fractional polynomial models when the distribution of true exposure given observed is log-normal. However, when we assume normality of the distribution of true

## 5. Performance of correction methods

---

exposure given observed then we cannot easily find analytical expressions for all of the expectations we require for fractional polynomial models. In this case we can use Gauss-Hermite quadrature to approximate the expectation. If  $X|W$  follows a standard normal distribution then

$$E(f(X)|W) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2} f(x) dx \approx \sum_{k=1}^m w_k f(x_k)$$

for  $m$  appropriately chosen quadrature points  $\{x_k, k = 1, \dots, m\}$  and weights  $\{w_k, k = 1, \dots, m\}$ . The approximation is exact if  $f(x)$  is a polynomial of order no greater than  $2m - 1$ . For non-polynomial functions we can achieve better approximations to the expectation by increasing the number of quadrature points.

### Methods for comparing estimated shapes for the exposure–disease relationship

When comparing correction methods for non-linear exposure–disease relationships we require both global and local measures of performance. This is because a good global fit may belie regions of exposure where the fit is poor. These regions of poor fit, such as in the tails of the exposure distribution, are often of most interest.

The root mean squared error (rMSE) of the estimated hazard ratio across the exposure range is

$$\text{rMSE}(x_0) = \sqrt{\frac{1}{n} \sum_{i=1}^n \left( \log \hat{\text{HR}}(x_i; x_0) - \log \text{HR}(x_i; x_0) \right)^2}$$

where  $\text{HR}(x; x_0)$  is the true hazard ratio for an exposure  $x$  relative to the reference value  $x_0$ , and  $\hat{\text{HR}}(x; x_0)$  is an estimate of this hazard. The rMSE is dependent on the choice of reference value  $x_0$ —different choices of reference value will give different values for the rMSE. The mean of the exposure is usually chosen as the reference value in practice, and will typically give a smaller value for the rMSE than other choices. This is because a large proportion of the exposure values are usually located about the mean.

We propose a measure of the average squared error that does not depend on the choice of reference value, by summing over all possible choices of the reference value; we call this the reference free root mean squared error (RFrMSE). We define the RFrMSE as

$$\begin{aligned}
\text{RFrMSE} &= \sqrt{\frac{1}{n^2} \sum_{j=1}^n \sum_{i=1}^n \left( \log \hat{\text{HR}}(x_i; x_j) - \log \text{HR}(x_i; x_j) \right)^2} \\
&= \sqrt{\frac{1}{n^2} \sum_{j=1}^n \sum_{i=1}^n \left( (\log \hat{\text{HR}}(x_i; x_0) - \log \hat{\text{HR}}(x_j; x_0)) - (\log \text{HR}(x_i; x_0) - \log \text{HR}(x_j; x_0)) \right)^2} \\
&= \sqrt{\frac{1}{n^2} \sum_{j=1}^n \sum_{i=1}^n \left( (\log \hat{\text{HR}}(x_i; x_0) - \log \text{HR}(x_i; x_0)) - (\log \hat{\text{HR}}(x_j; x_0) - \log \text{HR}(x_j; x_0)) \right)^2}.
\end{aligned}$$

We see from the final line of the equation above that RFrMSE is straightforward to calculate, as it can be calculated from the errors relative to a chosen reference value.

We also propose considering the gradient of the fitted function as an alternative method of comparison; this is a local measure of performance. The gradient of the fitted function does not depend on the choice of reference value because the baseline hazard is independent of the exposure, and therefore disappears upon differentiation with respect to the exposure. For fractional polynomials the derivatives are easy to compute analytically. For P-splines they are easy to compute if we are using a power spline basis, and if we use a B-spline basis then we can use the formula given by De Boor [98]. The derivatives of a B-spline basis are easily obtained in R using the `spline.des` function.

Govindarajulu *et al.* [109, 123] compared the performance of fractional polynomials and P-splines. We considered using the performance measure they described in their study. They proposed calculating an approximation to the mean integrated absolute error. This was done by splitting the exposure range into a number of equally sized intervals, and within each interval they constructed two rectangles. Both rectangles had width equal to the length of the exposure interval, and height equal to the hazard ratio at the right hand point of the interval for the estimated and true curves respectively. They then found the absolute difference between the areas of the rectangles for the estimated and true values, for each interval. The sum of these across the grid of exposure values is a crude approximation to the area between the curves. This measure is equivalent to the mean absolute error across a regular grid of exposure values multiplied by  $mw$ , where  $m$  is the number of rectangles over which the measure is calculated, and  $w$  the width of each rectangle.

Govindarajulu *et al.* also proposed weighting this measure, where the weights are the normalised inverse variances of the difference between the two curves. These variances have to be estimated using bootstrap resampling which is computationally intensive, especially when the model is a structural fractional polynomial or P-spline. This weighting scheme is dependent on the curves being considered, and makes comparison across curves difficult. We therefore do not consider this performance measure in our simulation study.

### 5.3.2 Simulation procedure

The data for this simulation were generated as in chapter 3. We additionally generated a replicate exposure measurement,  $W_{i2}^*$ , for 10% of individuals where  $W_{i2}^* = X_i + U_{i2}$  if individual  $i$  is in the validation substudy and is missing otherwise.  $U_{i2}$  was generated from a normal distribution with mean 0, variance  $\sigma_u^2$ , and was generated independently of  $X_i$ , and the measurement error  $U_i$ . A substudy size of 10% reflects what may typically be available in practice.

Sometimes epidemiological investigations find no exposure–disease relationship. We therefore investigate the null relationship given by

$$h(t|X) = \log(h_0(t)). \quad (5.17)$$

Under this model  $\rho = 0.01$  was found to give an event rate of approximately 10%. We include this relationship to check that our correction methods do not artificially create a relationship when in fact there is none.

For each level of measurement error, each shape for the exposure–disease relationship, and each modelling method we calculate the estimated shape of the exposure–disease relationship; the rMSE which we evaluate using the estimated true exposure from the regression calibration model; the gradient at each exposure value; and the RFrMSE which we evaluate at the percentiles of the estimated true exposure from the regression calibration model. We use FP2 models and P-splines constrained to have 4 degrees of freedom, as in chapter 3.

### 5.3.3 Results

Figures 5.2 and 5.3 show the mean exposure–disease relationship obtained from structural fractional polynomial and P-spline analyses. Both methods perform well at recovering the true exposure–disease relationship except for the two threshold relationships.

Structural fractional polynomials recover the relationship that would have been observed in the absence of measurement error for the two threshold relationships, although they are poor fits to the true exposure–disease relationship in each case. The structural P-spline method almost recovers the relationship that would have been observed in the absence of measurement error for the threshold relationship, although, the hazard is increasingly biased away from the null relationship below the mean. For the non-linear threshold shaped relationship, structural P-splines are unable to recover the relationship that would have been observed in the absence of measurement error, with over correction for higher levels of exposure; the threshold increasingly moves away from the true value.

The reason for the over correction can be seen in figure 3.3. Between exposure levels of 10–10.75 units the observed relationship is biased away from, and not towards unity. The bias in this region changes little with increasing measurement error, contrary to what we would

---

expect; presumably because of the rate at which those above the threshold, are being ‘mixed’ with those below the threshold, due to measurement error. When we apply the structural P-spline model, we expect to see an increasingly steeper gradient when we correct for increasing degrees of measurement error over this range. Above the threshold the observed relationship is attenuated, and as a result the difference between the corrected curves remains relatively constant upon correction.

Graphing the gradient of the average exposure–disease relationship allows us to compare the performance of the two methods without relying on a reference value, although it is hard to relate what we observe in the plot of the gradient, to the undifferentiated relationship. Any range of exposure over which the gradient of the fitted function is not the true gradient, produces an incorrect shape for the exposure–disease relationship. In figures 5.2 and 5.3 we saw that the structural correction methods performed well, and hence the gradients of the curves obtained from the simulation closely follow the true gradients, except for the two threshold relationships. We therefore only give plots for these two shapes in figure 5.4.

There is a stark contrast between the gradients of the two correction methods for both shapes. The structural fractional polynomial models are essentially quadratic and hence the gradients are linear. The gradient of the structural fractional polynomial models does not resemble the true gradient of the relationship. The structural P-splines perform better at picking out the non-linearities in the gradient, but because P-splines are composed of polynomial functions they are unable to recreate the sharp turns in the gradient of the true relationship. We might expect both P-splines and fractional polynomials to perform badly when the exposure–disease relationship is not suitably differentiable.

Figure 5.5 shows the rMSE under both modelling methods for each shape of the exposure–disease relationship. We see that the rMSE is consistently smaller and less variable for structural fractional polynomial models, compared with the structural P-spline models, except under the threshold relationships. Under the threshold relationships P-splines perform better when there is no measurement error, or the measurement error variance is small ( $RDR=4/5$ ). When the measurement error variance is larger, structural P-splines perform worse, and are considerably more variable than the structural fractional polynomial models. Surprisingly, there is little change in performance as the measurement error increases, except for under the threshold models.

The RFrMSE produced trends that are similar to those for rMSE, and can be seen in figure B.1 of appendix B. This suggests that the choice of reference value does not have a large impact on our results.

Table 5.5 shows the power to detect non-linearity under both correction methods for each of the shapes of relationship. Comparing table 5.5, with table 3.3 from chapter 3, we see that there is no gain in power to detect non-linearity associated with applying measurement error

## 5. Performance of correction methods

Method	RDR	Shape of exposure–disease relationship							
		Null	Linear	Threshold	U-shaped	J-shaped	Increasing quadratic	Asymptotic	Non-linear threshold
Structural	4/5	0.01	0.01	0.42	0.53	0.51	0.51	0.23	0.41
fractional	2/3	0.01	0.01	0.25	0.33	0.32	0.28	0.13	0.23
polynomial	1/2	0.01	0.01	0.09	0.12	0.14	0.09	0.08	0.08
Structural	4/5	0.06	0.08	0.62	0.68	0.67	0.67	0.41	0.70
P-spline	2/3	0.04	0.04	0.45	0.52	0.53	0.54	0.17	0.52
	1/2	0.04	0.05	0.25	0.31	0.30	0.35	0.14	0.30

Table 5.5: Estimated power to detect non-linearity under each shape for the exposure–disease relationship, using structural fractional polynomial and P-spline analyses.

correction. This is to be expected, as correcting for the effects of classical measurement error gives us no more information about the shape of the exposure–disease relationship.

### 5.3.4 Discussion

We have seen how structural fractional polynomials perform better in terms of rMSE than structural P-splines. Both methods, however, had a slight tendency to over correct. The methods were unable to recreate either of the true threshold relationships. This is due to the inability of the models to pick up the threshold when there is no measurement error, rather than a failure of the correction methods. The methods’ performance did not deteriorate significantly when we considered larger values for the measurement error variance, except under the threshold relationships. Under the non-linear threshold relationship, the structural P-spline method severely over corrected. We attribute this behaviour to the ‘softening’ of the threshold caused by measurement error.

Under the threshold relationships, both measurement error and the modelling method used can both lead to the threshold being obscured or lost completely. Correction for measurement error does not allow us to recover features of the relationship that have been lost. In chapter 3, we saw that our ability to detect non-linear relationships is severely hampered by exposure measurement error, and we have seen in this chapter that measurement error correction does not allow us to regain this power.

The performance of structural fractional polynomials and P-splines may vary with the degree of non-linearity—this is an area for further research. In this simulation study we assumed that the distribution of the observed exposure given the true is normally distributed, with constant variance. In practice this assumption may be violated, and we investigate the impact of erroneously assuming normality in chapter 7.

---

## 5.4 Conclusion

In this chapter we have considered the performance of many of the methods for correcting for the effects of exposure measurement error that were introduced in chapter 4.

In simulation study 1 we showed that MacMahon's method provides us with plots of the exposure–disease relationship that are similar, but not identical, to the true shape of the exposure–disease relationship, except under the threshold relationship. MacMahon's method reduces the range over which we view the exposure–disease relationship. We showed theoretically that MacMahon's method does not provide the correct shape for the true exposure–disease relationship unless the relationship is linear.

In simulation study 2 we considered methods for correcting a grouped exposure analysis of a linear exposure–disease relationship for the effects of both random and systematic measurement error, when the underlying exposure was continuous. We found that group-SIMEX, Natarajan's method, and another regression calibration based method gave biased estimates of the true parameter values. Moment reconstruction and multiple imputation gave unbiased estimates even when the validation substudy was small, and both when the measurement error was random and systematic. These methods are therefore candidates to take forward to consider when the exposure–disease relationship is non-linear—we leave this for future work.

In simulation study 3 we compared structural fractional polynomials and structural P-splines. We considered different methods of comparing the performance of the two approaches both locally and globally, and without dependence on the arbitrary choice of reference value. Structural fractional polynomials performed better than structural P-splines in terms of rMSE, but neither method performed well at recreating the true threshold relationships.

In chapter 6 we apply some of the methods evaluated in this chapter to data on the relationship between fasting blood glucose and coronary heart disease, and in chapter 7 we extend simulation study 3 to consider non-classical measurement error.

## 5. Performance of correction methods

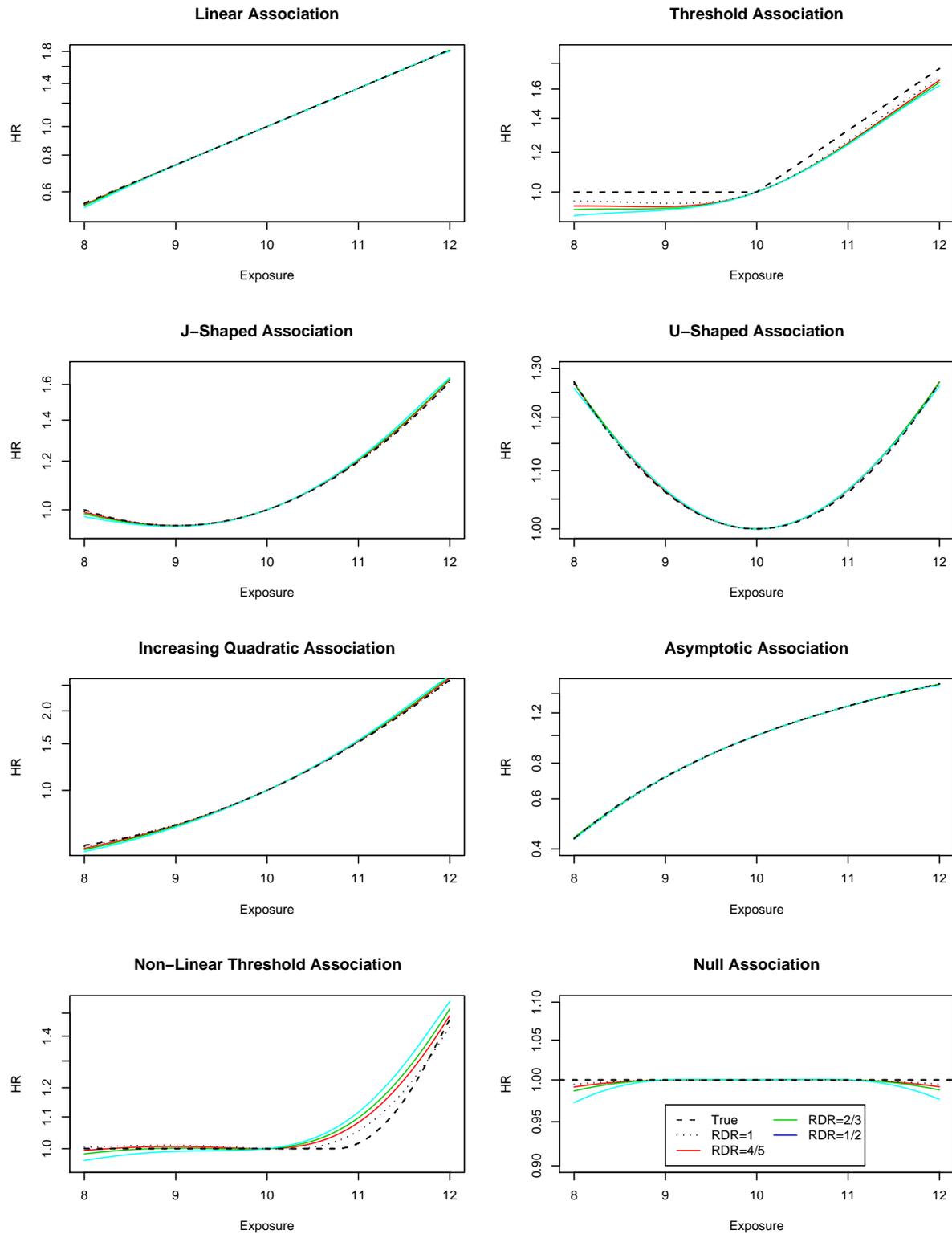


Figure 5.2: Structural P-spline analysis showing the measurement error corrected exposure-disease relationship.

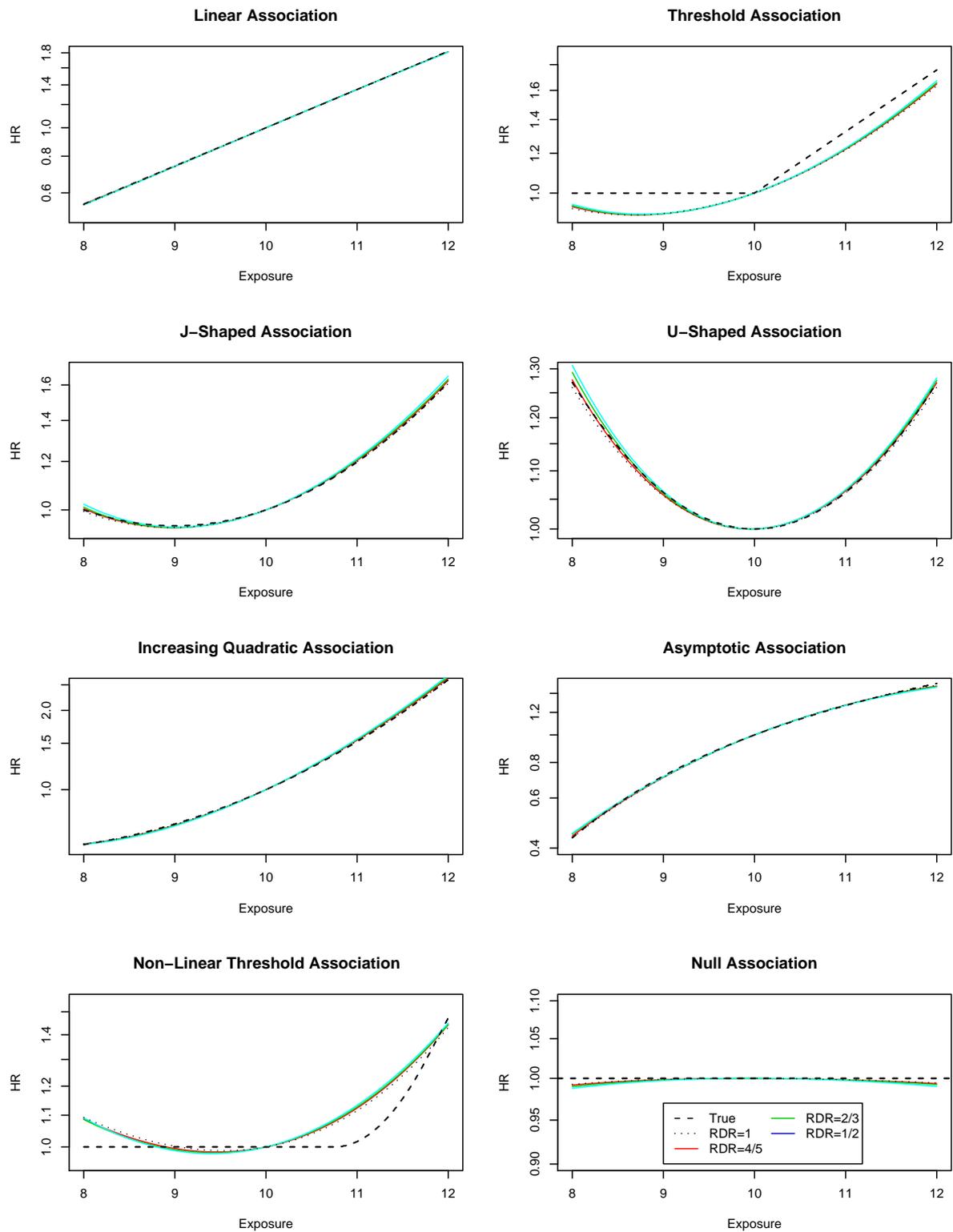


Figure 5.3: Structural fractional polynomial analysis showing the measurement error corrected exposure–disease relationship.

## 5. Performance of correction methods

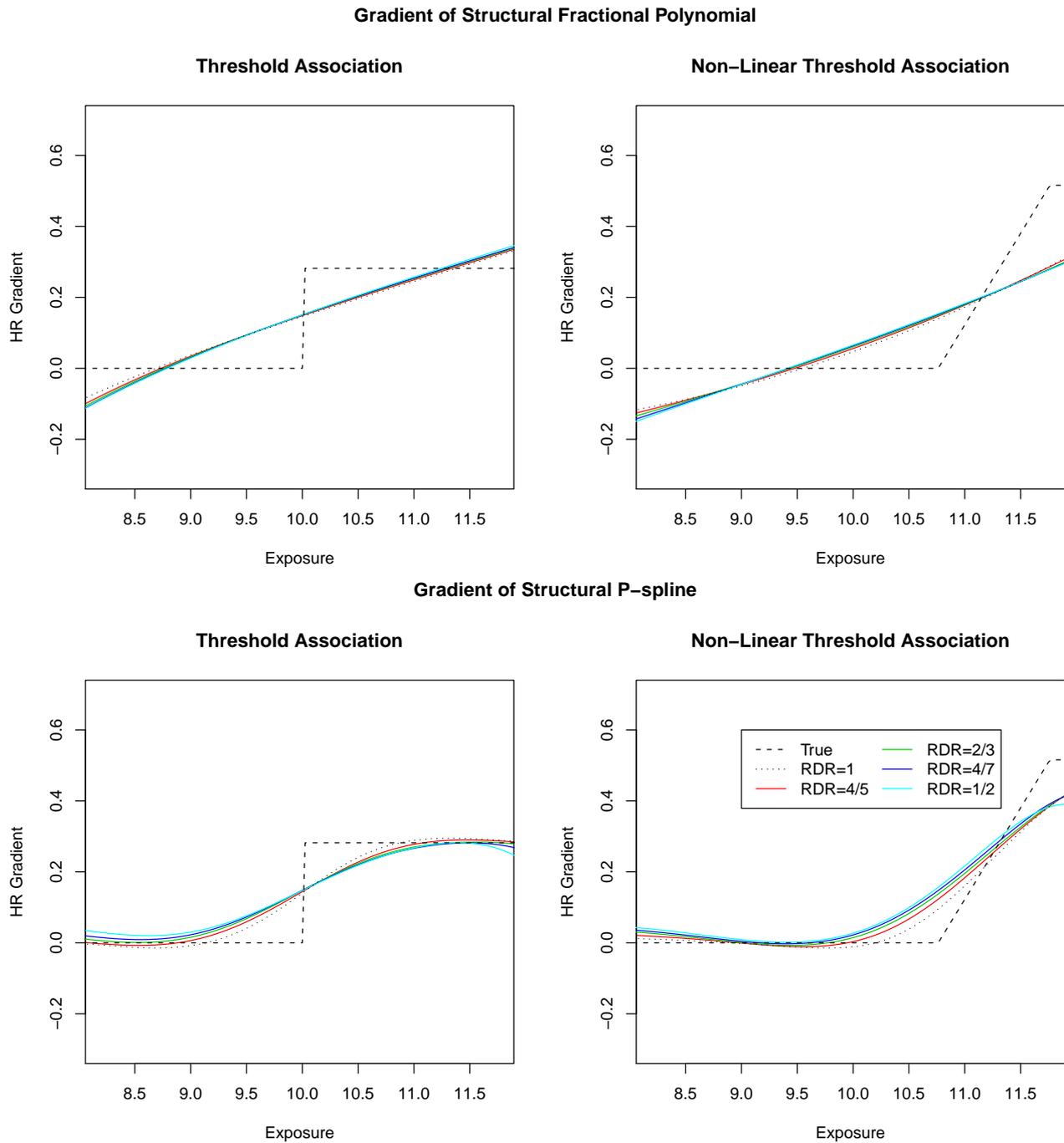


Figure 5.4: Gradient of structural fractional polynomial and P-spline models for measurement error corrected threshold exposure–disease relationship.

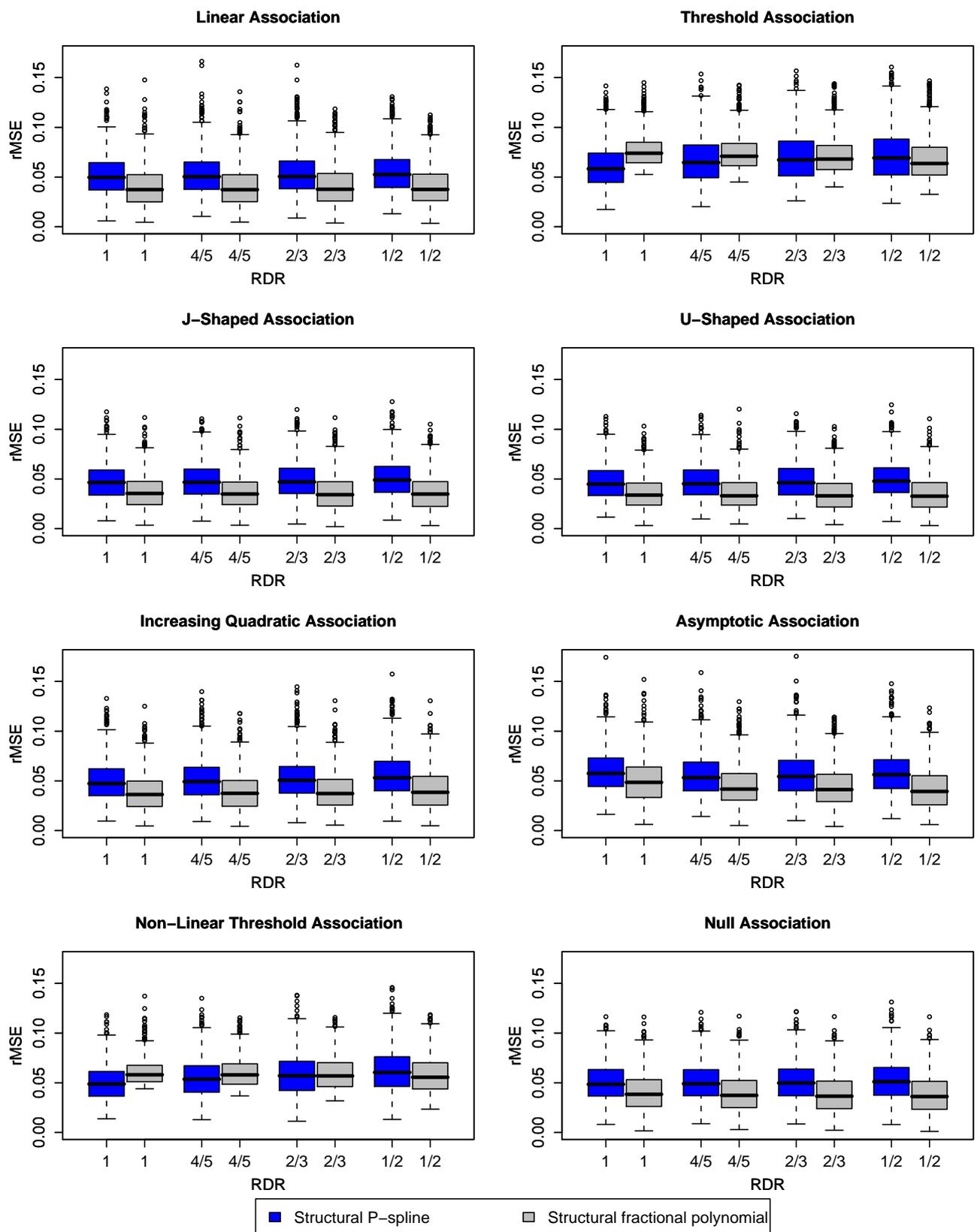


Figure 5.5: rMSE for structural fractional polynomial and P-spline simulations (with the mean exposure as reference value).



## Chapter 6

# Application — Emerging Risk Factors Collaboration

This chapter focuses on the motivating dataset for this thesis—the Emerging Risk Factors Collaboration (ERFC). We firstly describe the ERFC, before focusing on the relationship between fasting blood glucose (FBG) and the risk of experiencing a coronary heart disease (CHD) event. We provide background on what is known about the FBG–CHD relationship, the role of confounding variables, and provide a summary of the main features of the ERFC FBG dataset. We firstly perform grouped exposure, fractional polynomial and P-spline analyses using the observed FBG measurements. We then consider the measurement error in the observed FBG measurements before applying measurement error correction methods: MacMahon’s method, SIMEX, structural fractional polynomials and P-splines. We also discuss the differences between the observed and measurement error corrected relationships. In this chapter we do not allow for heterogeneity in the shape of the FBG–CHD relationship between studies; we return to this in chapter 8. The issues arising out of the analysis of the ERFC FBG data in this chapter shall provide motivation for chapter 7.

### 6.1 ERFC — Emerging Risk Factors Collaboration

As described in chapter 1 the ERFC [7] is a large individual participant data (IPD) meta analysis of over 1.1 million predominantly white Western individuals focusing on risk factors for CHD. There are over 69,000 incidents within the 11.7 million person years at risk. Individuals are not censored if they are known to have undergone cardiovascular investigations since this is not widely recorded amongst the participating studies. The studies are all prospective cohort studies, however, a small number of studies are randomised controlled trials or have supplied data from nested case-control studies. Full details of the collaboration can be found in the protocol paper for the collaboration [7] with further information available at <http://ceu.phpc.cam.ac.uk/research/erfc/>. We give a summary of some of

## 6. Application — Emerging Risk Factors Collaboration

---

the key features here.

The ERFC has obtained IPD from studies that had data available on: baseline measurements of at least one of the markers relevant to the ERFC's investigations; at least 1 year of follow-up; participants that were not selected on the basis of having previous cardiovascular diseases; and information on cause-specific mortality and/or major cardiovascular morbidity collected during the follow-up period. Studies were identified for inclusion primarily from previous meta-analyses, and additionally through literature searches, reference lists, and correspondence with authors of relevant reports. The teams looking at each risk factor have also performed checks to ensure that they have all available data. The data obtained from each study were checked for consistency and definitions of variables were harmonised across studies. For example, some studies may have categorised smoking into current smokers, past smokers and never smokers whereas others may have had categories of current smoker and non-current smoker. Any problems found with data obtained from each study were referred back to the collaborating study. Each investigation uses a subset of the whole dataset as not all covariates are measured for all individuals in all studies.

The Asia Pacific Cohort Studies Collaboration (APCSC) [13], another large IPD meta analysis of risk factors for CHD with data on 237,468 participants from 17 cohort studies, has similar aims to the ERFC. The incidence of CHD events within the Asia Pacific region is much lower than amongst western individuals. Therefore, because the ERFC is much larger, and will have many more CHD events, we should be able to better characterise the risk factor-CHD relationship.

Since the ERFC has observations on so many individuals it gives us better power to characterise the shape of the risk factor-CHD relationship than each study individually; especially at the extremes of the risk factor distribution where typically each study will only have a small number of individuals. The ERFC data are analysed principally by Cox proportional hazards regression models stratified by sex, undertaken in each study separately. Estimates of the risk factor-CHD relationship are combined over studies using random-effects meta-analysis.

Some covariates used in the ERFC's analyses can be assumed to be measured precisely such as age, height and sex but others such as FBG, cholesterol or blood pressure may fluctuate within individuals over time and/or are difficult to measure precisely. As we have seen in chapter 3, analyses that use mismeasured exposures can produce biased results and hence it is essential to correct for this.

In chapter 4, we discussed how an extra source of information is required to quantify the degree of measurement error present. There are no exposures within the ERFC for which we have a gold standard measurement as well as the mismeasured one. However, as of August 2007, 53 cohorts had provided one or more repeat exposure measurements, for one or more risk factors, on 339,808 participants.

---

We now look at the FBG–CHD relationship. We chose to focus on this relationship for two reasons: FBG is known to be subject to substantial measurement error, and the relationship with CHD appears to be non-linear.

## 6.2 Background — FBG and CHD

Glucose is the main source of energy for the body’s cells. The amount of glucose in the blood fluctuates throughout the day peaking after meals, especially those rich in carbohydrates [204]. The body uses the hormone insulin to control blood glucose levels. Fasting causes the body to produce glucagon which increases plasma glucose, and in healthy individuals insulin is produced to rebalance the increased glucose levels. Therefore, fasting blood glucose (the amount of glucose in the blood after a period of fasting, usually >8 hours) can be used as a measure of an individual’s ability to produce insulin and will typically be in the range 4.4–6.1 mmol/l for non-diabetics.

Several assay methods exist for estimating fasting blood glucose (FBG) from a blood sample using either chemicals or enzymes. The ERFC FBG data consists of FBG measurements, however, there are alternate measures of blood glucose including measuring peak blood glucose concentration two hours post glucose load, and measuring levels of HbA1c (glycosylated haemoglobin) which reflects long term average blood glucose levels [205]. HbA1c, therefore may be a better measure of blood glucose for use in epidemiological studies in the future, but currently it is not standardised, nor routinely measured.

Sustained high blood glucose levels (hyperglycaemia) or low blood glucose levels (hypoglycaemia) can have serious health consequences. Long term hyperglycaemia can cause nephropathy [206], neuropathy, retinopathy as well as cardiopathy [207], whilst hypoglycaemia can cause impairment of cognitive function. Diabetes mellitus can cause high levels of blood glucose as a result of insufficient insulin production (type 1 diabetes) or insulin resistance (type 2 diabetes).

FBG is typically split into three categories of interest—normal FBG, impaired FBG (IFG) and diabetic FBG. Usually there is little attention paid to those with low levels of FBG, although some studies have suggested increased risk amongst these individuals [208]. Disagreement exists over the definition of IFG with the two major authorities, the World Health Organisation (WHO) [205] and the American Diabetes Association [209] giving different ranges.

	WHO	American Diabetes Association
IFG	6.1 mmol/l to <7 mmol/L	5.6 mmol/l to <7 mmol/l
Diabetic FBG	≥ 7 mmol/l	≥ 7 mmol/l

There is an increasing focus on the relationship between IFG and CHD because IFG is a symptom of metabolic syndrome which is thought to affect one in five people in the United States

[210]. Bjørnholt *et al.* [211] found that FBG in the upper normal range was an independent predictor of cardiovascular death in non-diabetic healthy middle-aged men and Levitan [212] produced a meta-analysis of 38 studies of healthy non-diabetics and found an increased risk of CHD amongst those with IFG. It was also attributed to higher risk of CHD by Barr *et al.* [213] Upon full adjustment for known confounders Sung *et al.* [214] found that CHD risk was only substantially raised for those with diabetic levels of FBG. The DECODE study [215] showed a J-shape relationship between FBG and cardiovascular mortality with risk only increasing significantly for individuals with diabetic levels of FBG. The APCSC concluded that there was greatly increased hazard for those with impaired and diabetic levels of FBG. They also found that low levels of FBG (<5 mmol/l) conferred protection from CHD events [40].

### 6.3 ERFC fasting blood glucose dataset

The aim of this analysis of the ERFC FBG data is to better characterise the FBG–CHD relationship. This is important because there may be identifiable groups of non-diabetic individuals within the population who could reduce their risk of experiencing a CHD event by modifying their usual FBG levels through preventative measures such as changes in diet and/or lifestyle. The ERFC FBG dataset consists of 49 studies containing 470,454 individuals of which FBG was measured, on at least one occasion, in 278,346 individuals. There are a total of 186,433 repeat measurements on 79,842 individuals in 21 studies. There are up to 7 repeat measurements per individual but during this analysis we will only be using the first repeat measurement for an individual where this has occurred within the first 10 years of follow-up. The first CHD event is the outcome of interest. Although post first CHD event follow-up is available in many studies, we do not use this information to inform us about the FBG–CHD relationship because the relationship may differ between those who have previously experienced a CHD event and those who have not. The purpose of our investigation is to look at FBG as a risk factor for CHD in the general population so we shall exclude all known diabetics from our analysis. Diabetics will be on treatment and/or have made lifestyle choices (healthy diet, low alcohol intake, moderate exercise), to manage their condition which is likely to make them atypical compared with the general population. This reduces the dataset to 260,818 individuals with 59,144 first repeats.

Studies with fewer than eleven events were removed from the analysis. None of these studies had repeat observations on any participants. Two observations (without repeat measurements) from the Reykjavik study have been excluded from the analysis due to their implausibly low values of FBG (0.39,0.44); we suggest that these result from a data entry error. One outlying observation from the New Brunswick study was excluded from analysis because the first observation of FBG was 11.93 whilst the first repeat was 3.66 which represents a shift of more than ten standard deviations within that study. We shall ignore repeat measurements in any studies with repeat measurements on less than 11 individuals.

---

In this chapter we shall be treating the data as if it were one large study, stratifying by cohort and trial arm (where appropriate) to allow for non-proportional baseline hazards between cohorts as a result of factors such as geographical variation in prevalence of CHD related events. We shall take account of the heterogeneity in the shape of the FBG–CHD relationship between studies using meta-analytic approaches in chapter 8.

We stratify our analyses by sex because the risk profile for CHD differs between men and women [216]. Women are at lower risk of suffering a CHD event pre-menopause. It has been suggested that there is a link between sex and diabetes, with increased risk of ischaemic heart disease mortality associated with diabetes among women compared with men [217, 218]. In the general population survival in males tends to be worse than females.

As discussed in chapter 2, it is important to adjust for the effect of confounders. In our analyses we shall consider the effect of age, smoking status (current vs. non-current smoker), total cholesterol, systolic blood pressure (SBP) and body mass index (BMI). Measurements of one or more of these covariates are not available for 8,427 participants and hence the analyses are based on data from 241,390 individuals. Therefore, the final dataset for the analyses that follow consists of observations on 241,390 individuals with 58,568 first repeats who suffered 12,815 events during over 3 million person years of follow-up. Table 6.3 gives a summary of the ERFC FBG data used in the analyses. This is not the same dataset that was used in the papers published by the ERFC involving FBG [9, 10], but instead an earlier version. A table of study acronyms used in this chapter is given in appendix C.

The first panel of figure 6.1 shows a histogram of FBG which suggests that the distribution of FBG is non-normal with a positive skew (3.99) and is highly leptokurtic (excess kurtosis 38.5). A log transformation seemed to improve upon this and hence all analyses were performed using log-FBG. For presentational purposes we shall be plotting untransformed FBG on a log-axis to account for the transformation. Although the log transformation improves the normality of FBG, the distribution of log-FBG still has skew 0.85 and excess kurtosis of 5.89.

In general, outliers should be identified and analysed to ascertain their origin as part of the analysis scheme. However, in many epidemiological investigations, especially those that use large datasets, it is often not practicable to carry out a detailed outlier analysis and the most extreme observations are clipped (e.g. 0.5% of both tails) to remove any outliers that may unduly influence the analysis. We should be wary about adopting this approach when the exposure is subject to measurement error as the extreme values may have arisen due to the effects of measurement error; especially if the measurement error is multiplicative. Clipping may be justified if we believe that some process other than measurement error is causing outlying values (e.g. errors in data entry); we do not clip the FBG data.

## 6. Application — Emerging Risk Factors Collaboration

Study	Individuals	Individuals with repeats	Events	Male	Current smokers	Age mean (sd)	FBG mean (sd)	Repeat FBG mean (sd)
ALLHAT	12288	6771	464	6402	4045	66.02 (7.51)	5.39 (1.26)	5.67 (1.50)
ARIC	12699	11910	638	5521	3309	54.29 (5.71)	5.48 (0.53)	5.78 (1.00)
BHS	1571	457	90	723	430	43.26 (16.46)	5.07 (0.78)	5.32 (0.77)
BRUN	789	735	51	386	195	57.50 (11.27)	5.43 (0.68)	5.56 (1.14)
BUPA	19110		942	19110	7503	47.37 (7.65)	5.34 (1.03)	
BWHHS	2976		52	0	330	68.50 (5.45)	5.83 (0.89)	
CASTEL	2157	1026	60	843	296	73.42 (5.16)	5.73 (1.08)	5.33 (1.22)
CHARL	686	6	138	414	2	54.26 (9.35)	5.31 (1.55)	6.70 (2.09)
CHS1	3357	2135	480	1247	398	72.30 (5.21)	5.53 (0.53)	5.25 (0.85)
CHS2	366	253	36	138	63	72.30 (5.25)	5.44 (0.65)	5.50 (1.65)
DUBBO	1852		239	755	284	68.31 (6.68)	4.98 (0.58)	
FIA	1637	2	344	1350	405	53.92 (7.22)	5.38 (0.84)	5.95 (0.35)
FINE_FIN	248		63	248	33	76.33 (4.72)	5.70 (0.79)	
GOH	1115		29	550	342	51.34 (7.97)	5.53 (0.86)	
GOTO43	752		26	752	230	50.00 (0.00)	4.56 (0.76)	
GOTOW	1407		143	0	577	46.67 (6.19)	4.07 (0.66)	
HELSINAG	358		38	93	33	78.74 (4.12)	5.68 (1.85)	
HOORN	2004		62	886	651	61.02 (7.26)	5.41 (0.53)	
KIHD	1965	792	366	1965	604	52.43 (5.33)	4.64 (0.72)	5.00 (0.91)
MALMO	31139		1948	21555	14173	45.35 (7.35)	4.92 (0.74)	
MATISS83	2413	1594	75	1132	717	51.26 (9.64)	5.06 (0.79)	5.12 (0.97)
MATISS87	1929	1108	42	863	423	52.18 (9.47)	5.10 (0.68)	4.95 (1.09)
MATISS93	1122		13	547	312	49.04 (9.27)	4.76 (0.68)	
MRFIT	12665	12428	760	12665	8039	46.86 (5.96)	5.52 (0.87)	5.45 (0.93)
NCS3	310		14	212	226	41.71 (4.42)	5.62 (0.76)	
NHANES3	5843		167	2816	908	52.09 (15.86)	5.50 (1.18)	
OSLO	1396		148	1396	870	42.76 (6.58)	5.79 (0.97)	
PARIS1	6810	6525	321	6810	4596	47.14 (1.97)	5.67 (0.75)	5.65 (0.68)
PRHHP	3364	2883	101	3364	1684	53.93 (6.27)	4.99 (0.57)	5.26 (1.05)
RANCHO	1704		206	698	234	68.31 (11.13)	5.46 (0.85)	
REYK	16450		3151	7854	7805	52.25 (8.54)	4.45 (0.63)	
SHS	2065	1797	182	893	826	55.47 (8.09)	5.63 (0.59)	6.12 (2.14)
TARFS	2391	1781	136	1120	792	46.16 (13.11)	4.87 (0.85)	5.23 (1.13)
ULSAM	1603	627	436	1603	1094	49.63 (0.64)	4.90 (0.56)	4.97 (0.93)
VHMPP	66257		644	32130	12219	48.79 (13.10)	5.02 (1.46)	
VITA	7225		22	3248	1945	51.34 (8.04)	4.56 (1.22)	
WHITE2	7424	5746	156	5148	1408	49.47 (6.02)	5.23 (0.66)	5.17 (0.99)
ZARAGOZA	1943		32	821	348	59.18 (11.49)	5.41 (0.58)	
Total*	241390	58568	12815	146258	78349	51.23 (11.68)	5.12 (1.11)	5.50 (1.12)

Table 6.1: Summary of the ERFC FBG data by study. \* Excludes repeat measurements for those studies where they were obtained for less than 11 individuals (shaded grey).

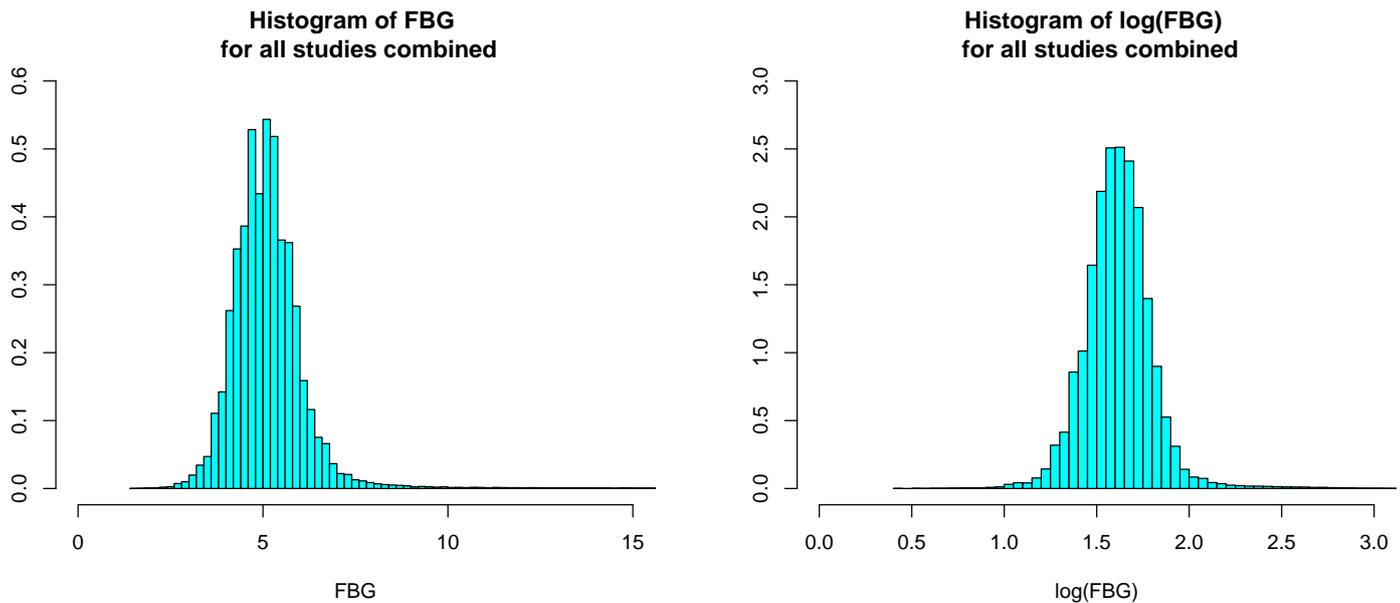


Figure 6.1: Histograms of FBG and log-FBG for all studies combined

## 6.4 Modelling — Ignoring the effects of measurement error

### 6.4.1 Grouped exposure analysis

In the ERFC, the main analyses of the shape of the FBG–CHD relationship are based on performing grouped exposure analyses [7]. We group the FBG data into nine groups,  $G$ , as in the ERFC FBG analysis

$$G = \{(0, 4], (4, 4.5], (4.5, 5], (5, 5.5], (5.5, 6], (6, 6.5], (6.5, 7], (7, 7.5], (7.5, \infty)\}$$

on the original scale. These groupings give a good balance between the number of groups and the number of individuals within groups. We then performed grouped exposure analyses using  $G$  as our predictor for FBG and adjusting for increasing levels of confounding: no adjustment for confounders (other than sex, cohort and trial arm through stratification), adjustment for age and smoking status; and adjustment for conventional cardiovascular risk factors at baseline (age, sex, BMI, smoking status, SBP and cholesterol). Each of the confounders was included as a linear term in the model.

Figure 6.2 shows a strong relationship between FBG and CHD which gets weaker upon adjustment for conventional cardiovascular risk factors. We show the relationship with respect to the third group as the reference category, and for each group points are plotted at the mean within-group exposure. Although the choice of reference category is arbitrary we have made this choice because it is the largest group. This choice will also aid comparison in the next section when we plot continuous methods of modelling the relationship, where it is usual to choose the mean exposure as the reference value. The confidence intervals in figure 6.2 use the

floating absolute risks that were introduced in chapter 2.

In the first plot of figure 6.2 we observe a strongly increasing relationship, with those individuals in the group with the highest levels of FBG exhibiting three times the risk of those in the group with the lowest levels of FBG. The relationship could be said to be a threshold with the threshold occurring between the third and fourth groups. When we adjust for age and smoking status, in the second plot of figure 6.2, we see that the FBG–CHD relationship has a similar shape; however the risk gradient is generally less steep with a much smaller difference in hazard between the top and bottom groups of FBG. In the third plot, where we have adjusted fully for confounding variables, there remains a considerable increase in risk associated with high levels of FBG. The shape of the relationship is now more J-shaped. It is of interest to note how the sixth group of FBG is now aligned with the fifth group suggesting that there is a large difference in the hazard amongst people in the lower and upper ends of the IFG range; although this could be an artefact of the choice of cutpoints. Figure 6.2 has shown that the effect of confounders on the FBG–CHD relationship is large, and since the confounders are widely accepted as independent predictors of CHD events, all analyses that follow will be fully adjusted for baseline levels of the conventional cardiovascular risk factors used in the third plot. To aid comparison of the FBG–CHD relationship across analyses we shall use the same axes in all plots of the FBG–CHD relationship that follow.

### 6.4.2 Fractional polynomial and P-spline models

In chapter 2 we described the shortcomings of using grouped exposure analyses. We now consider fractional polynomial and P-spline models. We fitted the best FP2, and a P-spline with four degrees of freedom to log-FBG; these are shown in figure 6.3. The fractional polynomial procedure chose an FP2 model with powers (0.5,1). The hazard across the exposure range is consistently greater under the P-spline model than that under the fractional polynomial model. The fractional polynomial model shows little increase in hazard below an FBG of 7 mmol/l, and above this the hazard only reaches 1.5 at an FBG level of 10 mmol/l. The P-spline model is more consistent with the trend observed in the grouped exposure analysis above (third panel of figure 6.2). This is probably because P-splines and grouped exposure analyses (to some degree) can better fit local features of the data. As we have seen in the histogram of log FBG in figure 6.1 most of the FBG observations are in the range 4–6 mmol/l. The fractional polynomial appears to be fitting this range well at the cost of infidelity in the tails of the FBG distribution. Both curves suggest the possibility of a slight upturn in the hazard at low levels of FBG; a moderate increase at levels consistent with IFG; and a greatly increased hazard at diabetic levels.

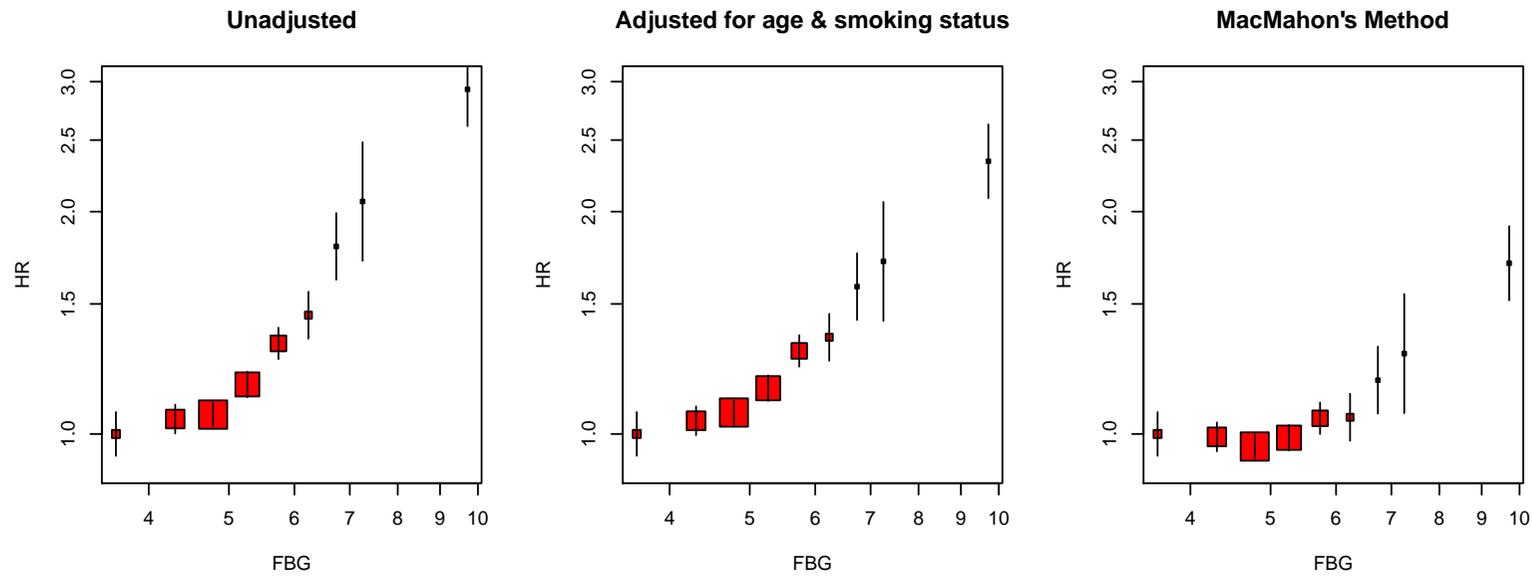


Figure 6.2: Grouped exposure analysis of the observed FBG–CHD relationship for increasing levels of adjustment for confounders. Points plotted at the mean within-group exposure, and 95% confidence intervals produced using the method of floating absolute risk.

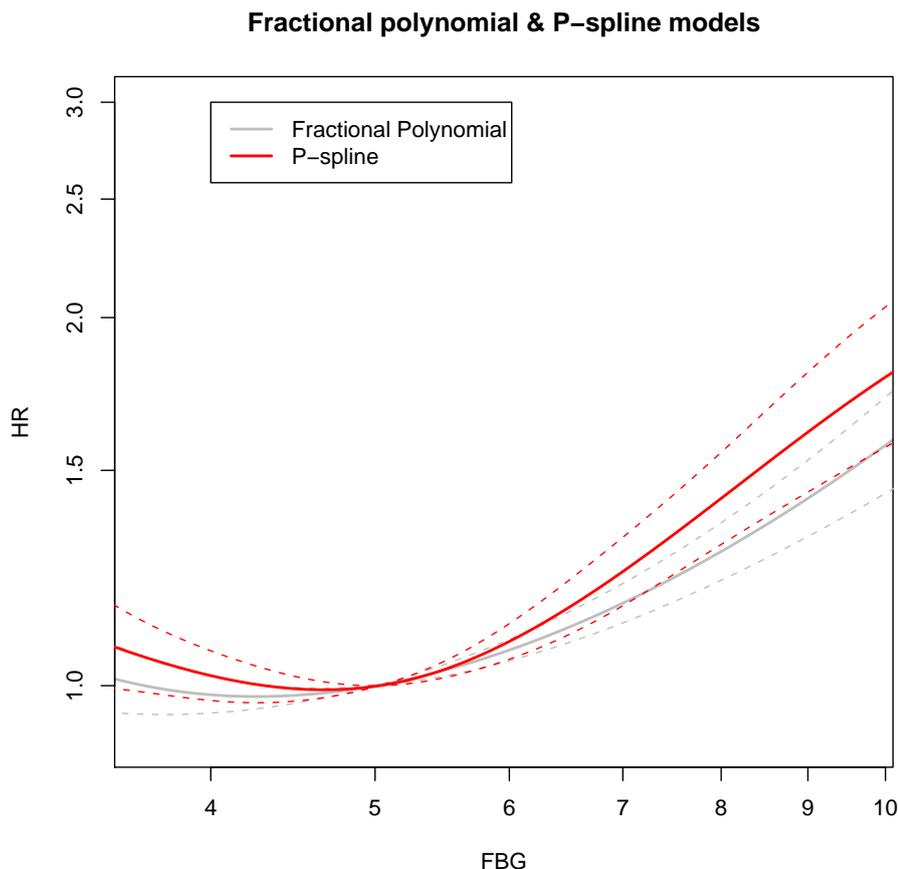


Figure 6.3: Fractional polynomial and P-spline models for the observed FBG–CHD relationship (with 95% confidence intervals given by dotted lines in the colour corresponding to each method).

## 6.5 Measurement error in FBG

Measurements of FBG are known to be subject to substantial measurement error which may result from a number of sources including: within-person variation, error from the assay procedure, and non-compliance with fasting. The APCSC found the RDR for measurements of FBG to be 0.6, whilst Rosner [28] found that within-person variability makes up nearly 50% of the total variance in FBG in the Framingham Heart Study. The analyses of section 6.4 fail to account for the measurement error in FBG and the effects it may have on the FBG–CHD relationship. In chapter 3 we saw that the effect of exposure measurement error is in general to attenuate the true relationship, hence the observed FBG–CHD relationship in section 6.4 will not reflect the true relationship.

A plot of RDRs for each study with repeats gives an indication of the variability in the amount of measurement error between studies. The plot may be used to highlight outlying studies that should be investigated to ascertain if there is an underlying reason as to why the amount of measurement error within that study is inconsistent with the other studies. Figure 6.4 shows

the RDR obtained from each study with greater than 11 repeats adjusted for baseline risk factors: age, sex, BMI, smoking status, SBP and cholesterol. There is a lot of variation in the RDRs across the 17 studies, which suggests that the approach taken by some authors of simply transporting an RDR from an external study may not be appropriate. However, the RDRs of the larger studies are in a similar range. The estimate of the mean RDR 0.56 (95% CI 0.51 - 0.61) is broadly consistent with those of the APCSC and Framingham Heart Study.

The variation in the RDR between studies could be caused by a number of factors including differences in the time between baseline and repeat measurements, and differences in unmeasured confounders, such as diet and exercise, between studies. The RDR is a function of not only the measurement error variance, but also the variance of the true exposure—the latter will almost certainly vary between studies.

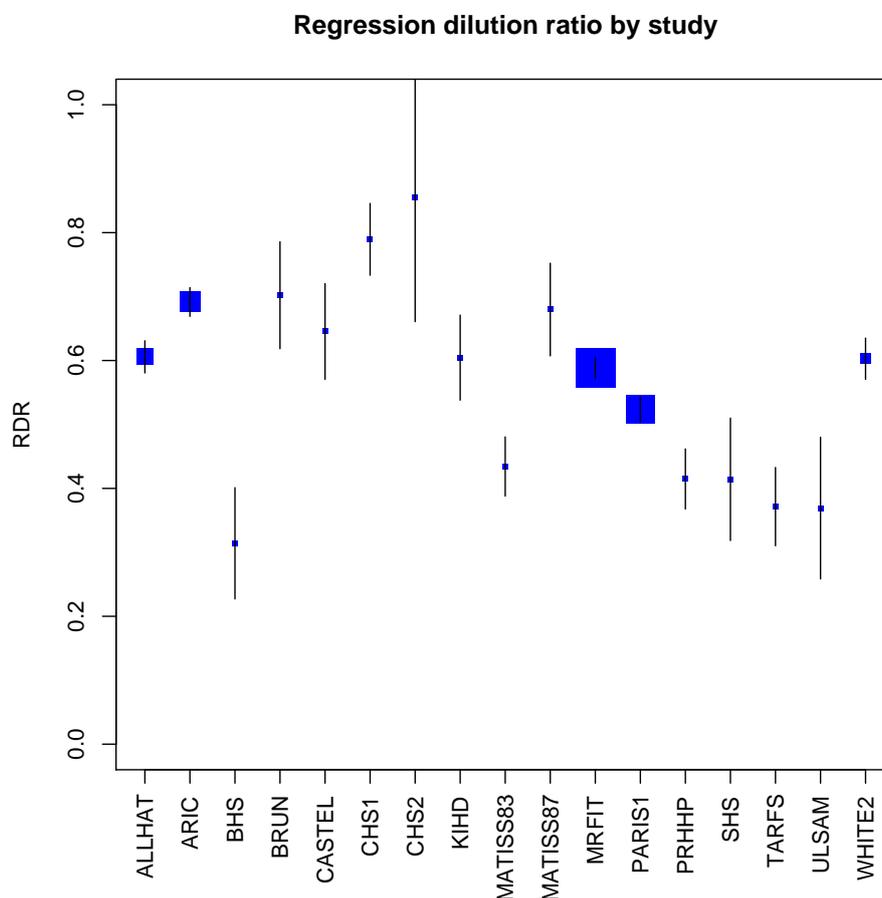


Figure 6.4: Estimated RDRs for FBG by study. Size of plotting symbol proportional to precision of RDR.

Most methods of correcting for the effects of exposure measurement error assume the classical measurement error model and often additionally assume that the true exposure given the observed is normally distributed. Carroll *et al.* [33] suggest plots that can be used to investigate the measurement error structure and whether these assumptions hold.

1. Produce a normal Q-Q plot of within-person differences. If the points fall on the Q-Q

## 6. Application — Emerging Risk Factors Collaboration

---

line then this suggests that the measurement error is normally distributed.

2. Plot the standard deviation of the observations on each individual against their mean. If the plot shows no obvious trends then this suggests that the measurement error variance is constant across the range of exposure.
3. Plot the standard deviation of the observations on each individual against each of the confounders. If the plot shows no obvious trends then this suggests that the measurement error variance does not depend on the value of confounding variables.

In this chapter we shall use these graphs to assess whether the classical measurement error holds, and whether the measurement error is normally distributed. In chapter 7 we shall talk more about the properties of these plots and how they can inform us about the structure of the measurement error in the case that the classical measurement error model does not hold.

Plots were produced for each of the cohorts with repeat measurements. Figure 6.5 provides an example from the PRHHP study and is representative of the plots obtained from the other sixteen studies with repeat measurements. The normal Q-Q plot shows a positive skew meaning that any assumptions of normality we make in modelling the measurement error may be invalid. The mean–variance plot of replicates strongly suggests that the measurement error variance is heteroscedastic with a higher variance amongst observations in the tails of the distribution of log-FBG, especially at high levels. The plot of the standard deviations for each individual against confounders suggest the within-person standard deviation does not appear to vary with the level of confounders. Note that there were no female participants in the PRHHP study. We shall ignore our reservations about the heteroscedasticity and non-normality of the measurement error in this chapter and we shall assume that the measurement error distribution is normal and homoscedastic (on the log scale). However, we will consider the possible impact of non-normality and heteroscedasticity of the measurement error in chapter 7.

As we saw in figure 6.4, the degree of measurement error varies widely between studies. For those studies where repeat measurements were available, we estimated the measurement error variance of log-FBG using a method of moments estimator. For those studies without repeats, we assumed that the measurement error variance was a sample size weighted average of the measurement error variance estimates obtained from those studies with repeat FBG measurements; this gave a value of 0.0069.

The average regression calibration model which we shall use for studies where repeat measurements are not available is:

	Intercept	log(FBG)	Smoking Status: Current	Age	SBP	BMI	Cholesterol	Sex: Female
Parameter Value	0.5299	0.5816	0.0053	0.0003	0.0004	0.0038	-0.0001	-0.0099

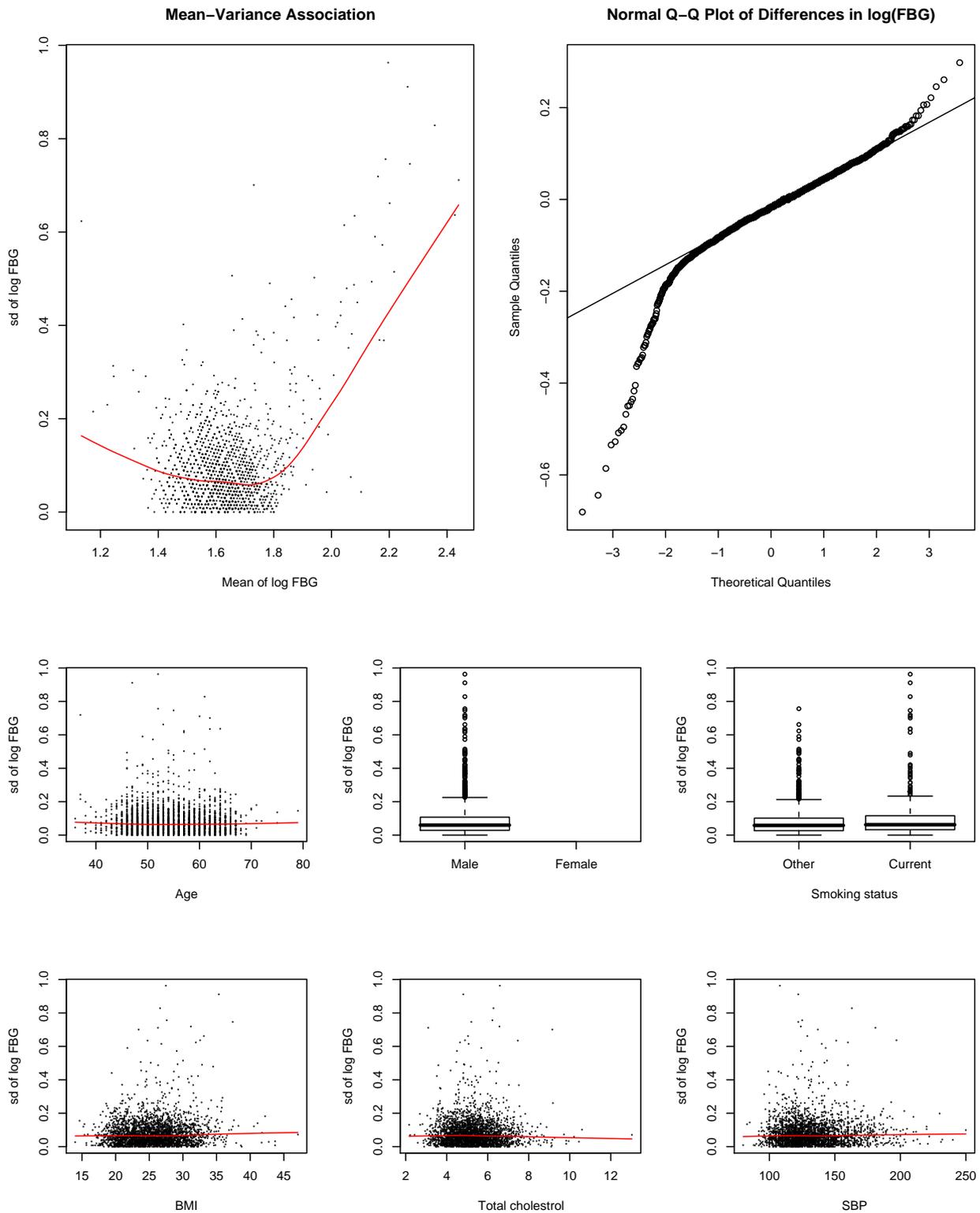


Figure 6.5: Graphs as proposed by Ruppert & Carroll for testing measurement error assumptions for studies with repeat measurements of FBG—PRHHP study. Note that there were no female participants in the PRHHP study.

As we use different regression calibration models for different studies the overall mean of usual log-FBG is not the same as that of observed log-FBG. We shall however continue to display all graphs with the mean of observed log-FBG as the reference value to aid comparison between the uncorrected and corrected analyses.

### 6.6 Correcting for measurement error

Firstly we shall consider correcting the grouped exposure analysis of section 6.4 using MacMahon's method, which although flawed, can give a good indication as to the shape of the exposure-disease relationship. We then use SIMEX and structural fractional polynomials and P-splines, to correct for the effects of measurement error in our continuous models for the FBG-CHD relationship.

#### 6.6.1 MacMahon's method

In figure 6.7 we have plotted the hazard ratios for each exposure group against the means of the repeat measurements in each baseline group. We can see that the means of the most extreme groups of FBG come in towards the centre of the distribution a long way from the unadjusted values and we therefore observe the relationship over a much smaller range. As we described in chapter 5, this is a graph of usual exposure in baseline groups of FBG versus hazard, rather than what we want which is hazard against groups of usual FBG.

Figure 6.6 compares the shape of the FBG-CHD relationship obtained without correction for the effects of measurement error, and using MacMahon's correction method. The group means have been connected to give an impression of the shape of the implied FBG-CHD relationship under each analysis. We can see that the shape of the corrected and uncorrected relationships appear to be very similar at lower levels of FBG. Only above an FBG of 6 mmol/l does the corrected relationship appear stronger than the uncorrected one, and then the difference appears to be rather modest, except for the highest exposure group.

#### 6.6.2 SIMEX

The SIMEX procedure was discussed in detail in section 4.3.4 and involves observing how parameter estimates change for increasing amounts of exposure measurement error and then extrapolating this trend back to the case of no measurement error. In this section we apply the SIMEX procedure to fractional polynomial and P-spline analyses.

We applied the SIMEX procedure to the P-spline model of section 6.4. When applying SIMEX to P-splines we need to ensure that we fit the disease model using the same basis for log-FBG in each dataset, otherwise the parameter estimates obtained for each pseudo-dataset will be incomparable. This can be achieved by extending the range over which the basis is defined,

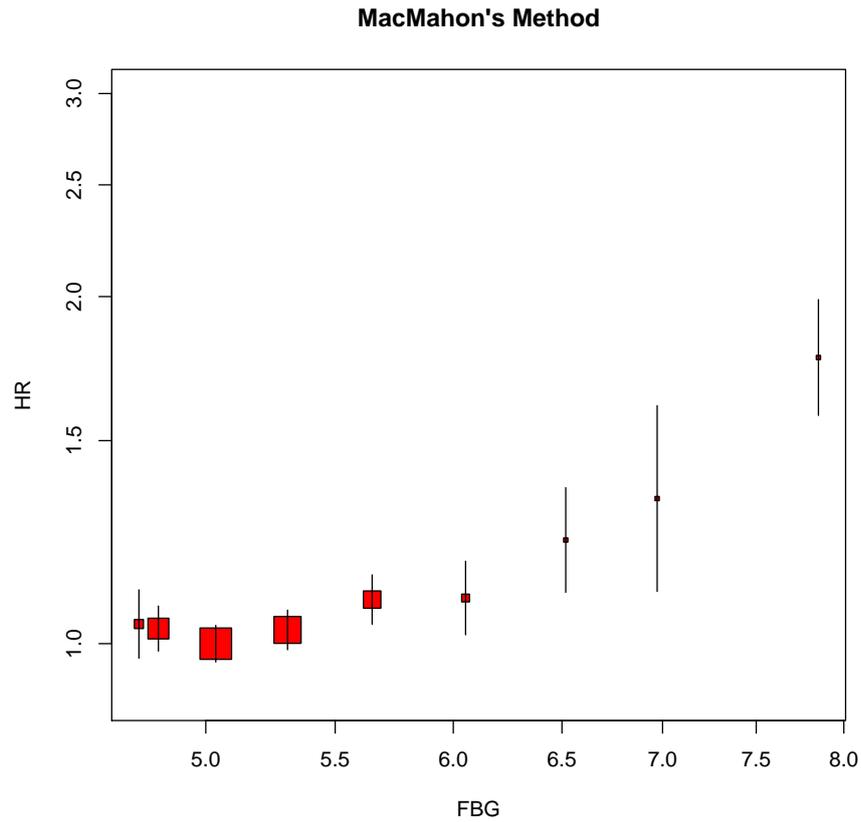


Figure 6.6: Measurement error corrected grouped exposure analysis of the FBG–CHD relationship using MacMahon’s method.

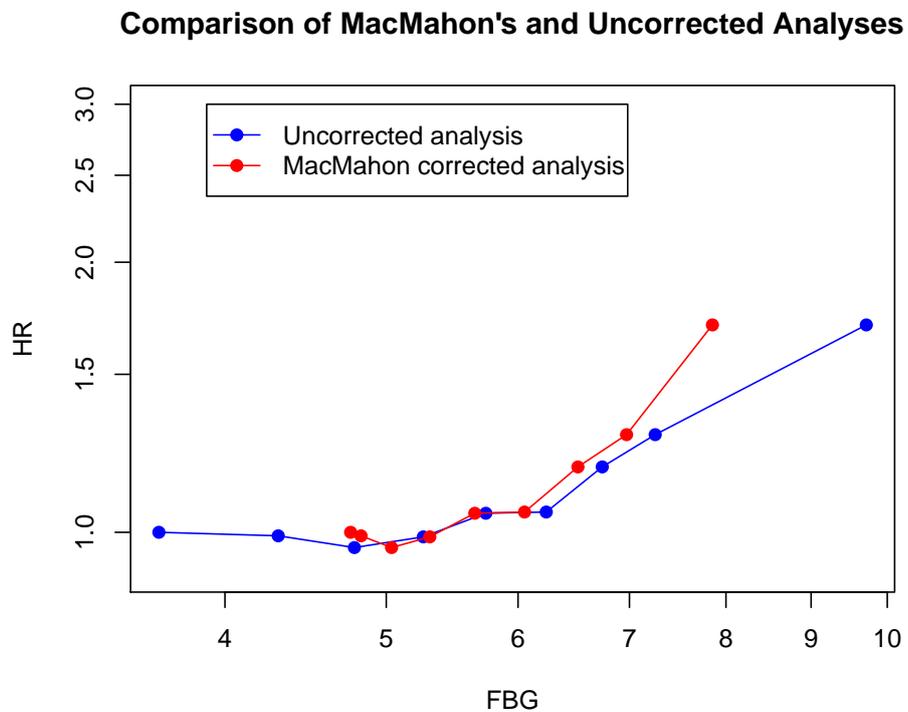


Figure 6.7: Comparison of the uncorrected and MacMahon corrected group exposure analyses, where the hazard ratio at the group means have been connected by linear segments to give an impression of the shape of the FBG–CHD relationship.

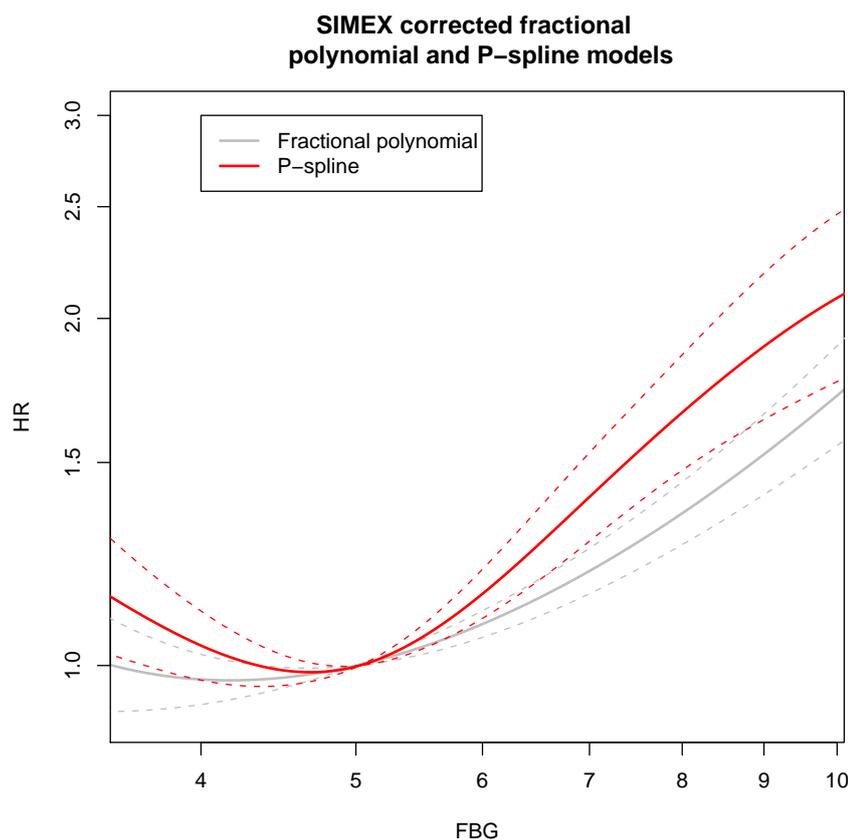


Figure 6.8: SIMEX corrected fractional polynomial and P-spline models for the FBG–CHD relationship.

beyond the range of the observed FBG measurements. In figure 6.8 we use 14 basis functions for log-FBG and constrain the P-spline to have four degrees of freedom. It is not clear how the penalisation of the individual P-spline models feeds through to the final model and hence whether the model obtained from this procedure will also have four degrees of freedom. We also applied SIMEX to the FP2 model of section 6.4.

In each case the SIMEX procedure was carried out using  $\zeta = \{0.5, 1, 1.5, 2\}$  with 100 datasets generated for each value of  $\zeta$ . Standard errors were calculated using the jackknife procedure [166] and both the parameters and their standard errors were extrapolated using a quadratic function. Figure 6.8 shows the corrected P-spline and fractional polynomial models for the FBG–CHD relationship. Comparing figure 6.8 with figure 6.3 we see that the curves obtained from the SIMEX procedure suggest that the effects of measurement error are modest, especially under the fractional polynomial model.

Figure 6.9 shows the extrapolation of the parameters of the fractional polynomial model in figure 6.8. The quadratic and rational linear extrapolants fit the parameter estimates  $\beta(\zeta)$  well and there is only a small difference between the extrapolated parameters. We do not show the extrapolation plots for the P-spline model due to the large number of parameters. Theoretically, and from our experience, SIMEX works well when the degree of measurement error is small because the extrapolation function is only approximately correct. In section 6.5 we saw that

---

the observed FBG measurements are subject to substantial measurement error and therefore we should view the results with an element of caution, although the measurement error correction is likely to be conservative [169].

### 6.6.3 Structural fractional polynomial and P-spline models for FBG

We now consider the results from structural fractional polynomial and P-spline models. The distribution of true log-FBG given the observed was assumed to be normally distributed with mean and variance parameters as calculated using the regression calibration models in section 6.5. The best fitting structural fractional polynomial of maximum degree two was the degree two model with powers (-2,-2). This is a different combination of powers than those for the unadjusted model.

One problem we encountered in applying the method of structural fractional polynomials method was that some of the quadrature points used to evaluate expectations of powers of usual FBG given observed FBG were located below zero. This was remedied by modifying the transformation of equation 2.9 to

$$x' = x - (x_{\min} + \sigma_x \omega_{\min}) + \gamma \quad (6.1)$$

where  $\sigma_x^2$  is the variance of  $x$ ,  $\omega_{\min}$  is the smallest quadrature point, and  $\gamma$  is the smallest difference between consecutive ordered exposure measurements. For 5 point Gaussian quadrature  $\omega_{\min} = -2.86$ . If the smallest value of  $x$  were 0.5, and the standard deviation of  $x$  were 0.2, then the smallest quadrature point would occur at  $0.5 + (0.2 \times -2.86) = -0.07$ . If the smallest difference between successive ordered observations were 0.05 so  $\gamma = 0.05$ , using equation 6.1 we shift the data by adding  $0.07 + 0.05 = 0.12$  to each observation so that all the quadrature nodes are positive. This example is illustrated in figure 6.10.

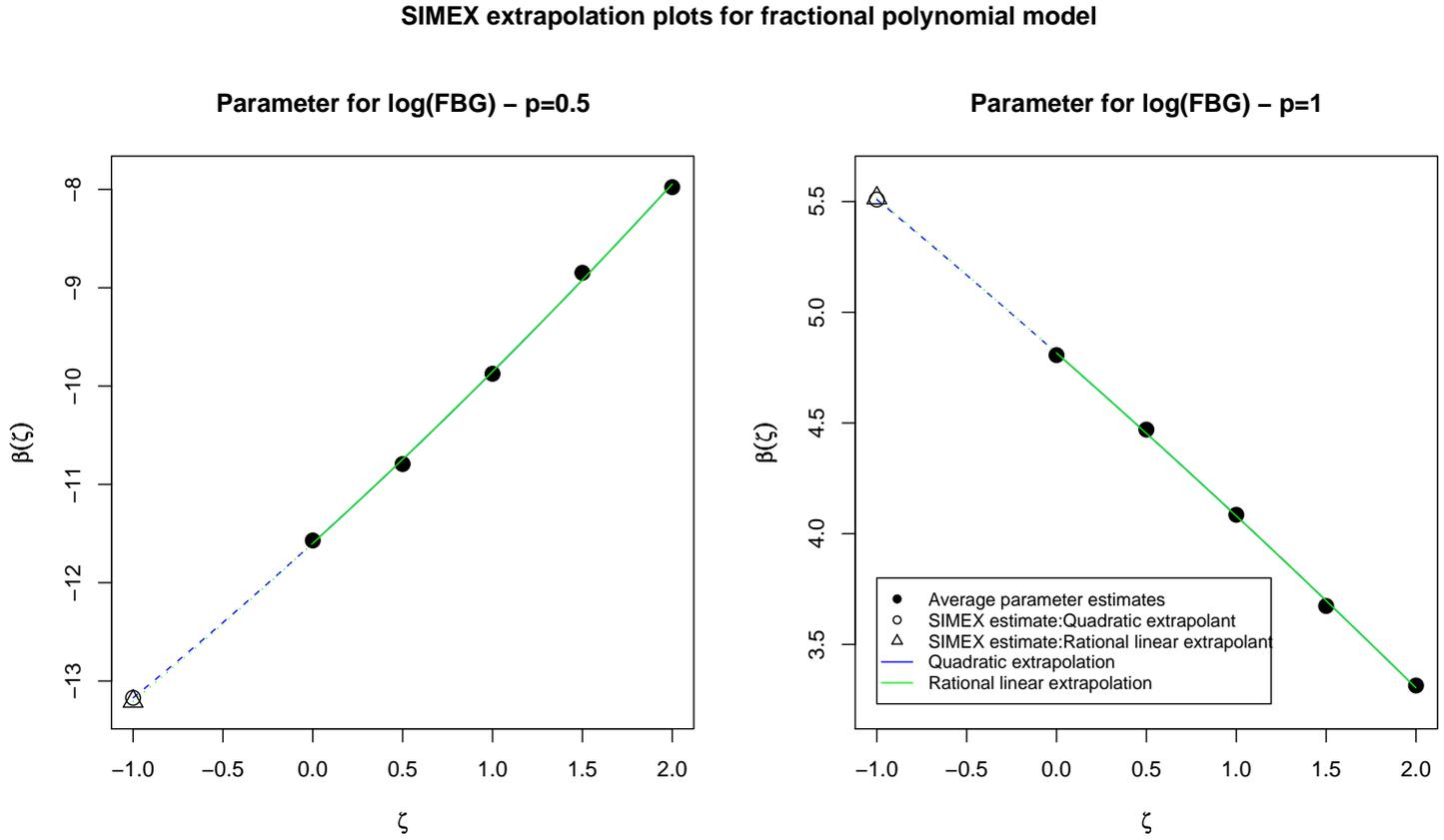


Figure 6.9: SIMEX extrapolation plots using the quadratic and rational linear extrapolants for the SIMEX corrected fractional polynomial model in figure 6.8.

---

Figure 6.11 shows a consistently and significantly higher risk of CHD events for those with FBG levels of 6–10 mmol/l under both the structural fractional polynomial and P-spline analyses, compared with those uncorrected for the effects of measurement error in FBG. This suggests that by ignoring the effects of measurement error we may be seriously underestimating the risk of CHD associated with levels of FBG consistent with IFG or diabetes. The structural fractional polynomial and P-spline analyses are in close agreement up until an FBG of 8 mmol/l; after this point the structural P-spline analysis suggests that the relationship begins to flatten off whilst the structural fractional polynomial analysis suggests the hazard continues to rise steeply. This difference is probably due to the P-spline models flexibility to fit local features of the data at the extremes of the exposure range.

We investigated the effect of using the model selection criterion AIC or cAIC that we introduced in section 2.2.4 to select the number of degrees of freedom in the P-spline models. We found that both methods of choosing the penalty led to severe undersmoothing. Both criterion chose a model with 8.1 degrees of freedom for the observed relationship, and 8 degrees of freedom for the measurement error corrected structural P-spline model. The structural P-spline model exhibited a turning point in the upper end of the FBG range, which implied that CHD risk decreased with increasing FBG beyond the turning point; this is epidemiologically implausible.

## 6.7 Discussion

There is a strong relationship between FBG and the hazard of experiencing a CHD event, with increased hazard for those with impaired and diabetic levels of FBG. The shape of the relationship is attenuated by the presence of substantial measurement error in the observed FBG measurements. The relationship between usual FBG and CHD is much stronger, with significantly higher risk associated with impaired and diabetic levels of FBG, in the measurement error corrected analyses. These results agree strongly with those of the APCSC [40]. The shape of the relationship appears to be S-shaped from the P-spline analyses however the fractional polynomial analyses suggest a J-shaped relationship. The fractional polynomial models are probably providing an extremely good fit in the normal range of FBG at the expense of some infidelity as you move outside this region, where less of the data lie. Figure 6.12 shows all of the continuous models that we have fit in this chapter for comparison.

The role of blood glucose in causing CHD events may also be related to other measures of how the body maintains blood glucose levels. For example, the DECODE study [215] found that the effect of FBG disappeared upon adjustment for blood glucose levels two hour post glucose load, suggesting that blood glucose peaks may also have a role to play in the risk of experiencing a CHD event.

Confounders appeared linearly in both our disease and regression calibration models. We in-

Example of transformation for structural fractional polynomial

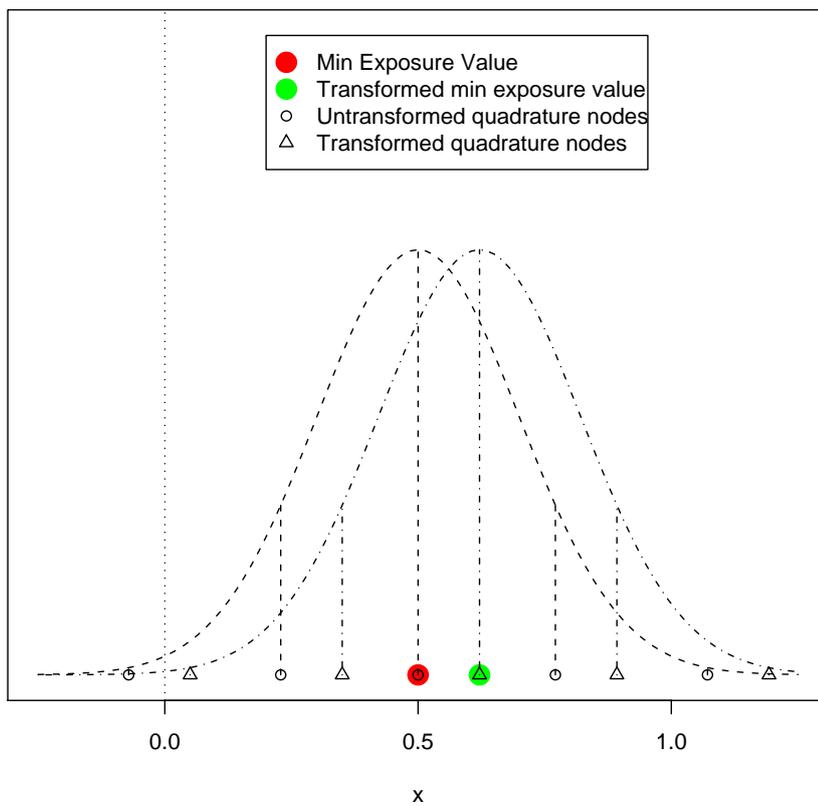


Figure 6.10: Illustration of data transformation given in equation 6.1 for structural fractional polynomials.

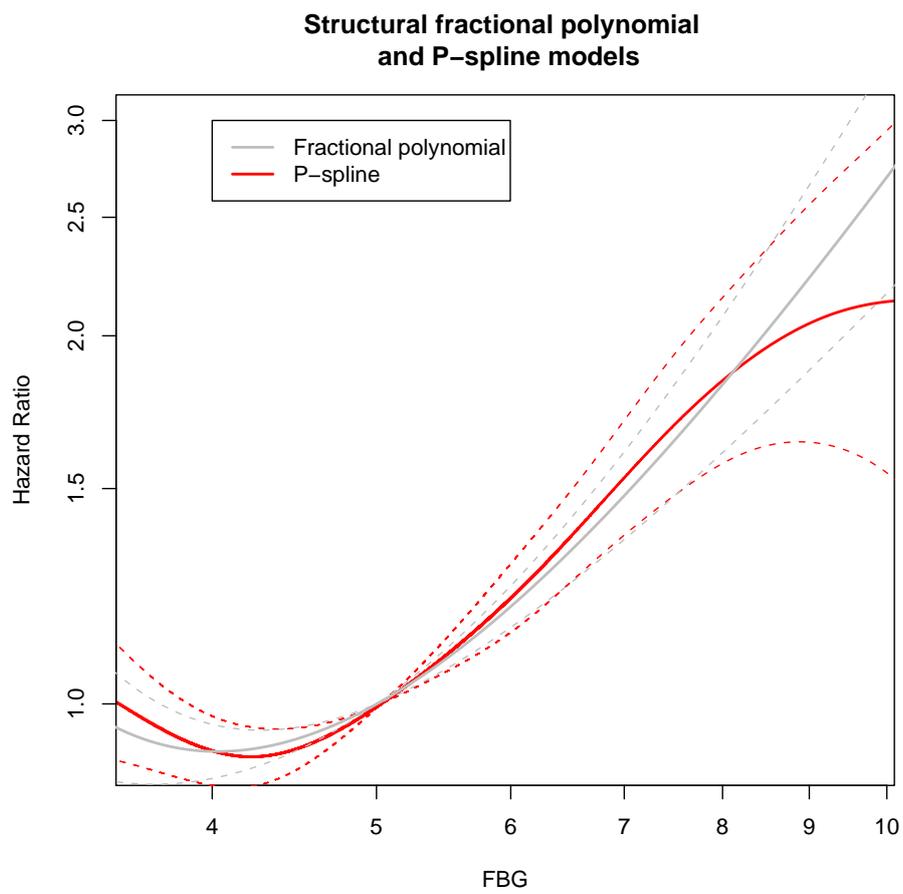


Figure 6.11: Structural fractional polynomial and P-spline models for the FBG–CHD relationship.

## 6. Application — Emerging Risk Factors Collaboration

---

investigated the effect of allowing them to appear non-linearly in our disease models by including P-spline terms for each of the confounders; this did not appear to substantially affect the results obtained. Misspecification of the regression calibration model is a potential problem but can be avoided by performing standard regression diagnostics. This can, however, be very time consuming when we have multiple studies because each study with repeats has its own regression calibration model.

In section 6.5 we used the measurement error plots proposed by Carroll *et al.* to assess the measurement error structure. We observed that the measurement error did not appear to adhere to the classical measurement error model. These plots do not seem to be widely used in practice and the assumption of classical measurement error is usually assumed. In the examples used by Carroll *et al.* [33] and Iturria, Carroll, and Firth [133], the plots show that the classical measurement error holds and that the error is normal. We have not found any examples in the literature where these plots (after possible taking a logarithmic transformation of the data) have suggested that the measurement is non-normal and/or heteroscedastic.

The confidence intervals we have presented in this chapter have not taken into account the uncertainty in the measurement error model, which is substantial for those studies for which we did have repeat measurements available. We have not acknowledged the possibility that the additional cardiovascular risk factors which we have adjusted for are subject to measurement error. Age can be assumed to have been measured without error. Some studies assume that BMI is measured without error although Davey Smith [45] believes that this assumption should not be made. Smoking status is likely to have some misclassification error due to individuals giving inaccurate information but we can probably assume this to be small [219–221]. It is well known, however, that measurements of blood pressure [35–37] and cholesterol [36–39] can be subject to a large amount of measurement error. We saw in figure 6.5 that the variance of the measurement error in log-FBG did not appear to be affected by the level of any of the confounders. We might therefore expect the difference between multivariate correction and correcting for measurement error in each variable univariately to be small. Day *et al.* [222] looked at the effect of correlated measurement error on a bivariate regression model and found that unless the measurement error in the exposures was highly correlated  $> 0.7$ , then the parameter estimates obtained tended not to be too different than those obtained when the measurement error was uncorrelated. Knuiman *et al.* [223] say that multivariate adjustment should be made after they obtained differing answers using univariate and multivariate adjustment. Repeat measures of SBP and cholesterol are available within the ERFC dataset so the error in the confounders could potentially be corrected for, but we do not consider this here.

Although the  $x$ -axis of all the FBG–CHD event associations shown in this chapter have been labelled FBG there is a difference between what we mean by FBG in section 6.4 and section 6.5. In the graphs of section 6.4 FBG refers to observed FBG levels which are subject to measurement error, whereas in the graphs of section 6.5 we are referring to usual exposure; the

---

long term average of FBG. The distribution of observed FBG will be much greater than that of usual FBG. The variance of usual FBG given observed FBG, which we use in the structural correction techniques, is smaller than the variance of usual FBG; we therefore view the relationship over a slightly reduced range. As discussed previously MacMahon's method severely reduces the range over which we can observe the FBG–CHD event relationship.

As is common in cohort studies we only used baseline values of risk factors in our analyses. In reality it is likely that the levels of risk factors will change over time. It is well known for example that BMI varies with age. The incidence of type 2 diabetes increases with age [224] hence we would expect FBG to increase with age. During short periods of follow-up the difference between current and baseline risk factors are likely to be small. When the follow up period is long, however, there may be large differences.

Our regression calibration models only used the first repeat measurement and hence we are throwing away any information that may be contained in the additional follow up measurements in the ERFC data. However, the longer the period between the baseline measure and that of the follow-up the less likely it is that the follow-up is going to be a true replicate of the baseline measure. Additional repeat measurements could be taken into account by including them in the regression calibration model [33].

## 6.8 Conclusion

The ERFC is a large IPD meta-analysis looking at risk factors for CHD. We focused on the relationship between FBG and CHD events because it is non-linear and FBG is subject to substantial measurement error. We fitted Cox proportional hazards models, stratified by sex, cohort (and trial arm where appropriate), to the data allowing for conventional cardiovascular risk factors: age, SBP, BMI, smoking status and total cholesterol. We considered the observed relationship using a grouped exposure analysis, fractional polynomials and P-splines. These analyses suggested that FBG has a relatively modest effect on the hazard of suffering a CHD event except for those in the upper diabetic range. We then considered MacMahon's method for correcting the grouped exposure analysis for the effects of measurement error in the FBG measurements; and SIMEX and structural P-spline and fractional polynomial analyses. The SIMEX procedure suggested that the effect of measurement error is modest, whereas the structural methods suggest that there is a significant increase in hazard associated with FBG levels consistent with IFG and diabetes and that the hazard is minimal for FBG somewhere in the range of 4–4.2 mmol/l.

In section 6.5 we saw that the measurement error appeared to be heteroscedastic and non-normal. In chapter 7 we look at the effects of non-normal and heteroscedastic error on the observed exposure–disease relationship and the corrected relationship when we incorrectly assume that the measurement error is normal and heteroscedastic. Then in chapter 8 we consider

## 6. Application — Emerging Risk Factors Collaboration

how to properly account for the heterogeneity in the shape of the exposure–disease relationship between studies using meta-analysis.

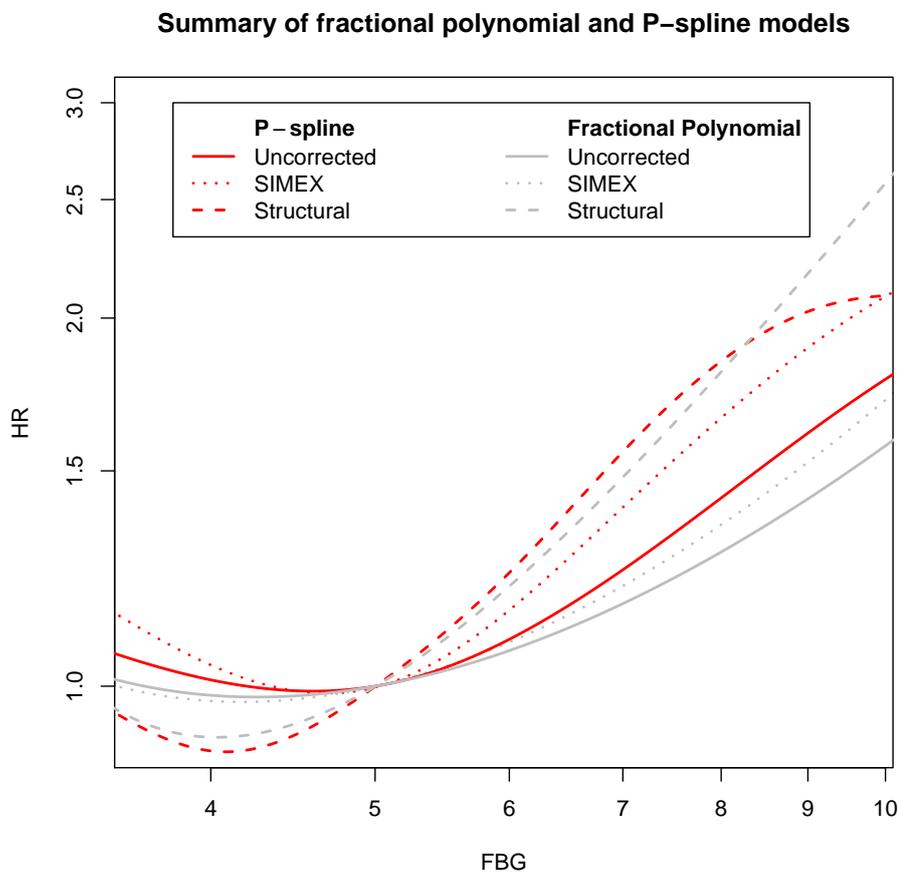


Figure 6.12: Summary of the fractional polynomial and P-spline models for the FBG–CHD relationship considered in chapter 6.

# Chapter 7

## Non-classical measurement error

Thus far in this dissertation we have assumed that either the measurement error, or the distribution of the true exposure given observed, is normal and homoscedastic. In this chapter we investigate non-classical measurement error and non-normally distributed true exposure, since our structural correction methods can be sensitive to the distribution of true exposure given observed. In chapter 6 we saw that the ERFC FBG measurements appeared to be subject to non-normal and heteroscedastic measurement error.

In the first section we extend the simulation study of chapter 3 to consider the effects of non-classical measurement error, and non-normal true exposure, on the observed exposure–disease relationship. We then look at the impact on structural fractional polynomials and structural P-splines of erroneously assuming that the distribution of the true exposure given observed is normally distributed. In the second section we consider methods for correcting for non-classical exposure measurement error. We use a mixture of normals for the distribution of true exposure given observed in order to allow for non-normality in structural fractional polynomial and P-spline models for modelling the FBG–CHD relationship within the ERFC.

### 7.1 Introduction

When using structural fractional polynomials and P-splines in chapter 6 we assumed that the distribution of usual FBG given observed FBG was normally distributed. Figure 6.5 suggested that this assumption may be inappropriate; the measurement error appeared to be non-normal and heteroscedastic. In general the distribution of the true exposure given observed may be non-normal if the true exposure is non-normally distributed; the measurement error is not independent of the true exposure; or the measurement error is non-normally distributed.

In many practical situations the distribution of true exposure given observed is unlikely to be normal. Few exposures are actually normally distributed; often a simple transformation of the data can be found, such that the transformed data are more normally distributed. When our

## 7. Non-classical measurement error

---

exposure is subject to measurement error, transformation of the observed data is not, in general, equivalent to transforming the true exposure; a point rarely appreciated. For example, normal measurement error can be seen to attenuate the skewness of the true exposure distribution

$$\begin{aligned} \frac{E((W - \mu_w)^3)}{(E((W - \mu_w)^2))^{\frac{3}{2}}} &= \frac{E((X + U - \mu_x)^3)}{(E((X + U - \mu_x)^2))^{\frac{3}{2}}} \\ &= \frac{E(((X - \mu_x) + U)^3)}{(E(((X - \mu_x)^2 + U)))^{\frac{3}{2}}} \\ &= \frac{E((X - \mu_x)^3)}{(E(X - \mu_x)^2 + E(U^2))^{\frac{3}{2}}} \end{aligned}$$

i.e.  $\text{skew}(W) = \frac{\sigma_x^3}{\sigma_w^3} \text{skew}(X) = \lambda^{\frac{3}{2}} \text{skew}(X)$ . The skewness of the observed exposure distribution reduces quicker than the exposure–disease relationship is attenuated. This makes identifying non-normality of the true exposure more difficult, especially when the measurement error variance is large.

Many exposures are subject to measurement error where the variance is not constant across the range of exposure. Examples include serum creatinine [27] and serum cotinine [225]. Non-constant variance may be due to the exposure measurement procedure being designed to accurately capture the level of exposure for those with high or low levels of exposure (i.e. diseased individuals) and therefore, those with normal exposure levels are subject to greater measurement error. Alternatively, those with extreme levels of exposure may be subject to greater within-person variation than those in the centre of the exposure distribution.

It seems plausible that the measurement error distribution for many exposures may be skewed and/or have heavier tails than the normal distribution. There seems, however, to be few examples of non-normal measurement error in the literature perhaps as mentioned in chapter 6, because investigations of the measurement error structure are rarely carried out in practice.

### 7.1.1 Introduction of non-linearity

It is well known that non-classical measurement error can introduce non-linearity into the observed exposure–disease relationship when it is truly linear and distorts the shape of non-linear relationships [33]. We now consider theoretical results regarding the introduction of non-linearity; firstly when our exposure is subject to multiplicative measurement error, an example of heteroscedastic measurement error; and then when the true exposure distribution is non-normal.

The multiplicative measurement error model, which was introduced in chapter 3, is perhaps the simplest heteroscedastic measurement error model. It is heteroscedastic because the measurement error variance increases with increasing exposure. Suppose that  $X$  and  $W$  are log-

normally distributed with

$$\log X \sim N(\mu_{\log x}, \sigma_{\log x}^2)$$

and

$$\log U \sim N\left(-\frac{\sigma_{\log u}^2}{2}, \sigma_{\log u}^2\right)$$

so that the measurement error is unbiased on the original scale. Under a multiplicative model  $W = XU$  so  $\log W = \log X + \log U$  meaning that

$$\log W \sim N\left(\mu_{\log x} - \frac{\sigma_{\log u}^2}{2}, \sigma_{\log x}^2 + \sigma_{\log u}^2\right).$$

This implies

$$\log X|W \sim \log N\left((1 - \lambda)\mu_{\log x} + \lambda\left(\log W + \frac{\sigma_{\log u}^2}{2}\right), \lambda\sigma_{\log u}^2\right)$$

where  $\lambda = \frac{\sigma_{\log x}^2}{\sigma_{\log w}^2}$  i.e. the RDR on the log scale. Then,

$$E(X|W) = W^\lambda \exp((1 - \lambda)\mu_{\log x} + \lambda\sigma_{\log u}^2).$$

Note how this is proportional to  $W^\lambda$  and not  $W$ . Hence, for a simple linear model where the exposure–disease relationship is linear the observed relationship is

$$\begin{aligned} E(Y|W) &= \beta_0 + \beta_1 E(X|W) \\ &= \beta_0 + \beta_1 W^\lambda \exp((1 - \lambda)\mu_{\log x} + \lambda\sigma_{\log u}^2) \end{aligned}$$

which is non-linear in  $W$ . Similarly for  $X^k$  we obtain  $E(X^k|W) \propto W^{k\lambda}$ . As  $W$  increases (for  $W > 1$ ) the observed exposure–disease relationship is increasingly attenuated.

Chesher [226] showed that non-normal true exposure can also introduce non-linearity in to true linear relationships or distort non-linear relationships, by giving a small variance approximation for the linear model

$$E(Y|W = w) = E(Y|X = w) + \frac{\sigma_u^2}{2} \left( 2 \frac{\partial E(Y|X = w)}{\partial X} \frac{\partial \log f_X(w)}{\partial X} + \frac{\partial^2 E(Y|X = w)}{\partial X^2} \right) + o(\sigma^2)$$

where  $f_X(\cdot)$  is the distribution function for the true exposure,  $X$ . When the exposure–disease relationship is linear this gives

$$E(Y|W = w) = \beta_0 + \beta_1 w + \sigma_u^2 \beta_1 \frac{\partial \log f_X(w)}{\partial X} + o(\sigma^2)$$

which is only linear in  $w$  when the true exposure is normally distributed; otherwise non-

linearity is introduced through the derivative of the distribution of the true exposure in the third term of the equation.

### 7.2 Diagnosing non-classical error

As discussed above, non-normal exposure and heteroscedastic measurement error can introduce non-linearity into linear relationships or distort non-linear relationships. We therefore need methods for identifying when our exposure is subject to non-classical measurement error or when our true exposure is non-normal.

In this section we shall consider how we can investigate the properties of the measurement error, and true exposure distributions, using the plots introduced in chapter 6. We shall focus on the common case where we have replicate exposure measurements for a random subset of individuals. It is essential that replicate measurements are obtained from across the exposure range otherwise we may miss important distributional features. Also, more repeats, on more individuals, allow us to better assess the distributional features of our exposure, and will allow better parameter estimation in our measurement error models.

In chapter 6 we described four plots for investigating the exposure and measurement error distributions. Here we do not consider plots of within-person standard deviation against confounders, but give a more detailed commentary on the other three plots:

1. *Normal Q-Q plot of within-person differences*—We may be able to discern that the measurement error distribution is heavy or light tailed, compared with the normal distribution. The symmetry induced by taking differences means that we are unable to tell from this plot whether the distribution is skewed. The Q-Q line should pass through the origin; if it does not then this implies that there is a difference in systematic bias between the baseline and repeat measurements.
2. *Plot of within-person standard deviation against within-person mean*—If the exposure is subject to classical measurement error we would expect this plot to show no trend [133]. If instead a linear trend is present it suggests a multiplicative error structure (i.e. the standard deviation of the measurement error increases linearly with exposure) and taking logarithms may give no relationship between exposure and standard deviation [133].

Inference from this plot is not affected by the distribution of the true exposure, although under different distributions of true exposure the cloud of data points will have a different shape. The relationship we observe between the within-person means and the measurement error variance is attenuated by the remaining error in the within-person means. The within-person means contain  $\frac{1}{k}$  times the measurement error of the individual observed values, where  $k$  is the number of repeat measurements. This attenuation prevents accurate modelling of the variance function. A shortcoming of this plot is that

---

the measurement error variance cannot be read off directly. This plot is preferred over a Bland-Altman plot [227] because trends in the standard-deviation across the range of exposure are more easily visible.

3. *Normal Q-Q plot of the within-person means*—This plot gives an indication of the normality of the true exposure. When appraising this plot one must bear in mind that non-normality may be due either to non-normality of the true exposure, or due to the remaining measurement error if it is non-normal. Non-normality of the true exposure could mask non-normality of the measurement error if they are skewed in opposite directions. If the measurement error in the observed values is large then the within-person means can still be subject to substantial measurement error if  $k$  is small. In this situation relatively small departures from normality in the normal Q-Q plot could belie large departures in normality of the true distribution. The slope of the Q-Q line is approximately the RDR of the within-person means if the measurement error is normally distributed.

## 7.3 Simulation study

We extend the third simulation study of chapter 5 to consider a non-normally distributed true exposure, exposure measurement error that is non-normal and/or heteroscedastic, and combinations of these. Firstly we look at examples of the measurement error plots we might observe under each of the ten scenarios; this allows us to investigate how easy it is to diagnose the source of non-classical error models. Then we look at the observed exposure–disease relationship under fractional polynomial and P-spline analyses. We then consider the exposure–disease relationships we obtain using structural fractional polynomial and structural P-spline correction techniques but erroneously assume that the distribution of the true exposure given observed is normally distributed with constant variance. This allows us to see how sensitive our structural methods are to violation of this assumption.

### 7.3.1 Simulation procedure

The simulations were carried out using the data generation process and shapes of relationship described in chapters 3 and 5. We shall continue to assume that our measurement error satisfies an additive model

$$W = X + U.$$

Heteroscedasticity is introduced into the observed exposure by allowing the measurement error variance to depend on a non-negative function  $g^2(x)$  such that

$$U \sim N(0, \sigma_u^2 g^2(X)).$$

## 7. Non-classical measurement error

---

$g^2(X) = 1$  corresponds to homoscedastic measurement error and  $g^2(X) = X^2$  corresponds to multiplicative measurement error. In our previous simulations we allowed the measurement error variance to take the values  $\sigma_u^2 = 0.25, 0.5, 1$ . In this simulation we use the same values for  $\sigma_u^2$  and choose  $g^2(x)$  such that

$$\int_{-\infty}^{\infty} g^2(x) f(x) dx = 1$$

where  $f(\cdot)$  is the appropriate density function of the true exposure to the scenario. This means that  $\sigma_u^2$  is the average measurement error variance and this allows comparison with previous simulations. We shift and scale the true exposure and measurement error distributions as appropriate so that  $X$  maintains mean 10, variance 1; and  $U$  mean 0, variance  $\sigma_u^2$ .

We shall consider ten scenarios for the distribution of the true exposure  $X$ , measurement error  $U$ , and the measurement error variance function  $g^2(x)$ ; these are listed in table 7.1.

Scenario 1 is the classical measurement error model which was considered in chapters 3 and 5 and is included here for comparison. Scenarios 2 and 3 consider non-normal true exposure. Under scenario 2 the true exposure follows a shifted Gamma distribution, which is positively skewed, and defined on  $[10 - \sqrt{6}, \infty)$ , hence the smallest possible true exposure will be 7.55 units. Under scenario 3 the exposure is  $t$ -distributed, which means the exposure will have heavier tails than the normal distribution.

Scenarios 4, 5, 6 consider heteroscedastic measurement error dependent on the level of true exposure. Scenario 4 is a multiplicative measurement error model, where the standard deviation of the measurement error increases linearly with  $X$ . Heteroscedastic error where the measurement error variance increases with the mean is commonly observed in practice. Under scenario 4 the variance two standard deviations above the mean of the true exposure is 2.25 times that two standard deviations below the mean. Scenario 5 is more extreme with the measurement error variance two standard deviations above the mean of the true exposure 9 times that two standard deviations below the mean. Scenario 6 was motivated by the ERFC FBG data where we observed that the measurement error variance increased about the mean, with those individuals with extreme true exposure values subject to measurement error with larger variance than those in the middle of the true exposure distribution. Under scenario 6 the measurement error variance two standard deviations above and below the mean is 13 times that at the mean.

Scenarios 7 and 8 consider non-normal distributions for the measurement error similar to those used for the true exposure under scenarios 2 and 3.

Scenario 9 is a combination of non-normal true exposure and heteroscedastic measurement error. After preliminary investigation of scenarios 1-9, we found that they were not as extreme as the trends exhibited by the ERFC FBG data. We therefore, searched for a scenario (scenario 10) that replicated some of the features of the ERFC FBG data. Under this scenario the

Class	Scenario	Distribution of $X$	Distribution of $U$	$g^2(x)$
Classical	1	$N(0, 1)$	$N(0, 1)$	1
Non-normally distributed $X$	2	$\Gamma(6, \sqrt{6})$	$N(0, 1)$	1
	3	$t_8$	$N(0, 1)$	1
Heteroscedastic measurement error	4	$N(0, 1)$	$N(0, 1)$	$x^2/101$
	5	$N(0, 1)$	$N(0, 1)$	$(x - 6)^2/17$
	6	$N(0, 1)$	$N(0, 1)$	$0.2 + 0.75(x - 10)^2$
Non-normal measurement error	7	$N(0, 1)$	$t_8$	1
	8	$N(0, 1)$	$\Gamma(6, \sqrt{6})$	1
Combination	9	$\Gamma(6, \sqrt{6})$	$N(0, 1)$	$x^2/101$
	10	$\chi_1^2$	$N(0, 1)$	$\frac{1}{0.6844} x - 10 $

Table 7.1: Description of each of the scenarios considered in the simulation study. Note that the distributions are shifted and scaled so that  $X$  maintains mean 10, variance 1 and  $U$  mean 0, variance  $\sigma_u^2$ .

minimum possible true exposure is 9 units.

We accept that these scenarios are not exhaustive, and that other scenarios not given here may be of interest, however we feel that this provides a sufficiently wide range of scenarios to give an indication of the possible effects of non-classical measurement error and non-normally distributed true exposure.

For each scenario we take a single sample of baseline and a single repeat measurement for 1,000 data points for  $\sigma_u^2 = 1$  to illustrate typical measurement error plots. We fit fractional polynomial and P-spline models to investigate the degree of non-linearity/distortion introduced into the observed exposure–disease relationship. We also fit structural fractional polynomial and P-spline models where we erroneously assume that the distribution of the true exposure given observed is normal to allow us to assess the sensitivity of our correction methods to violation of normality.

## 7.3.2 Results

### Measurement error plots

Figures 7.1 and 7.2 are plots of a single sample of baseline and a single repeat measurement for 1,000 randomly generated individuals taken from each of the ten scenarios described above when  $\sigma_u^2 = 1$ . They are meant to illustrate what may be observed under each scenario, however, due to sampling variation different trends may be seen in other samples.

In the first row of figure 7.1, where we have a classical measurement error model, we see no trend in the within-person standard deviation, and the normal Q-Q plots of within person differences and mean exposure suggest normality as expected. Rows 2 and 3 show that the non-normality of the true exposure makes little difference to the plots we obtain, and is only

## 7. Non-classical measurement error

---

reflected in slight non-linearity in the normal Q-Q plots of within-person means at the extremes. As described in section 7.1 this may be expected, since normal homoscedastic measurement error attenuates higher moments of the exposure distribution more than the mean exposure. In practice we would probably assume normality under both these scenarios. In rows 4 and 5 we see the linearly increasing trend in the standard deviation of the measurement error as we would expect.

In the first row of figure 7.2 we see a clear V-shape in the measurement error standard deviation. The effect of this severe heteroscedasticity is also to make the tails of the normal Q-Q plots of both the within-person differences *and* within-person means appear non-normal, giving the impression that the true exposure is non-normally distributed although this is not the case. Rows 2 and 3 shows that non-normality of the measurement error is hard to discern from the normal Q-Q plots and in fact manifests itself as trends in the standard deviation of the measurement error variance. Row 4 shows slight non-normality in the within individual means; as discussed above this reflects much greater non-normality in the true exposure. Row 5 is the most extreme of the scenarios considered, we see the V-shape in the measurement error standard deviation, the normal Q-Q plot of within-person differences shows non-normality in the tails, and the Q-Q plot of within-person means shows a positively skewed true distribution, all of these are features of the ERFC FBG data. We note that nearly 20% of observations lie below 9 units despite this being the minimum observable true exposure under this scenario.

Scenarios 2, 4, 5, 8-10, involve positively skewed true exposure or measurement error, or the measurement error variance increases with exposure, or combinations of these. In the next section, we therefore analyse these scenarios on the log scale; the usual tactic employed when dealing with positively skewed variables, and reduces the effects of outliers. Measurement error plots for the data on the log-scale for these scenarios are given in appendix D.

### Uncorrected exposure–disease relationships

Analyses were performed on all eight shapes for the exposure–disease relationship considered in chapter 5. For reasons of space, we shall concentrate on the linear and asymptotic relationships here, and briefly comment on the other shapes afterwards.

Figure 7.3 shows P-spline models for the observed relationship for a true linear exposure–disease relationship under each scenario. We see that the non-linearity introduced into the relationship is relatively small for most of the scenarios. When the measurement error is heteroscedastic (scenarios 4-6) we observe that there is greater attenuation in ranges where the measurement error variance is greater than the mean error variance, and less attenuation where the measurement error variance is less than the mean error variance; this is particularly noticeable under scenarios 5 and 6. The degree of non-linearity *decreases* as the measurement error variance increases (i.e. as the RDR decreases), this is due to the simultaneous attenuation of the measurement error variance function. Under scenarios 2 and 9 we see greater non-linearity

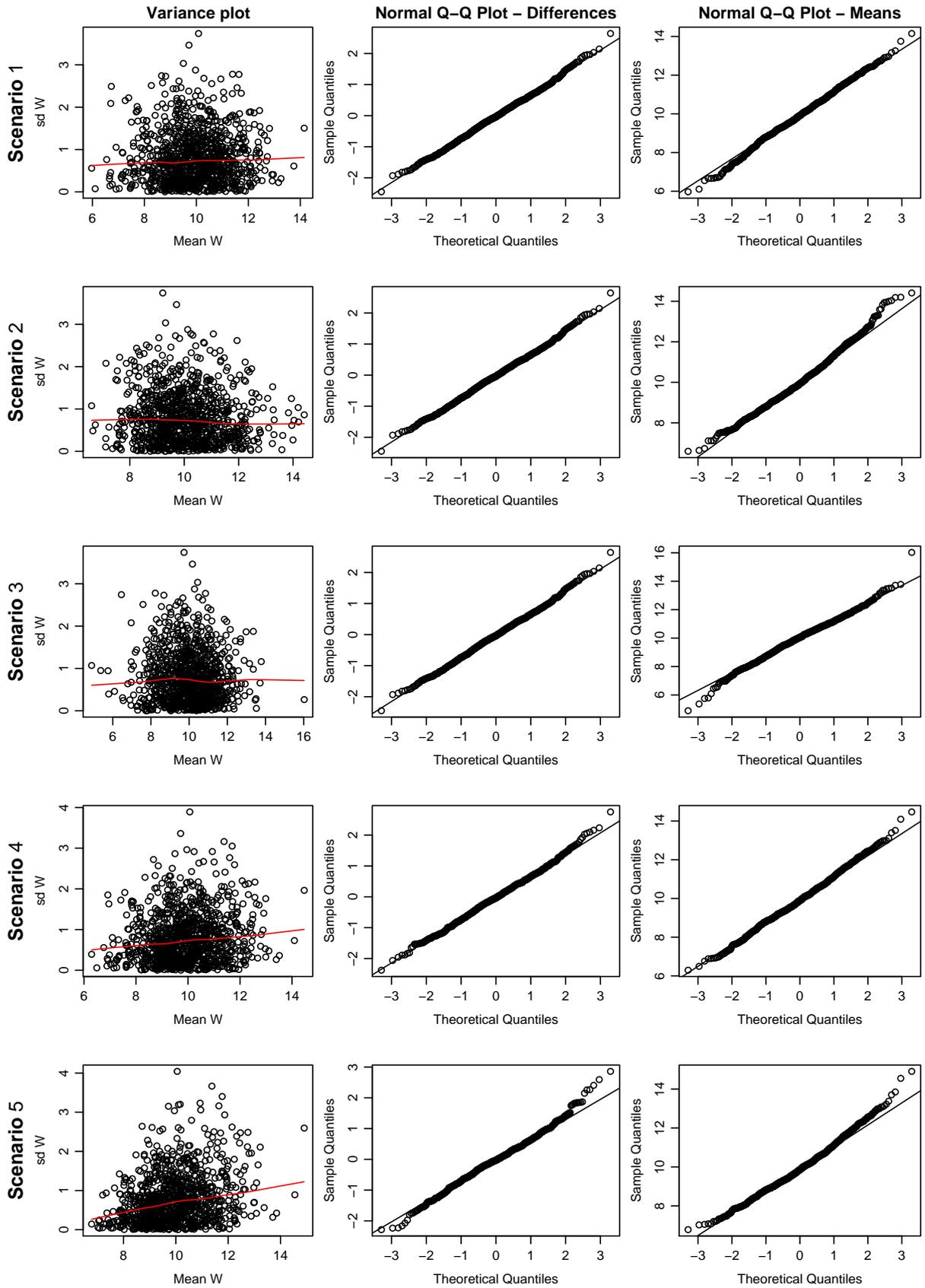


Figure 7.1: Example measurement error plots for a simulated sample of 1,000 individuals under scenarios 1-5,  $\sigma_u^2 = 1$ . Loess smooths of the data (red lines) are shown to aid trend identification in the left hand column.

## 7. Non-classical measurement error

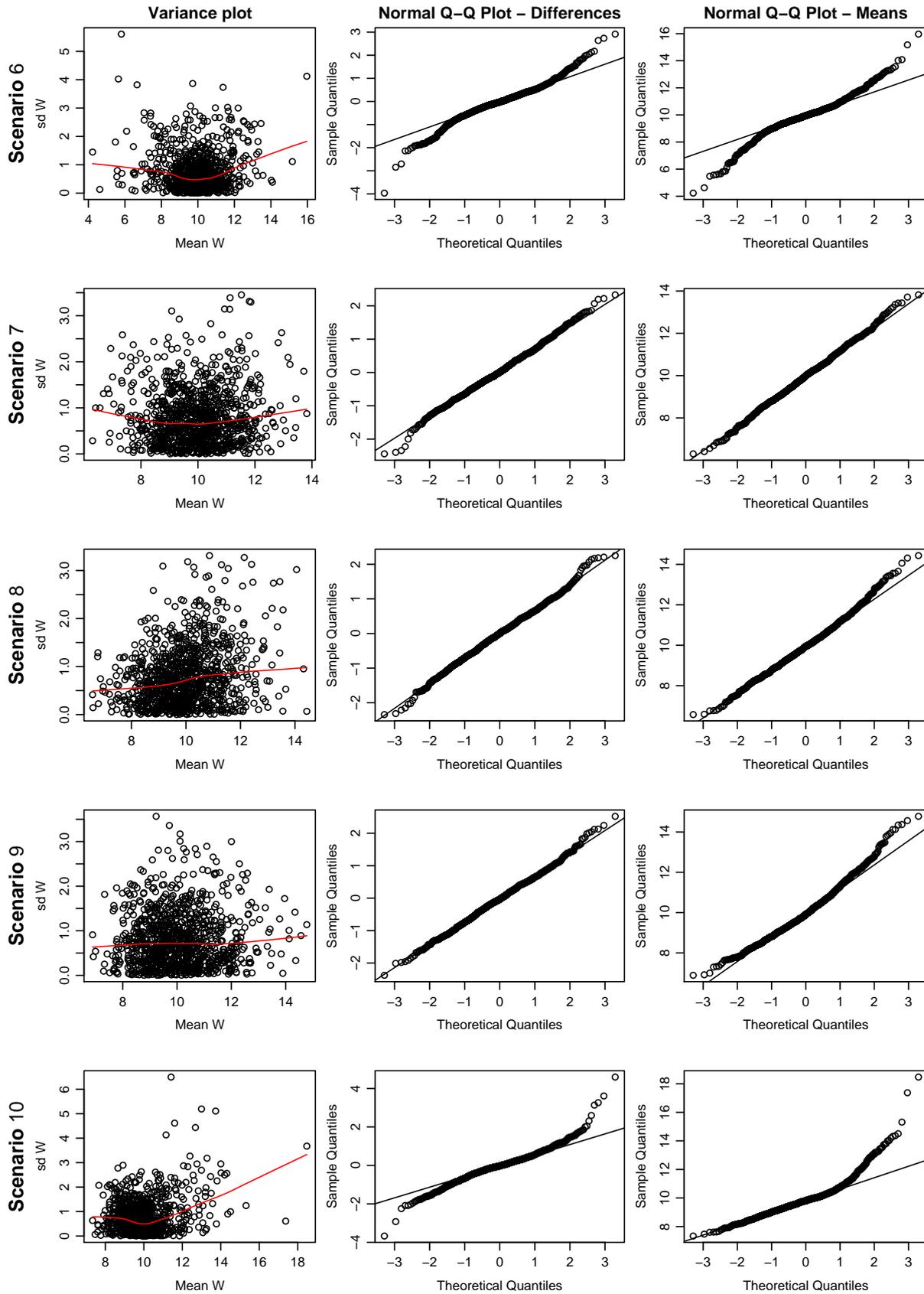


Figure 7.2: Example measurement error plots for a simulated sample of 1,000 individuals under scenarios 6-10,  $\sigma_u^2 = 1$ . Loess smooths of the data (red lines) are shown to aid trend identification in the left hand column.

---

of the relationship in the tails of the exposure and under scenario 10 we observe that there is considerable non-linearity introduced, the relationship could almost be described as exhibiting a threshold; this is a result of increasing numbers of observations located below the minimum observable true exposure due to measurement error.

Figure 7.4 shows the observed exposure–disease relationship under a true asymptotic relationship under each scenario. We observe similar trends as in figure 7.3. Under scenarios 2 and 9 the observed relationship is essentially linear, and under scenario 10, we observe a convex relationship. Similar trends are observed for each of the other shapes for the exposure–disease relationship, and there is little difference between fractional polynomials and P-splines except for under the threshold and U-shaped relationships. The differences for the threshold relationship are similar to those observed in chapter 5; P-splines pick up the threshold much better than fractional polynomials. Fractional polynomial and P-spline models under the U-shape relationship are illustrated for scenario 10 in figure 7.5. Similar, although not as extreme differences can be seen between the two methods under scenarios 2 and 9 as well. When there is no measurement error, fractional polynomials and P-splines give very different shapes for the U-shaped relationship below the mean exposure. Outside the range of the exposure the two methods behave quite differently; fractional polynomials retain the same functional form, whereas P-splines become linear. When the exposure measurements are subject to measurement error, all values below 9 units are the result of measurement error. Almost all these measurements will have true hazard ratios of less than 1.05. This is why we observe such a weak relationship at low exposure levels, which the fractional polynomial model has problems picking out. In this scenario, although we observe exposure values less than 9 units, there is no true relationship at this level.

### **Corrected exposure–disease relationships**

We shall now consider the effect of correcting for measurement error assuming that the distribution of the true exposure given observed is normal with constant variance. Under all scenarios, except for scenario 1, either or both of the assumptions of normality or constant variance of the true exposure given observed will be violated.

Figure 7.6 shows that although the normality and homoscedasticity assumptions are invalid under scenarios 2–10 in many situations our correction procedure may return the correct form for the true exposure–disease relationship. Under scenarios 3–7 there is little bias in the corrected relationship. Under scenario 8 there is some bias but this is localised to the extremes of the exposure range. Only scenarios 2, 9 and 10 exhibit significant bias. Under each of these three scenarios the corrected relationship is J-shaped; although under scenario 10 the bias is not large when only considering exposure levels of 9 units (the minimum true exposure observable) and above.

Under the true asymptotic relationship, figure 7.7, we see similar trends as under the linear

## 7. Non-classical measurement error

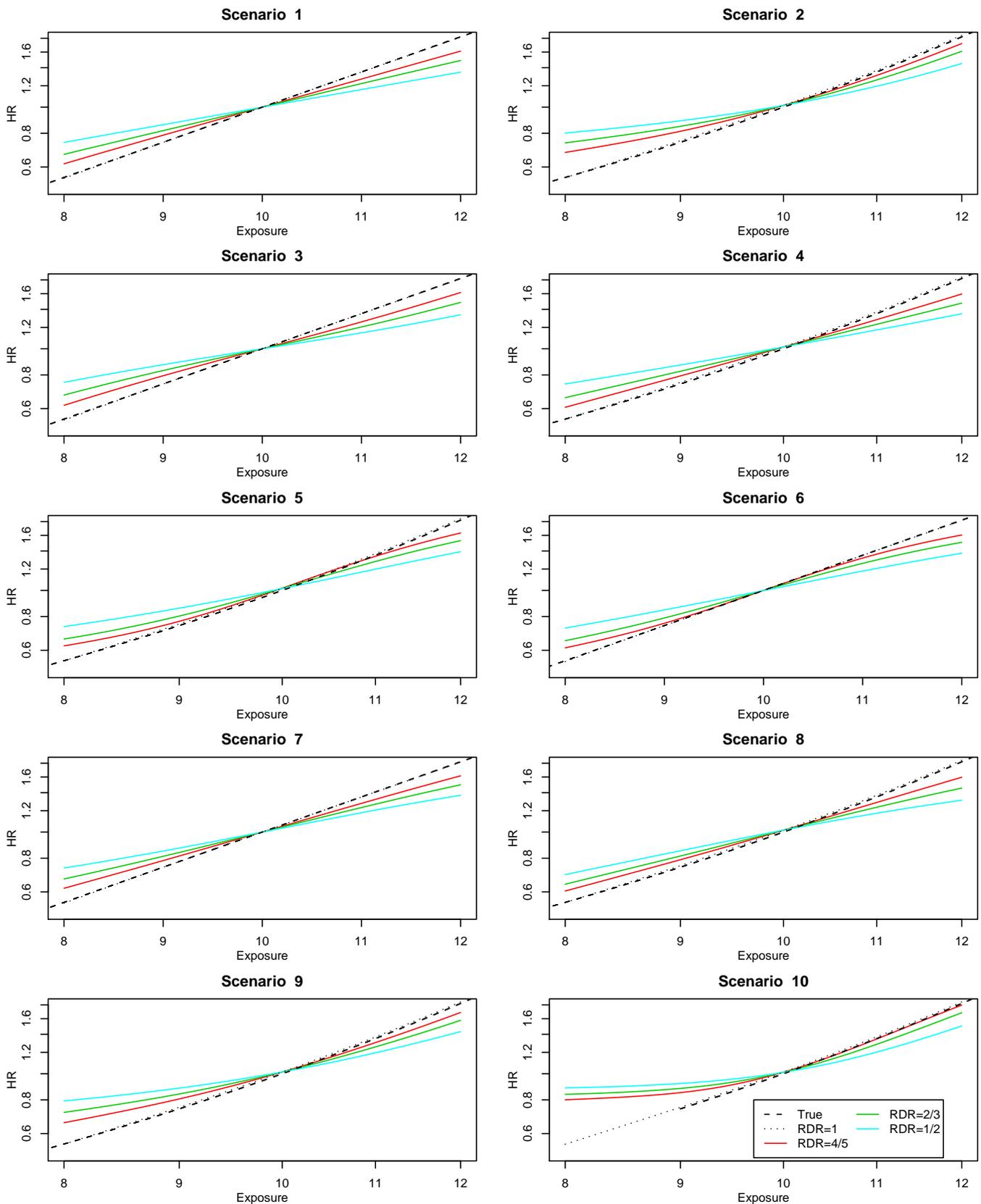


Figure 7.3: P-spline analysis showing the observed linear exposure–disease relationship under each of the ten scenarios.

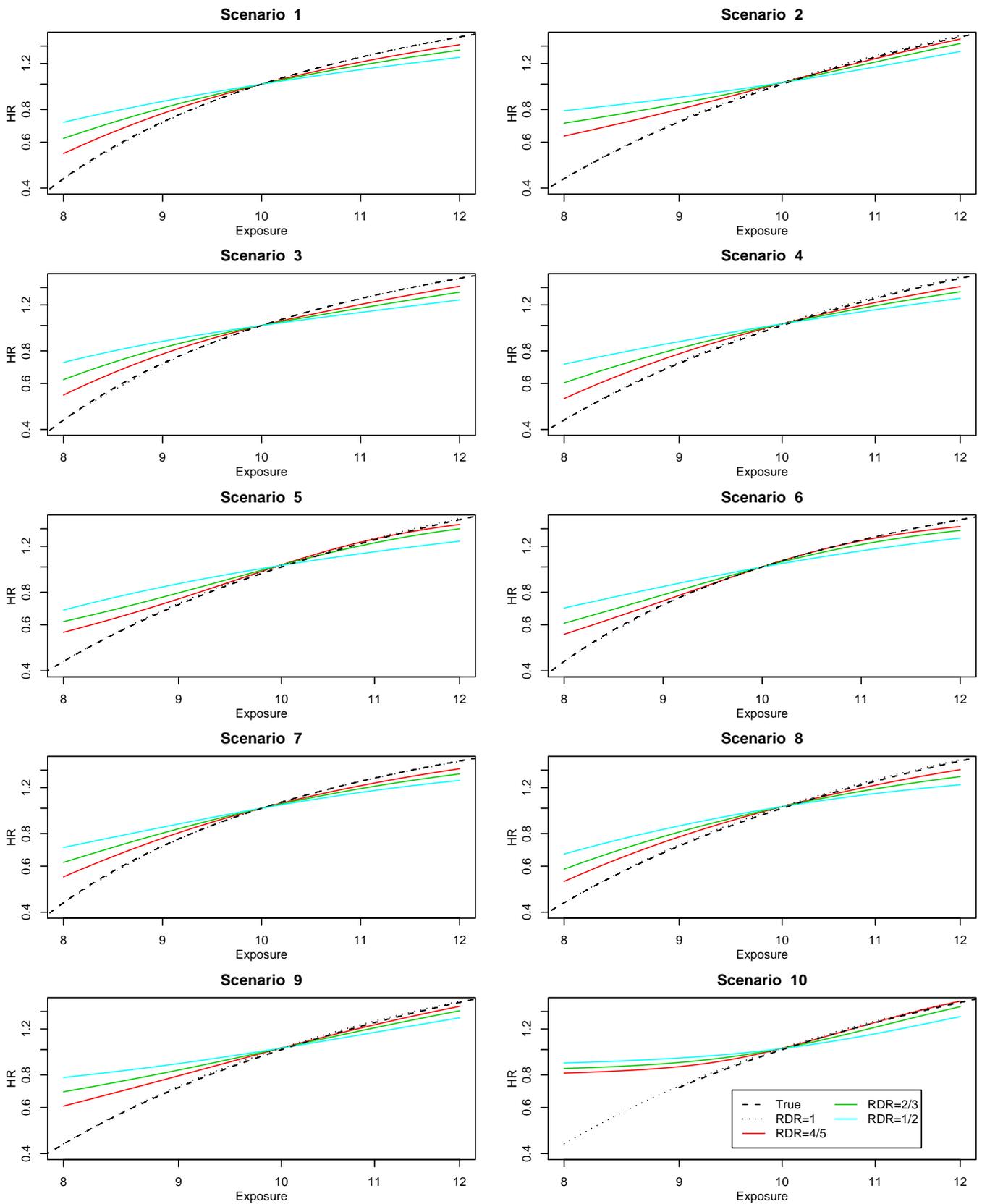


Figure 7.4: P-spline analysis showing the observed asymptotic exposure–disease relationship under each of the ten scenarios.

## 7. Non-classical measurement error

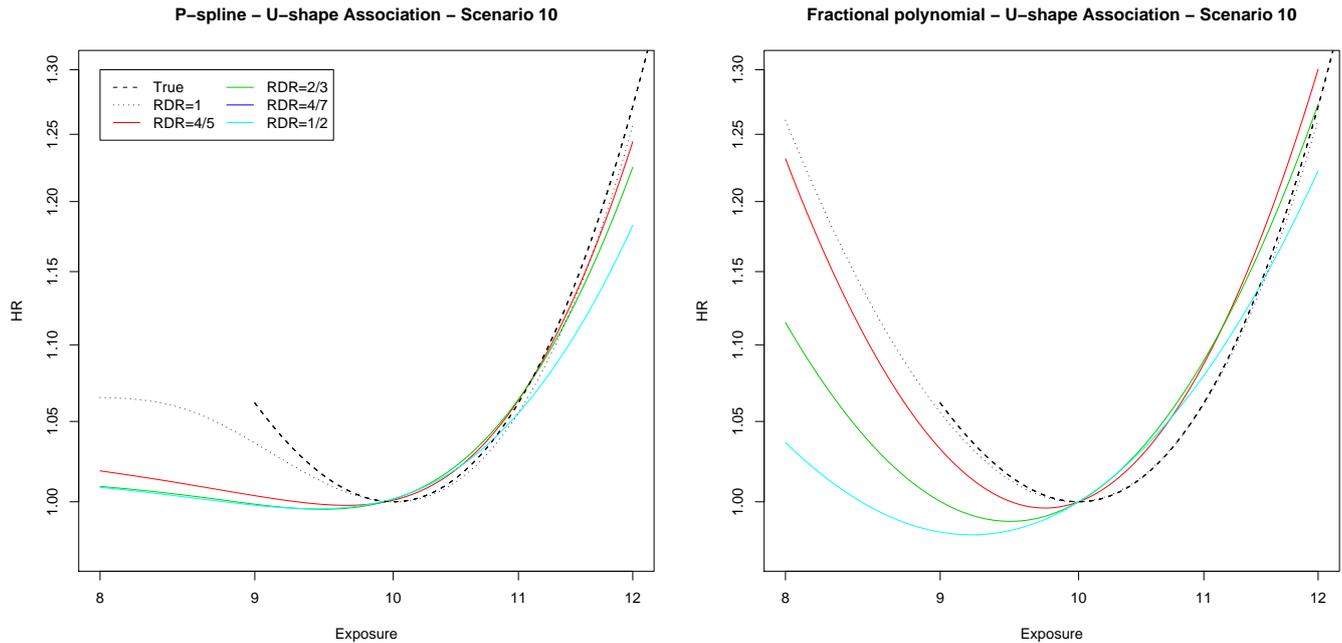


Figure 7.5: P-spline and fractional polynomial analyses of the observed threshold exposure-disease relationship under scenario 10.

relationship. Under scenarios 2, 9 and 10 the corrected relationship appears almost linear when the measurement error is low, before increasingly becoming a J-shaped relationship. The discrepancy we noted in figure 7.5 between fractional polynomial and P-spline analyses of a U-shaped relationship under scenario 10, can be seen in the corrected relationships (figure 7.8). Under the structural fractional polynomial model we overcorrect and the nadir increasingly moves away from the mean as increasing numbers of observations are observed with exposure values less than 9 units. Under the structural P-spline model there is severe over correction above the mean whilst at lower exposure levels the relationship is under corrected.

### 7.3.3 Discussion

We have considered data generation scenarios for the observed exposure involving non-classical measurement error, and non-normally distributed true exposure. We picked these scenarios based on intuition of what might be observed in practice and were surprised to find that they were very different to what we observe in the ERFC FBG data. It might, however, be that the FBG data are extreme compared with other exposures. This difference between the scenarios we initially included and the FBG data led us to also include scenario 10 in our simulation study. Scenario 10 aimed to recreate many of the features of the FBG data, and we obtained measurement error plots (figure 7.2) that were similar in appearance to those obtained from studies in the ERFC. Although scenario 10 was generated to resemble the FBG data it is probably more extreme in the sense that a high proportion of the true exposure values will have been located near to the minimum exposure value of 9 units, which caused us to observe a relation-

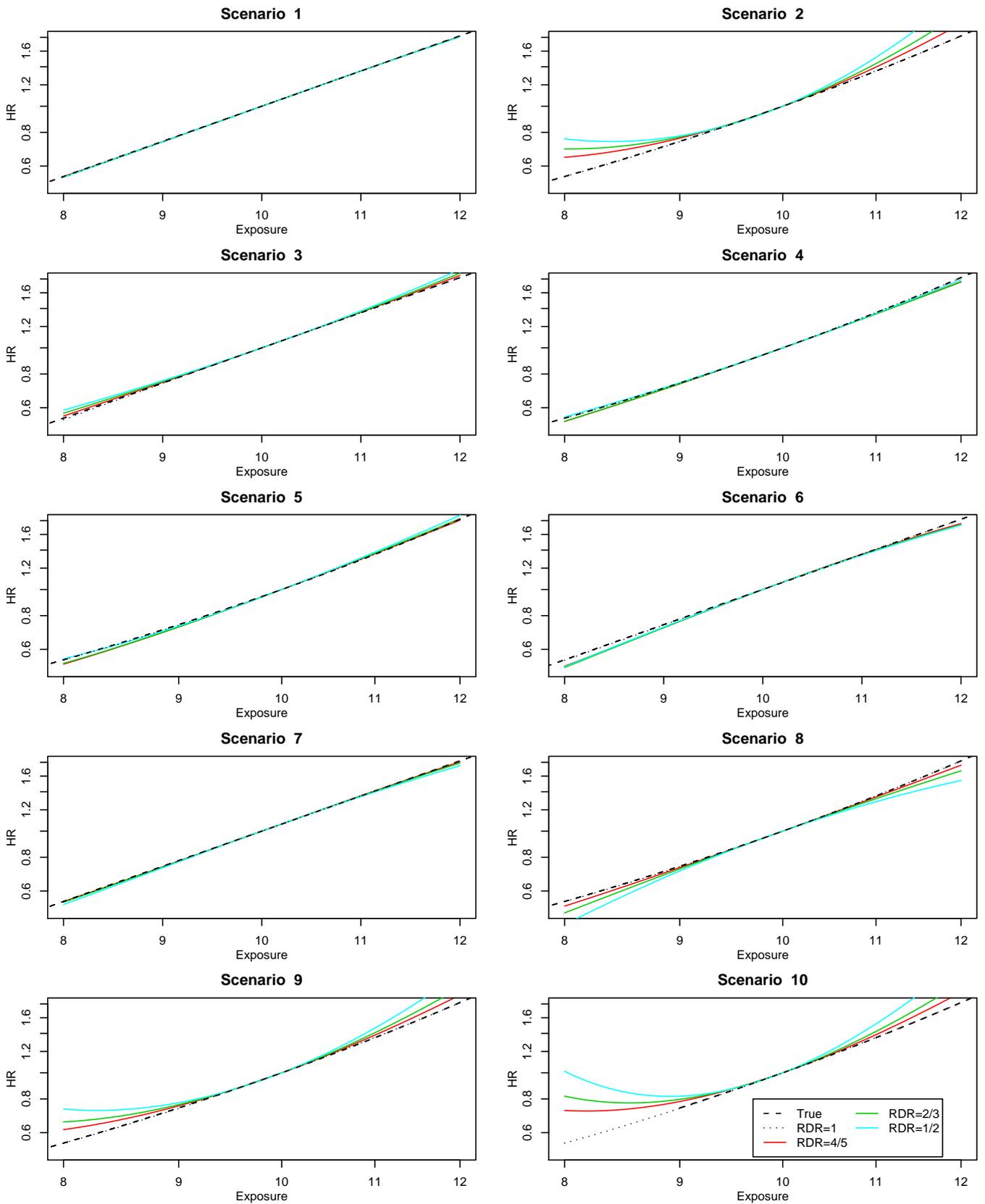


Figure 7.6: Structural fractional polynomial analysis of the corrected linear exposure–disease relationship under each of the ten scenarios.

## 7. Non-classical measurement error

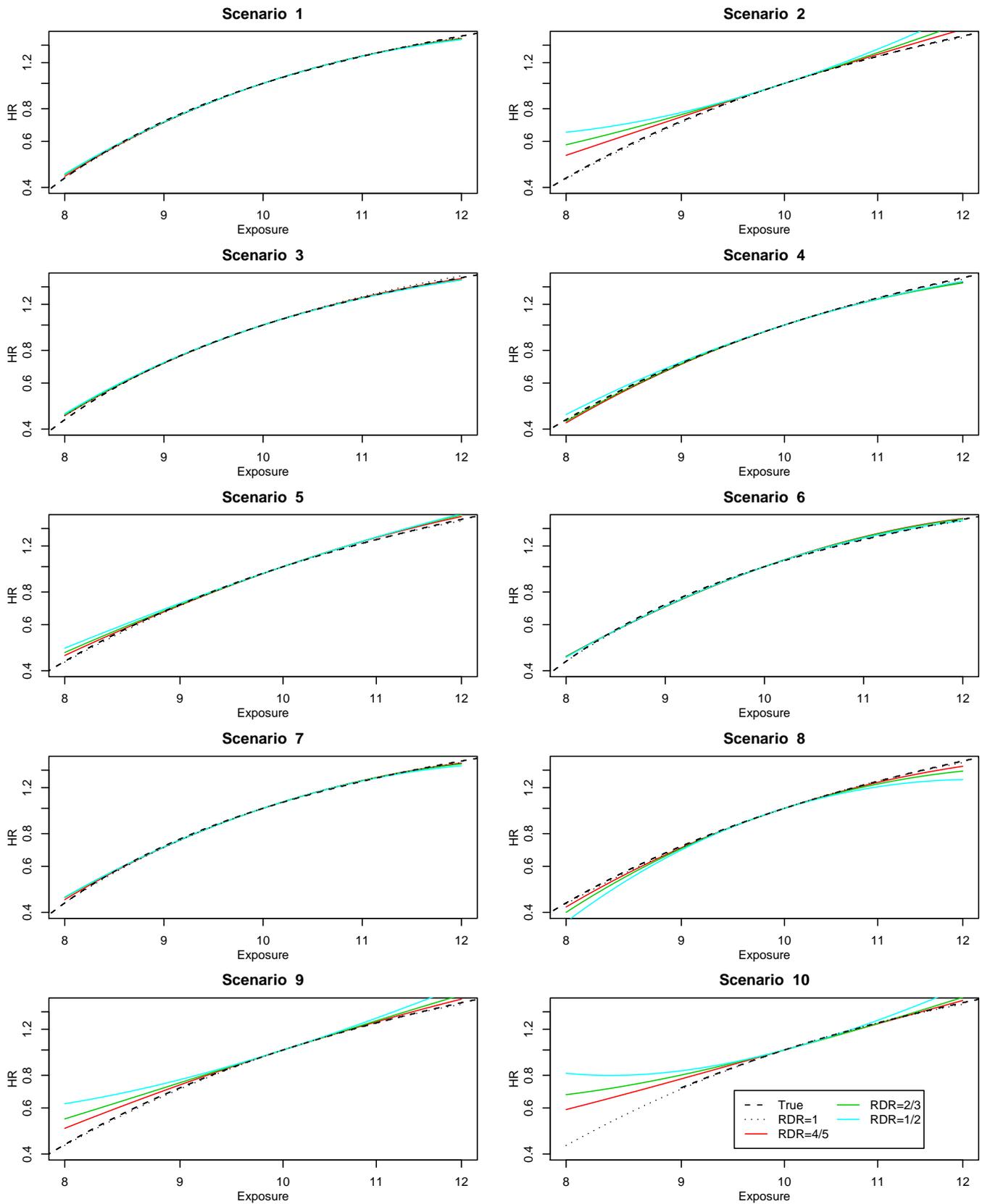


Figure 7.7: Structural fractional polynomial analysis of the corrected asymptotic exposure–disease relationship under each of the ten scenarios.

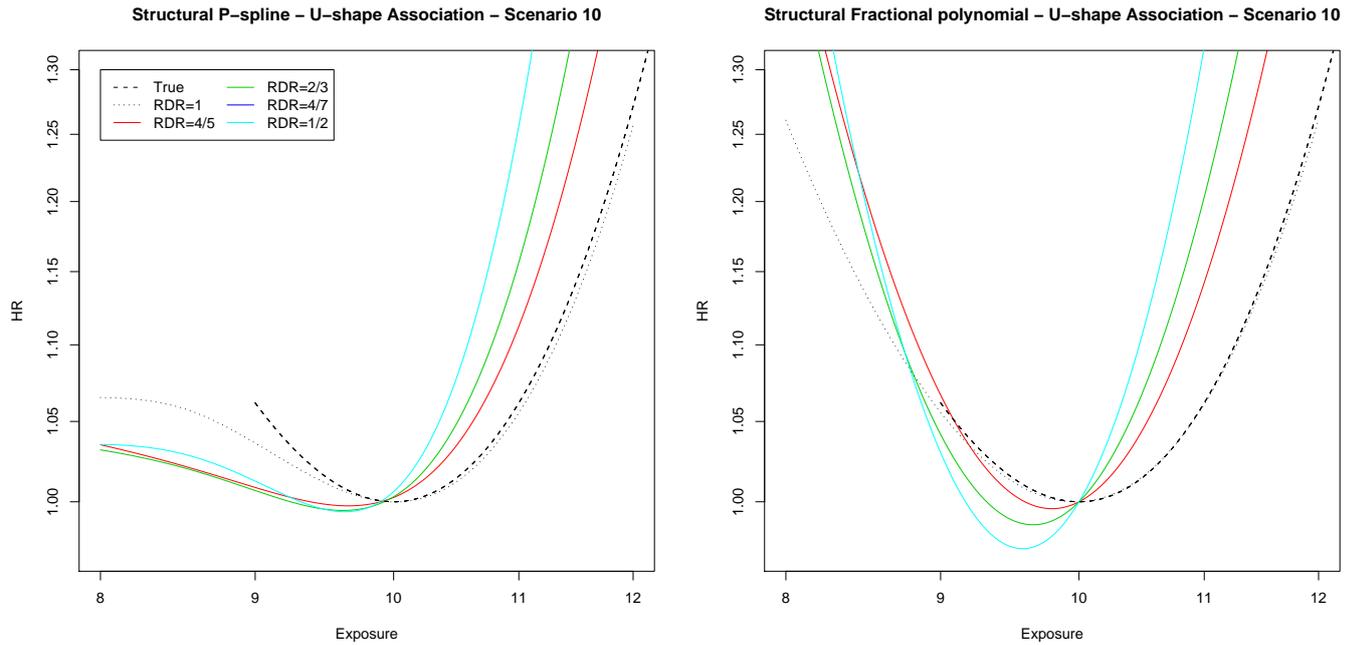


Figure 7.8: Structural P-spline and fractional polynomial analyses of the corrected threshold exposure–disease relationship under scenario 10.

ship where there was no true relationship (figure 7.5), and to mis-correct for measurement error (figure 7.8). It seems biologically plausible that for some exposures, including FBG, there is a minimum level of true exposure and that all observations below this level are the result of measurement error alone. We have observed that observations below the minimum observable true exposure can significantly bias our results and is an issue which we believe has not been previously discussed.

We have seen that structural fractional polynomials and structural P-splines are relatively robust to heteroscedasticity and non-normality of the measurement error distribution. However, non-linearity is introduced into a true linear relationship, and non-linear relationships are distorted, when the true exposure is non-normal. Although a log transformation of the data can help reduce the skewness of the exposure, the data in scenarios 2, 9 and 10 were still very non-normal.

In practice we may be saved from very non-normal distributions. Firstly, if the output of a laboratory analysis gave a wildly implausible value (as the result of measurement error) then in many cases the sample will be reanalysed. Secondly, gross outliers tend to be removed before the data are analysed.

In reality we do not know the true data generating mechanism. The diagnostic measurement error plots described in section 7.2 along with the plots of within-person standard deviations against confounders described in chapter 6 can help us to explore the distribution of the measurement error, assess whether the measurement error is constant, get an idea of the distribution

of the true exposure, and whether the measurement error is correlated with confounders. We have seen that it can be difficult to assess certain features of the measurement error graphically. Either we may misdiagnose the measurement error as not being heteroscedastic when in fact it is. Alternatively, as in the case of the ERFC, we may find that no simple transformation returns us to an additive model which leaves us in a situation which does not fit the standard models. It can also be difficult to identify the cause of what we observe; for example we have observed that non-normal measurement error can give plots of within-person standard-deviation against within-person mean that appear to show heteroscedasticity of the measurement error.

### 7.4 Correcting for non-classical measurement error

We shall briefly describe some of the correction methods that appear in the literature for correcting for non-classical measurement error, before considering a robust method for estimating the distribution of true exposure given observed which we apply to a re-analysis of the ERFC FBG–CHD relationship.

#### 7.4.1 Methods proposed for correcting for non-classical error

If the measurement error variance appears to increase linearly with exposure, then analysis on the logarithmic scale may be appropriate; this is the approach we took in the previous section. The measurement error plots described in section 7.3.2 allow us to check whether taking the logarithm is effective.

Guo and Little [53] propose that when the measurement error variance increases with exposure level we can approximate the distribution of  $X|W$  by  $X|W \sim N(\lambda_0 + \lambda_1 W, \pi^2 W^{2\gamma})$  where  $\gamma$  is estimated as the slope of the regression of the log squared residuals of the regression of  $X$  on  $W$  on the log of  $W^2$ . This method requires that we have observed a gold standard exposure measurement for a subset of individuals.

Augustin, Doring, and Rummel [164] use a modification of regression calibration for a linear exposure–disease relationship that allows for heteroscedastic measurement error. Augustin comments that calculating regression calibrated values for other powers of the true exposure requires further research.

Spiegelman, Logan, and Grove [54] recently proposed an adapted efficient regression calibration estimator, that used gold standard data to take account of heteroscedastic measurement error. They found that ordinary regression calibration actually performed better than any of the methods they used to correct for heteroscedastic measurement error. This seems to agree with our findings in the last section.

Empirical-SIMEX [175] deals with the problem of heteroscedastic error by forming pseudo datasets from linear combinations of the observations made on each individual. This method

requires replicate observations to be available for all individuals.

Non-normality of the true exposure, which was the main cause of distortions of the exposure–disease relationship in our simulation study, is only problematic for structural correction methods. For functional correction methods, such as SIMEX, we do not need to assume anything about the true exposure distribution. Guolo [150] gives a comprehensive review of robust techniques for measurement error correction.

As we mentioned in our description of structural P-splines in chapter 4 Carroll *et al.* [157] suggested that we use a mixture of normals to estimate the distribution of the true exposure given observed to robustify structural P-spline models. Similarly, this approach can also be used for structural fractional polynomials. We pursue this approach in the next section since it is a natural extension of the methods we have previously considered.

### 7.4.2 Mixture regression calibration modelling

Misspecification of the distribution of the true exposure given observed, can lead us to make invalid inference, as we saw in some of the scenarios considered in the previous section. Semi-parametric approaches allow greater flexibility for this distribution and can be more robust. Semi-parametric models can lead to a major loss of efficiency if a suitably chosen parametric model is approximately correct. A solution to this problem is to use a parametric model for the distribution that offers a high degree of flexibility.

Finite mixture models are flexible parametric models that are able to accurately reproduce a wide range of distributions [228]. Mixture models assume that observations belong to one of  $K$  latent classes, each of which has its own distribution e.g. normal with different parameters for each of the  $K$  latent classes in a normal mixture model. Ruppert, Carroll, and Maca [157] suggested using a mixture of normals to build robustness into structural regression splines. Mixture models have also been used in other measurement error problems by Carroll, Roeder, and Wasserman [229] and Richardson, Jaussent, and Green [230]. Two approaches to fitting mixture models are possible, Bayesian and classical likelihood—in this chapter we consider a classical likelihood approach.

Suppose that we wish to fit a normal mixture model with  $K$  components. The likelihood contribution of the  $k$ th component of the mixture regression calibration model is given by

$$f_k(\mathbf{x}|\mathbf{w}_1, \mathbf{z}, \boldsymbol{\alpha}_k, \sigma_k^2) = \prod_{i=1}^n \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_k^2}(x_i - (\alpha_{0k} + \alpha_{k1}w_{1i} + \boldsymbol{\alpha}_{zk}^T \mathbf{z}_i))^2\right) \quad (7.1)$$

and the likelihood of the  $K$ -component normal mixture model is given by the weighted com-

bination of the  $K$  components

$$f(\mathbf{x}|\mathbf{w}_1, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}) = \sum_{k=1}^K \pi_k f_k(\mathbf{x}|\mathbf{w}_1, \mathbf{z}, \boldsymbol{\alpha}_k, \sigma_k^2) \quad (7.2)$$

where the weights  $\boldsymbol{\pi}$  are constrained so that  $\pi_k \geq 0$  and  $\sum_{k=1}^K \pi_k = 1$ . This likelihood can be maximised directly, or via the EM algorithm [231].

### 7.4.3 Application — ERFC FBG data

In this section we apply the mixture modelling approach described above to the ERFC FBG data. We shall restrict ourselves to considering a mixture of two normals in this case because of the complexities introduced by considering multiple studies. It has also been suggested that the distribution of blood glucose is bimodal with peaks in the normal blood glucose and diabetic ranges [232].

For those studies with repeat measurements we fit a two component mixture regression calibration model by regressing the repeat FBG measurements on the baseline measurements, where each component takes the form of the regression calibration models used in chapter 6. We can then obtain the weight for each component, as well as fitted values and the variance of the residuals for each component. We adjust the residual variances for each component using a method of moments estimate of the measurement error within that component, in a similar way to section 5.2.1.

For those studies where we do not have repeat FBG measurements available for a subset of individuals we use the same approach as described in chapter 6, whereby we take a weighted average of the mixture calibration models from each study. In this case it is not obvious how to combine the regression calibration models for those studies with repeats across studies to obtain an ‘overall model’. This is because the regression calibration model for each study with repeats has two components which have no natural ordering.

We have reordered the components of the mixture models from each of the regression calibration models for studies with repeats according to the mean fitted values within each component of the mixture. The ‘overall model’ was obtained by taking the inverse variance weighted average of the models for each component across studies.

In figure 7.9 we see the result of fitting two component flexible structural fractional polynomial and P-spline models to the FBG data. Under the P-spline model of figure 6.11 where we assumed normality of the true FBG given the observed the curve flattens out between an FBG of between 8–10 mmol/l. Under the mixture model the curve continues to rise, although at a slower rate than before 8 mmol/l. The minimum risk occurs at a value of about 4.8 mmol/l under the mixture model and the risk rises as the level of FBG decreases. This is in contrast

---

to the normal model where the minimum occurred at a lower level and the model suggested that low levels of FBG had a protective effect. The structural fractional polynomial mixture model is quite different from the structural P-spline mixture model, especially for values of FBG above the mean where the fractional polynomial suggests a much lower hazard; although by 10 mmol/l the fractional polynomial model is similar to the P-spline.

In this section we only considered a two component mixture model, ideally however, we would use model selection criteria to decide on the number of components. We chose to order the components of the mixture models within each study with repeats according to the mean fitted values of the individual components. Although this is a reasonable approach to have taken it is *ad hoc* and the resulting ‘overall model’ we used for those studies without repeat measurements is clearly subject to large uncertainty; both because we are estimating a greater number of parameters in our measurement error model than when we used a single component, and because of the issues related to the ordering of the mixture’s components. To obtain a mixture model for those studies without repeat measurements a better approach would be to fit a single hierarchical model to all the studies with repeats; doing this would allow us to get around the problem of component ordering. The problems encountered with mixture modelling highlight the problems that can be introduced when we consider multiple studies.

In this section we have seen that a number of methods have been proposed for non-classical measurement error. We saw that a normal mixture model can be used to allow for non-normality of the distribution of the true exposure given observed. When we applied this to as part of estimating the FBG–CHD relationship, it suggested that the results of chapter 6 *may* be sensitive to the normality assumption we made about true FBG given observed FBG. This model is, however, subject to great uncertainty because of the problems associated with the mixture of normals used in the measurement error model for those studies without repeats. For this reason we do not take this model forward in later chapters, but we do suggest it is an area for future research.

## 7.5 Conclusion

In the first section of this chapter we considered the effects of non-normal true exposure, and non-normal and/or heteroscedastic measurement error on the observed exposure–disease relationship. We also considered the effect on structural fractional polynomial and P-spline methods for the exposure–disease relationship where we erroneously assumed that the distribution of the true exposure given observed was normally distributed. We saw that plots using the means and differences between replicate observations within individuals allow us to gain an insight into the measurement error structure and the distribution of the true exposure, although these plots were often hard to interpret.

Heteroscedastic measurement error can cause non-linearity in the observed exposure–disease

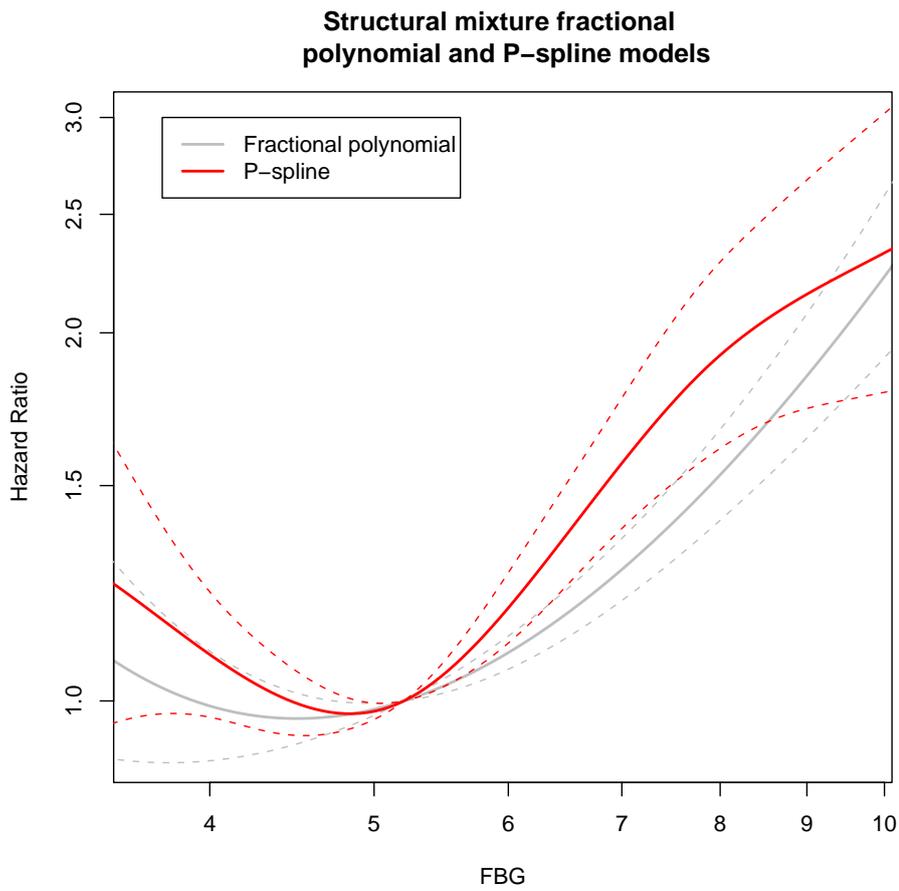


Figure 7.9: Structural fractional polynomial and P-spline models for the FBG–CHD relationship where the distribution of true FBG given the observed is modelled using a two-component normal mixture model.

---

relationship when it is truly linear, and distorts the shape of the relationship when it is non-linear, but this effect decreases as the measurement error variance increases, since the measurement error variance is simultaneously attenuated. In our simulation, however, structural fractional polynomials and P-splines appeared to be relatively robust to heteroscedastic measurement error. Non-normality also has a similar effect on the shape of the relationship but the effect remains when the measurement error variance increases. Structural P-splines and fractional polynomials retain the non-linearity introduced by heteroscedasticity and non-normality of the true exposure. We also noted that significant bias can be introduced when observed exposure values lie outside of the range of the true exposure distribution due to measurement error.

In the second section we briefly discussed methods for correcting for non-classical error. We then investigated using a two-component mixture model for the distribution of the true exposure given observed for structural fractional polynomial and P-spline models. We applied this approach to the FBG–CHD relationship because in chapter 6 we saw that the measurement error appeared to be heteroscedastic, and the distribution of true FBG non-normal. The relationship observed under the two-component structural P-spline model was similar to the single component model except at levels of FBG below the mean; however, the two-component structural fractional polynomial was much less steep above the mean level of FBG. These results suggest that our models may be sensitive to the assumption of normality of true FBG given observed. We did however, note that there was considerable uncertainty surrounding the measurement error model for those studies with repeat exposure measurements.

In chapter 6 we ignored heterogeneity in the shape of the FBG–CHD relationship between studies. In chapter 8 we consider meta-analysis, which will allow us to properly account for this variation, so that we are able to ascertain our best estimate of the shape of the FBG–CHD relationship.



# Chapter 8

## Meta-analysis

When modelling the FBG–CHD relationship in chapter 6 we ignored heterogeneity in the shape of the relationship between studies. In this chapter, we consider methods to take account of this heterogeneity. We start by giving an introduction to meta-analysis; this sets the background and introduces the methods we shall use later in the chapter. We then consider individual participant data (IPD) meta-analysis which is the gold standard approach to meta-analysis. Sauerbrei and Royston [67] recently proposed an approach for performing a 2-stage IPD meta-analysis using fractional polynomials, taking into account heterogeneity in the shape of the exposure–disease relationship between studies. We describe their approach, consider how it might be extended to P-spline models, and consider alternative methods for pooling relationships across studies. The methods described in this chapter allow us to obtain the ultimate goal of this dissertation: to produce measurement error corrected non-linear exposure–disease relationships in a large IPD meta-analysis. Hence, we conclude this chapter by reanalysing the FBG–CHD relationship.

### 8.1 Evidence synthesis

Evidence synthesis is the combination of multiple sources of data on a research question. The initial stage of an investigation is to clearly define the research question, and the strategy to find relevant studies. The search usually uses queries of search engines such as PubMed and Embase to identify papers that are potentially relevant. The abstracts of these papers will be read to assess their relevance to the research question. The identified papers will then be read in full. Once the relevant papers have been identified the references within those papers and those citing them will usually also be checked because both are likely to be relevant. Other ways of identifying relevant data include *ad hoc* searches, databases of trials, conference abstracts, and discussions with researchers in the field. Usually more than one researcher will be involved in this process to ensure that no studies ‘slip through the net’. The methodology used should be published, with a diagram showing the identification process often proving helpful to the reader.

One of the greatest challenges of quantitative evidence synthesis is extracting an effect size estimate  $Y_i$  from each paper, and its associated variance  $V_{Y_i}$ . This estimate may be an estimated slope, odds ratio, or log hazard ratio for a linear, logistic, or Cox regression respectively. Different studies may have used different analysis methods, different levels of adjustment for confounding, different study design or different ways of reporting the outcome. For example, different studies may have used different forms for the linear predictor, or different groupings under a grouped exposure analysis, which can make it difficult to obtain the parameter of interest, and its variance, from each study. A process of standardisation between studies therefore needs to be performed.

### 8.2 Advantages/disadvantages of meta-analysis

Meta-analysis is the statistical synthesis of results from a number of studies [233]. As opposed to a narrative review that draws conclusions in a qualitative manner, the aim of meta-analysis is to do so quantitatively. Meta-analyses often allow conclusions to be drawn from a set of studies that could not be drawn from any one individually.

There are many advantages to meta-analysis including:

- They are much cheaper than conducting a new trial or cohort study.
- They make best use of existing data.
- They can yield results much more quickly than waiting for the completion of a new study. This is especially the case in epidemiological studies where follow-up periods may be many decades long.
- They allow greater statistical power to find a significant relationship.
- They can help prioritise areas where further research should be conducted. Similarly in some instances they may indicate areas where sufficient research has already been conducted.

Despite the many advantages of meta-analysis, many criticisms have also been made including [234]:

- Garbage — The use of poor-quality studies leads to poor-quality results. The studies that go into a meta-analysis must be of good quality.
- Apples and Oranges — Often meta-analyses are accused of comparing ‘apples and oranges’ i.e. comparing studies that are not measuring the same outcome or differ in other significant ways. Heterogeneity between studies can arise in a number of ways such as differences in study design and adjustment for confounders. This can be prevented by clearly defining the research question, using judgement, and by exploring the sources of heterogeneity.

- File drawer problem — In meta-analysis of clinical trials there is a tendency for studies that do not show a significant effect not to be published but instead to be ‘filed away in the drawer’. Similarly, in epidemiology when studies consider multiple exposure–disease relationships, significant relationships are more likely to be published. The risk of failing to include all relevant studies can be reduced partially by performing a thorough evidence synthesis. We shall see in section 8.4.2, that we can also test whether publication bias is evident within our collection of studies.

## 8.3 Univariate meta-analysis

Having obtained a set of effects size estimates  $Y_i$ , and their variances  $V_{Y_i}$ , from an evidence synthesis we need to be able to combine these estimates across studies allowing for the relative precision of the estimates. In this section we consider fixed effect and random effects approaches to combining these estimates to produce an overall estimate  $\hat{\theta}$  of the true effect, and  $\hat{\mu}$  of the mean effect respectively.

### 8.3.1 Fixed effect meta-analysis

In a fixed effect meta-analysis we believe that all studies were measuring the same true effect size,  $\theta$ , and that the variation in the observed effect sizes,  $Y_i$ , between individual studies is due to random sampling error,  $\epsilon_i$  i.e. we assume that the observed effect size is the sum of the true effect and within-study sampling error

$$Y_i = \theta + \epsilon_i, \quad \text{Var}(\epsilon_i) = V_{Y_i}.$$

For example  $Y_i$  might be an estimate of the log-hazard ratio from the  $i$ th study, and  $\theta$  its true underlying value.

In a fixed effect meta-analysis each study,  $i$ , is given a weight

$$W_i = \frac{1}{V_{Y_i}}$$

which is the inverse variance of the estimated effect size,  $V_{Y_i}$ ; for example this might be the estimated variance of the log-hazard ratio,  $Y_i$ . Large studies will typically receive large weights as they will estimate the true effect more precisely, while small studies will tend to receive low weights as they will have measured the true effect size less precisely.

The summary effect  $\hat{\theta}$  is the weighted mean of the effect sizes from each of the  $K$  individual studies

$$\hat{\theta} = \frac{\sum_{i=1}^K W_i Y_i}{\sum_{i=1}^K W_i}.$$

The variance of the mean effect size is given by the inverse of the sum of the weights

$$V_{\hat{\theta}} = \frac{1}{\sum_{i=1}^K W_i}.$$

### 8.3.2 Random effects meta-analysis

Under the random effects model we assume that the true effect size under each study is drawn from a distribution of possible true effect sizes. The differences in true effect size between studies could be due to factors such as differences in study protocols (e.g. studies may have studied groups of people with different characteristics) and in observational studies unmeasured confounding variables. The observed effect size in the  $i$ th study is expressed as the mean effect size  $\mu$  plus a deviation from this  $\eta_i$ , plus the within-study sampling error

$$Y_i = \mu + \eta_i + \epsilon_i.$$

The variance of  $Y_i$  is therefore given by the sum of the sampling error variance plus the variance of the study specific true effect sizes,  $\tau^2$ . There are several ways to estimate  $\tau^2$ ; we shall consider the method of moments or DerSimonian and Laird [235] method below and discuss other methods in section 8.3.4. There is some difficulty in how we interpret the results from a random effects analysis. Borenstein *et al.* [233] describes the summary effect as the mean of all the relevant true effects.

It is argued that fixed effect meta-analysis is rarely justified [236] and that random effects analyses should always be carried out since it will collapse down to a fixed effect analysis if the observed heterogeneity between studies is zero. Visually, non-overlapping confidence intervals give an indication that there is heterogeneity between studies and that a random effects meta-analysis may be more appropriate. Statistical tests such as the  $Q$  statistic, described below, should be used to test for heterogeneity. If significant heterogeneity is present we may wish to look at the studies for any important differences between the studies that may have been missed earlier in the evidence synthesis process.

### 8.3.3 DerSimonian and Laird estimate of $\tau^2$

The observed weighted amount of variation between studies is given by

$$Q = \sum_{i=1}^K W_i (Y_i - \hat{\theta})^2 \quad (8.1)$$

$$= \sum_{i=1}^K W_i Y_i^2 - \frac{\left(\sum_{i=1}^K W_i Y_i\right)^2}{\sum_{i=1}^K W_i}. \quad (8.2)$$

$Q$  is the weighted sum of squared differences between the individual study effect sizes and the mean effect size. Under the null hypothesis that all  $K$  studies have the same true effect size,  $Q$  has a chi-squared distribution with  $K - 1$  degrees of freedom. Hence we can test for heterogeneity between studies by comparing the value of  $Q$  against a  $\chi_{K-1}^2$  distribution. We should not be over reliant on this test of heterogeneity, as non-significant values of  $Q$  could result from low power to detect heterogeneity due to the meta-analysis being of a small number of studies and/or the studies having large within study variance.

We estimate  $\tau^2$  by

$$\hat{\tau}^2 = \frac{Q - (K - 1)}{\sum_{i=1}^K W_i - \frac{\sum_{i=1}^K W_i^2}{\sum_{i=1}^K W_i}}.$$

$\tau^2$  is a variance and therefore cannot take values less than zero. Due to sampling error  $\hat{\tau}^2$  can be negative if  $Q - (K - 1) < 0$  in which case  $\hat{\tau}^2$  is set equal to 0.

Once  $\tau^2$  has been estimated we can calculate the random effects meta-analysis using weights

$$W_i^* = \frac{1}{V_{Y_i}^*}$$

where

$$V_{Y_i}^* = V_{Y_i} + \hat{\tau}^2.$$

The mean effect size,  $\hat{\mu}$ , is calculated as under the fixed effect model, but using  $W^*$  which takes into account the between-studies variance

$$\hat{\mu} = \frac{\sum_{i=1}^K W_i^* Y_i}{\sum_{i=1}^K W_i^*}, \quad V_{\hat{\mu}} = \frac{1}{\sum_{i=1}^K W_i^*}. \quad (8.3)$$

The range of weights within a random effects meta-analysis will be smaller than in a fixed effect meta-analysis. This is because each study tells us something about the distribution of the true effect sizes even if it has measured the true effect size for that study imprecisely. This, however, can be a disadvantage if smaller studies are of lower quality. The consequence is that the standard error of the summary effect will be larger under a random effects meta-analysis than under a fixed effect model.

Higgins *et al.* [237] proposed a statistic  $I^2$  that tells us about the proportion of the observed variance that relates to true differences in the effect size between studies

$$I^2 = \left( \frac{Q - (K - 1)}{Q} \right) \times 100\%.$$

Values of  $I^2$  close to 0% tell us that the true effect sizes are similar between studies and a value close to 100% tells us that the heterogeneity between studies dwarfs that within studies.

### 8.3.4 Other measures of $\tau^2$

Another popular method, although a lot more computationally intensive and less intuitive to non-statisticians, is the use of maximum-likelihood based procedures to estimate  $\tau^2$  and  $\mu$  simultaneously. The likelihood for  $\mu$  and  $\tau^2$ , assuming that the observed effect sizes  $Y_i$  are normally distributed, is given by

$$\log L(\mu, \tau^2 | \mathbf{Y}) = -\frac{1}{2} \sum_{i=1}^K \log(V_{Y_i} + \tau^2) - \frac{1}{2} \sum_{i=1}^K \frac{(Y_i - \mu)^2}{V_{Y_i} + \tau^2}$$

and can be maximised iteratively for  $\mu$  and  $\tau^2$ . Maximum likelihood estimates of variance parameters are known to be negatively biased in many situations [238]. The restricted maximum likelihood (REML) given by

$$\log L(\mu, \tau^2 | \mathbf{Y}) = -\frac{1}{2} \sum_{i=1}^K \log(V_{Y_i} + \tau^2) - \frac{1}{2} \log \sum_{i=1}^K \frac{1}{V_{Y_i} + \tau^2} - \frac{1}{2} \sum_{i=1}^K \frac{(Y_i - \mu)^2}{V_{Y_i} + \tau^2}$$

does not have this problem, and as above these equations are maximised iteratively.

Other estimators for  $\tau^2$  have been proposed including Hunter-Schmidt [239], Hedges [240], and empirical Bayes [241] estimators. Viechtbauer compared the bias and efficiency of these methods and concluded that the REML method is to be preferred, whilst the Hunter-Schmidt and maximum likelihood methods are to be avoided [238]. Although each method will typically produce a different estimate of  $\tau^2$  the effect this has on the overall effect size estimate is generally small.

Meta analytic methods have been implemented in most standard statistical software packages such as the `metan` [242] package in STATA and `meta` [243], `rmeta` [244], and `metafor` [245] packages in R.

## 8.4 Graphs for meta-analysis

Anzures and Higgins [246] give an overview of graphs that can be used for meta-analysis and some useful tips on presentation. Here we give details of the three most commonly used graphs.

### 8.4.1 Forest plot

The forest plot is used to display the effect size and confidence interval for each study in a meta-analysis along with the mean effect size. The effect size is usually on the horizontal axis. The size of the plotting symbol for the effect size in each study is usually proportional to the inverse variance of the study. This is a clever way to draw the eye to the studies with the most precisely estimated effect sizes since otherwise the eye is naturally drawn to those with the

---

largest confidence intervals which are less precisely estimated. The mean effect size is also usually plotted using a different symbol, typically a diamond, where the width of the symbol corresponds to the width of the confidence interval for the parameter of interest. This is to differentiate the mean effect size from those of the individual studies.

For illustration we show the forest plot obtained from fitting a Cox model with a linear predictor for FBG (adjusting for age, sex and other conventional cardiovascular risk factors and stratifying by trial arm where appropriate) to each of the studies within the ERFC in figure 8.1. We shall also use this example to illustrate the two other graphs below. We observe that fixed effect and random effects meta-analyses give very similar mean effect sizes in this case, only differing in the second decimal place, although the confidence interval for the random effects meta-analysis is larger. There is moderate heterogeneity in the effect sizes between studies,  $\hat{\tau}^2 = 0.13$ .

### 8.4.2 Funnel plot

The funnel plot allows us to assess publication bias. It is a scatterplot of effect sizes against a measure of study size, usually the inverse standard error of the effect sizes. Typically the  $y$ -axis is inverted and dotted lines showing the 95% confidence limits based on a fixed effect meta-analysis added, which gives an indication of the heterogeneity in the data. Asymmetry in the plot suggests that there is a relationship between effect size and its precision. This could result from subsets of studies having different effect sizes, publication bias, or a systematic difference between smaller and larger studies [246]. If asymmetry is present then this should be investigated to ascertain the probable cause, and the appropriateness of carrying out a meta-analysis on the whole set of effects should be considered. We illustrate this in figure 8.2 where we observe that the points show no clear asymmetry and the points lie almost within the confidence limits.

### 8.4.3 Galbraith plot

An alternative to the funnel plot is the Galbraith plot [247]. The Galbraith plot is a scatter plot of the effect sizes divided by their standard error, against the inverse of the standard error of each study. Under this plot better estimated effect sizes lie further from the origin. The mean fixed effect can be obtained by regressing the effect sizes divided by their standard error, on the inverse of the standard error where the model is constrained to have zero intercept. A test of zero intercept in the unconstrained model can indicate whether publication bias is present. The constrained line and its 95% confidence are normally added to the plot and aid in detecting heterogeneity. Anzures and Higgins suggest that this plot is used when there are more studies than can be sensibly displayed in a forest plot. This plot is illustrated in figure 8.3.

## 8. Meta-analysis

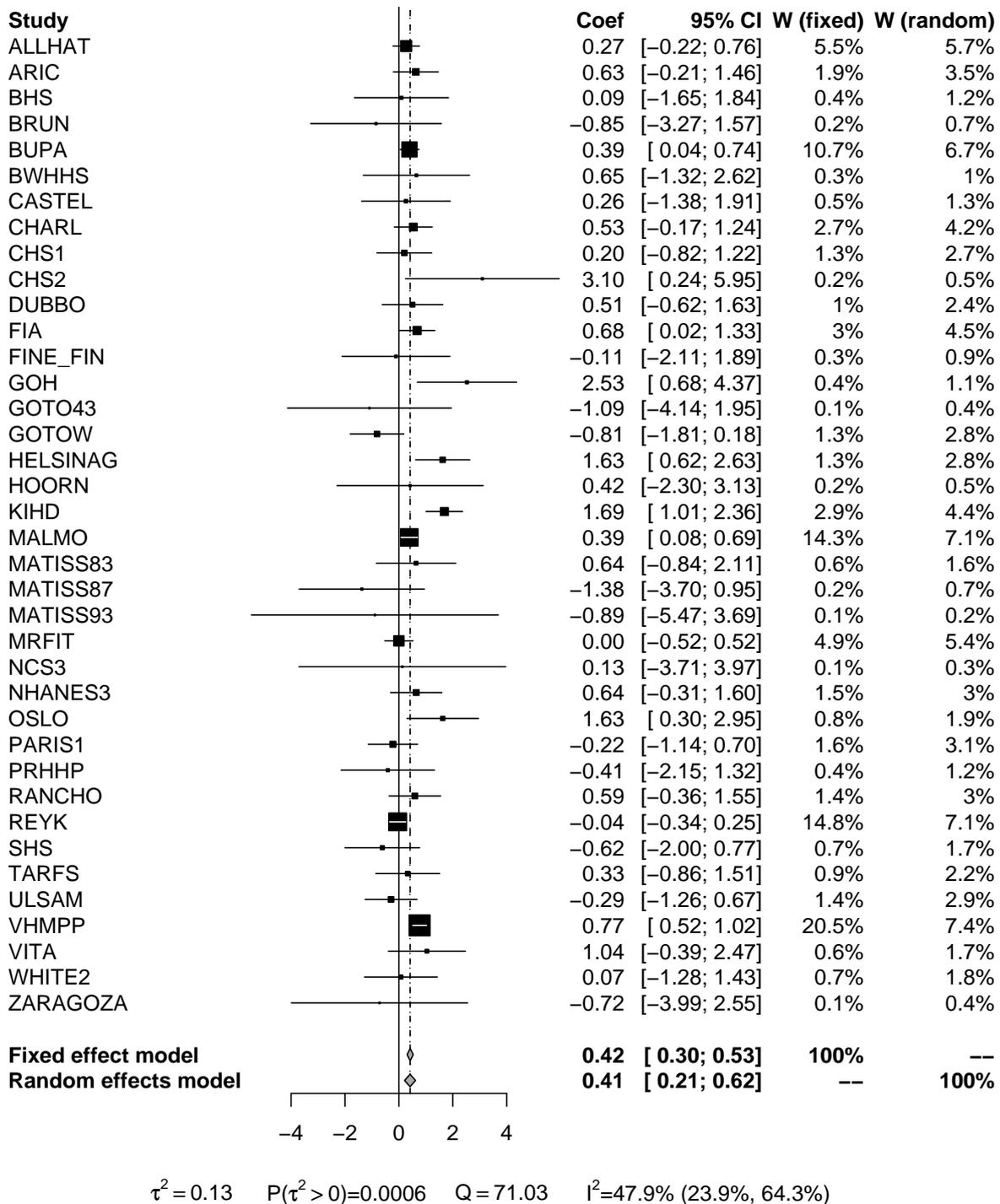


Figure 8.1: Forest plot for the slope parameter for a linear FBG-CHD relationship, with 95% confidence interval (CI) and fixed and random effects weights (W) for each study.

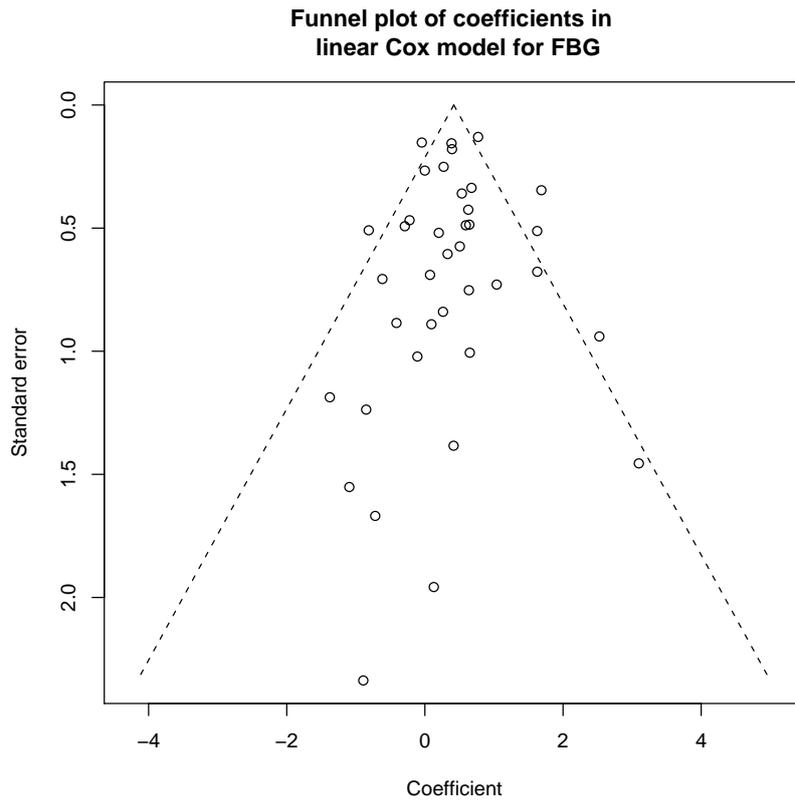


Figure 8.2: Funnel plot of the slope parameter for a linear FBG–CHD relationship.

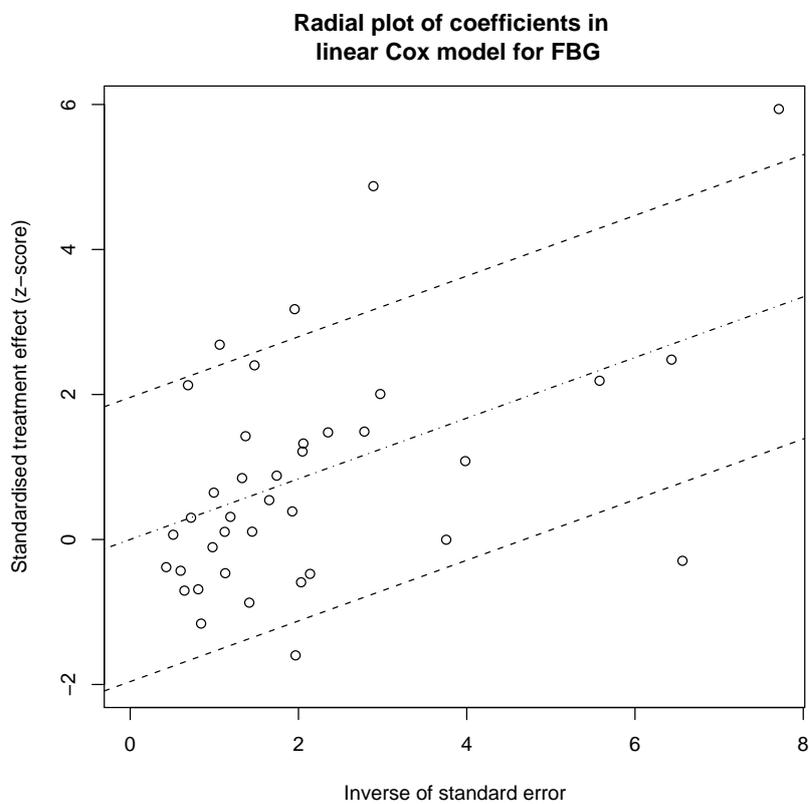


Figure 8.3: Galbraith plot for the slope parameter for a linear FBG–CHD relationship.

## 8.5 Multivariate meta-analysis

There are occasions when more than one parameter is of interest; for example Hartung, Knapp, and Sinha [248] analyse data from Kalaian and Becker where the two parameters of interest are performance on the Scholastic Aptitude Tests (SAT) mathematics and verbal reasoning examinations. In section 8.7.2 we shall be dealing with multiple model parameters from each study.

The easiest approach to meta-analysis of multiple parameters is simply to ignore the multivariate nature of the data and meta-analyse each parameter univariately. This ignores the information that is contained in the correlation between the parameters. Another approach is to form a measure that is the composite of the parameters [249]; this is clearly fraught with difficulties. A more appealing approach is to perform a multi-variate meta-analysis. A fixed effect approach was proposed by Raudenbush, Becker, and Kalaian [250] and implemented in the `gap` package [251] in R whereby the outcomes are stacked and are regressed using generalised least squares on a set of indicators, one for each parameter, which equal one if the indicator relates to that parameter and zero otherwise. To assume a fixed effect is a strong assumption in univariate meta-analysis but becomes even stronger in multivariate meta-analysis because it seems implausible that there is no between-study heterogeneity in any of the parameters [252]. Therefore, a true multivariate random effects meta-analysis approach is required.

Under the multivariate model, where we have  $m$  parameters, we assume that the parameter estimates from  $i$ th study  $\mathbf{Y}_i = \{Y_{i1}, \dots, Y_{im}\}$  are distributed according to a multivariate normal distribution

$$\mathbf{Y}_i \sim N(\boldsymbol{\theta}_i, \boldsymbol{\Delta}_i)$$

where the true effect size is also assumed to come from the multivariate normal distribution

$$\boldsymbol{\theta}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

For example, if we have estimates of two parameters,  $Y_1$  and  $Y_2$ , with within-study variance-covariance matrix for the  $i$ th study,

$$\boldsymbol{\Delta}_i = \begin{pmatrix} \sigma_{1i}^2 & \rho\sigma_{1i}\sigma_{2i} \\ \rho\sigma_{1i}\sigma_{2i} & \sigma_{2i}^2 \end{pmatrix}$$

and between-study variance-covariance matrix,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \tau_1^2 & \kappa\tau_1\tau_2 \\ \kappa\tau_1\tau_2 & \tau_2^2 \end{pmatrix}$$

then we get the bivariate normal model for the  $i$ th study

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 + \tau_1^2 & \rho_i \sigma_{1i} \sigma_{2i} + \kappa \tau_1 \tau_2 \\ \rho_i \sigma_{1i} \sigma_{2i} + \kappa \tau_1 \tau_2 & \sigma_{2i}^2 + \tau_2^2 \end{pmatrix} \right).$$

To estimate the parameters  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  we can maximise the restricted likelihood

$$\log \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{Y}) = -\frac{1}{2} \sum_{i=1}^K \log |\boldsymbol{\Sigma} + \boldsymbol{\Delta}_i| - \frac{1}{2} \log \left| \sum_{i=1}^K (\boldsymbol{\Sigma} + \boldsymbol{\Delta}_i)^{-1} \right| - \frac{1}{2} \sum_{i=1}^K (\mathbf{Y}_i - \boldsymbol{\mu})^T (\boldsymbol{\Sigma} + \boldsymbol{\Delta}_i)^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}).$$

Once we have obtained an estimate for  $\boldsymbol{\Sigma}$  we can calculate our overall effect size estimate  $\hat{\boldsymbol{\mu}}$  and its variance as

$$\hat{\boldsymbol{\mu}} = \left( \sum_{i=1}^K (\hat{\boldsymbol{\Sigma}} + \boldsymbol{\Delta}_i)^{-1} \right)^{-1} \left( (\hat{\boldsymbol{\Sigma}} + \boldsymbol{\Delta}_i)^{-1} \mathbf{Y}_i \right)$$

$$\widehat{\text{Var}}(\hat{\boldsymbol{\mu}}) = \left( \sum_{i=1}^K (\hat{\boldsymbol{\Sigma}} + \boldsymbol{\Delta}_i)^{-1} \right)^{-1}.$$

Maximising the restricted likelihood is computationally intensive, especially as the number of parameters increases. In section 8.7.2 we perform a multivariate meta-analysis on twelve parameters which would take considerable time using this approach. An alternative to REML is a DerSimonian and Laird based approach [252] which is computationally simpler, especially when the number of parameters is large. This approach is used in all of the multivariate meta-analyses that follow in this chapter. Under this method we calculate the  $Q$  matrix where the  $(j, k)$ th element (corresponding to the  $j$ th and  $k$ th outcomes) is given by

$$Q_{jk} = \sum_{i=1}^k \frac{(Y_{ji} - \bar{Y}_j)(Y_{ki} - \bar{Y}_k)}{\sigma_{ji} \sigma_{ki}}$$

where

$$\bar{Y}_j = \sum_{i=1}^K \frac{Y_{ji}}{\frac{1}{\sigma_{ji} \sigma_{ki}}}.$$

The expectation of  $Q_{ij}$  is given by

$$E(Q_{jk}) = \left( \sum_{i=1}^K \rho_j - \frac{\sum_{i=1}^K \frac{\rho_{ji}}{\sigma_{ji} \sigma_{ki}}}{\sum_{i=1}^K \frac{1}{\sigma_{ji} \sigma_{ki}}} \right) + \left( \sum_{i=1}^K \frac{1}{\sigma_{ji} \sigma_{ki}} - \frac{\sum_{i=1}^K \frac{1}{\sigma_{ji}^2 \sigma_{ki}^2}}{\sum_{i=1}^K \frac{1}{\sigma_{ki} \sigma_{ki}}} \right) \kappa \tau_j \tau_k$$

which is a linear function of the parameter of interest  $\kappa \tau_j \tau_k$ . Hence, by equating  $Q_{ij}$  with its expectation we can calculate  $\kappa \tau_j \tau_k$ . Note that this can be done pairwise which significantly reduces the amount of computation required.

In univariate meta-analysis the estimated between study heterogeneity  $\hat{\tau}^2$  is truncated so that

## 8. Meta-analysis

---

it takes a positive value. Similarly, in multivariate meta-analysis we need to ensure that our estimate of the between study heterogeneity matrix  $\hat{\Sigma}$  is positive semi-definite. This can be achieved by setting any negative eigenvalues of the spectral decomposition of  $\hat{\Sigma}$  to zero. Let  $\Phi = \{\phi_1, \dots, \phi_m\}$  be the eigenvalues of  $\Sigma$ , and  $e = \{e_1, \dots, e_m\}$  the corresponding eigenvectors then

$$\hat{\Sigma}_+ = \sum_{i=1}^m \max(0, \phi_i) e_i e_i^T$$

will be positive semi-definite as required.

When we have studies that have not measured all parameters, and if we can assume that the parameter values are missing at random, then we can replace the outcome measure with zero (or with any other arbitrarily chosen value) with an extremely large within-study variance, and zero covariance. This has the effect of weighting out those outcome measures that we did not observe. For example, pupils from a certain school may only have sat the mathematics paper of the SAT.

Both REML and DerSimonian and Laird approaches are implemented in `STATA` in the `mvmeta` package [253], and REML is implemented in the package of the same name in `R` [254], we give our own `R` code for both REML and DerSimonian and Laird approaches in appendix F.

Different ways of graphically presenting the results of a multivariate meta-analysis are required to those given for the univariate case in section 8.4. For example a bubbleplot can be used instead of a forest plot to display the parameter estimates from each study in a bivariate meta-analysis or two dimensions from a higher dimensional meta-analysis [255].

### 8.6 Individual participant data meta-analysis

IPD meta-analysis is the gold standard approach to combining data across multiple studies allowing us to control for confounders consistently between studies, carry out the same analysis on all studies using common criteria for inclusion/exclusion, and allowing us to perform additional analyses such as sub-group analyses that we would otherwise be unable to do. IPD meta analyses can lead to different conclusions being reached than if the meta-analysis had been conducted on aggregate information; for some examples see Riley *et al.* [62].

There are increasing numbers of large consortia such as the Emerging Risk Factors Collaboration [7] and the Asia Pacific Cohort Studies [13] as well as multi-centre studies such as the European Prospective Investigation into Cancer and Nutrition [256] which combine IPD from large numbers of studies/centres looking at similar research areas.

---

### 8.6.1 Advantages and disadvantages of IPD meta-analysis

Thompson [257] gave a table of some of the advantages that IPD analysis of data collated by large consortia allows, which we reproduce below:

- Enhanced comparability across studies
  - Ability to check and harmonise data to a common format
  - Use of standardised exposure, covariate and outcome definitions
  - Use of individual-level rather than study-level eligibility criteria
- Minimisation of biases
  - Ability to use common approaches to missing data
  - Allowance for measurement error
  - Use of consistent adjustment for potential confounders
- Greater insight
  - Ability to address novel hypotheses
  - Increased statistical power compared to individual studies, and the ability to assess associations with rarer outcomes
  - More precise estimation of associations under different circumstances
  - Detailed exploration of potential sources of heterogeneity with less vulnerability to ecological fallacies such as can occur with aggregated data
  - Ability to update and extend duration of follow-up
  - Permits more reliable sensitivity analyses to assess robustness of findings
- Consequential benefits
  - Provides a network of investigators to help further common research interests
  - Can help minimise duplication of research effort
  - Can help promote standardisation of research practices
  - Can stimulate advancement of methods that maximise the value of available data

There are some downsides to IPD meta-analysis. These include:

- Large IPD meta-analyses are expensive because of the need to employ data managers to look after the collation and harmonisation of the data from all the contributing studies.

- They can take considerable time because the data harmonisation process can be hindered by slow response to data queries with collaborating studies and because it is usual that papers resulting from such consortia are circulated amongst the collaborators before publishing.
- There can be issues regarding how authorship of manuscripts is attributed.
- Detail may be lost that was present in the original studies. For example smoking status may have been recorded in some studies as *current smoker*, *past smoker* or *never smoker* whereas some studies may have just recorded *current smoker* or *non-current smoker*. It may however be possible to perform analyses on a subgroup of studies if sufficient studies have records at the more detailed level. Biological measurements can also vary between studies due to the equipment or assay procedure used. Careful analysis of the results obtained from different methods is necessary to ensure that the measurements from different studies are compatible.

### 8.6.2 Approaches to IPD meta-analysis

There are two approaches to carrying out an IPD meta-analysis: the one stage approach and the two-stage approach. In the one-stage approach all the data are used in a single hierarchical model that takes account of the clustering within studies; for an example see Higgins *et al.* [258]. The random effects Cox model was developed by Ma *et al.* [259] and considered by Tudur Smith *et al.* [260]. These models can be fit in  $\mathbb{R}$  using the `coxme` [261] package. In the two-stage approach a model is fitted to each of the studies and then the estimates from these models are combined using the summary meta-analysis techniques described in section 8.3. These two approaches generally lead to similar results [262]. We consider a two-stage approach because of the computational complexity of fitting a one-stage model to large IPD [15].

In some situations it may not be possible to get the IPD from some studies, in which case aggregate level data may have to be used, Riley *et al.* [263] give a review of this area. The simplest approach is within a two-stage meta-analysis where the IPD are analysed in the first stage study by study and then combined with the aggregate level data in the second stage. This is not a problem that we have with the ERFC in which we have IPD available for all studies.

## 8.7 IPD meta-analysis for continuous exposure measures

In this section we consider approaches to meta-analysis of non-linear exposure–disease relationships; we start by considering combining grouped exposure analyses across studies, and then consider methods for combining continuous models for the exposure–disease relationship across studies.

---

### 8.7.1 Multivariate meta-analysis of point estimates

Perhaps the most obvious way to combine continuous exposure-disease relationships across studies is to discretise the continuous exposure by splitting the exposure into  $R$  groups, fit a grouped exposure analysis to each study to obtain parameter estimates  $\hat{\beta}_i = \{\hat{\beta}_{1i}, \dots, \hat{\beta}_{Ri}\}$  and associated variance-covariance matrix  $V_{\hat{\beta}_i}$ , and combine the results across studies using multivariate meta-analysis. This was the approach taken by White [253] who performed a multivariate meta-analysis of IPD on a continuous exposure, fibrinogen, and the risk of coronary heart disease. As we discussed in chapter 2 there are many problems with using grouped exposure analyses and we would prefer to model the exposure-disease relationship in each study using a continuous model.

### 8.7.2 Continuous approach

We now consider meta-analysis of continuous functions for the linear predictor. Sauerbrei and Royston [67] noted that of the 281 references in Sutton and Higgins' [264] review of recent developments in meta-analysis, none of them appeared to deal with a continuous exposure without using groups to discretise the problem. Although this was incorrect, as one of those references by Schwartz and Zanobetti [265] does in fact pool LOESS (locally weighted scatterplot smoothing) smooths across studies, this is a research area that has received relatively little attention.

Sauerbrei and Royston [67] have recently provided an approach to meta-analysis of fractional polynomial models allowing for heterogeneity in the shape of the exposure-disease relationship between studies. Their approach consists of two steps—model selection, and pooling of models across studies.

For the first step, Sauerbrei and Royston give three model selection rules for choosing a fractional polynomial model for each study. We discuss these rules and consider possible selection rules for P-spline models. For the second step they suggest a pointwise pooling rule to combine the relationship across studies allowing for heterogeneity in the shape of the relationship. We discuss this approach, and consider alternatives. We use simulated examples to illustrate some of the differences between the methods.

#### Model selection rules for fractional polynomial models

Sauerbrei and Royston's three approaches to obtaining a fractional polynomial model for each study within an IPD meta-analysis are:

- **Overall Model**

Select the best FP2 fit to the overall dataset, stratifying by study to allow for possible differences in baseline hazard between studies. We have much greater power to detect

## 8. Meta-analysis

---

non-linearity in the overall dataset so Sauerbrei and Royston suggest using a smaller nominal significance level (e.g. 0.01 or 0.001) so only models that are highly supported by the data are selected. We then use the powers from the selected fractional polynomial model and refit this model to each of the individual studies.

- **Studywise Model (fixed  $df$ )**

In each study the best fitting FP2 is chosen for each study.

- **Studywise (varying  $df$ )**

In each study the best fitting fractional polynomial model of maximum degree 2 is chosen for each study. Typically we will have lower power to detect non-linearity in individual studies so we should use a larger nominal significance level.

Fractional polynomial fitting requires that all the exposure values are positive. If this is not the case then a constant is added to the exposure values to ensure positivity (c.f. section 2.2.3). When using the second and third methods above this must be done with respect to the range of exposure values for *all* studies since otherwise we may get asymptotes in the range over which we wish to obtain fitted values from the model. The studywise (varying  $df$ ) selection rule was found by Sauerbrei and Royston to not work well, tending to give linear exposure-disease relationships in smaller studies which results in pooled relationships that do not exhibit much non-linearity, despite non-linearity being supported by the larger, more powerful studies. The flaw in this approach is that it does not allow us to borrow strength about the shape of the exposure–disease relationship across studies.

### **Model selection rules for P-spline models**

Splines have been used in meta-regression [147, 266] but do not appear to have been combined before in an IPD meta-analysis. We consider how we could select appropriate P-spline functions to combine in a meta-analysis as described above for fractional polynomials. We propose three different model selection rules for choosing the degrees of freedom for P-spline models which can be viewed as having similarities with the three rules for fractional polynomials:

---

- **Overall model**

We can fit the same model to each cohort. This can be achieved by setting the level of penalisation to be constant across models. We are then left with the choice of a suitable value for  $\lambda$ , the smoothing parameter. Using the value of  $\lambda$  which gives us four degrees of freedom in the model fitted to the whole dataset will lead to oversmoothing in the individual models. We found that using the smoothing parameter that gave four degrees of freedom to the largest study worked well. In this case the individual models for each cohort will have between one and four degrees of freedom. This differs from the fractional polynomial approach because under this method each study will have a different number of degrees of freedom.

- **Studywise model (fixed  $df$ )**

Fit a P-spline with four degrees of freedom to each study, this is directly comparable to the studywise FP2 approach where each study has approximately four degrees of freedom.

- **Studywise model (varying  $df$ )**

For each study fit a P-spline model where the degrees of freedom are chosen by a model selection criterion such as Hurvich's cAIC [108].

The overall model approach relies on us using the same basis for the model fitted to the whole dataset as well as the individual studies, since the smoothing parameter  $\lambda$  is specific to the basis. As discussed in section 2.2.4 Govindarajulu *et al.* [109] found constraining a P-spline model to have 4 degrees of freedom appeared to be preferable to selecting the number of degrees of freedom by AIC. We found that the studywise (varying  $df$ ) rule suffered from similar problems as with fractional polynomials, producing pooled relationships that did not exhibit much non-linearity.

## Pooling rules

Once a suitable model has been selected for each study we then need to consider how we can pool the models across studies. Firstly we describe the pointwise pooling rule of Sauerbrei and Royston, which was also used by Schwarz and Zanobetti [265], before considering two alternative methods:

- **Pointwise pooling**

At each point over a fine grid of exposure values we obtain a fitted value and standard error (relative to the reference value  $x_0$ ) from the model fitted to each of the individual studies. So that for the  $j$ th grid point in the  $i$ th study we have

$$y_{ij} = \hat{\beta}_i^T(x_j - x_0), \quad V_{ij} = (x_j - x_0)\mathbf{V}_{\hat{\beta}_i}(x_j - x_0)^T$$

## 8. Meta-analysis

---

where  $\hat{\beta}_i$  is the estimated parameter vector for the  $i$ th study and  $V_{\hat{\beta}_i}$  is its variance-covariance matrix. At each point  $j$  on the grid, we perform a univariate meta-analysis on  $\mathbf{y}_j = \{y_{1j}, \dots, y_{Kj}\}$  and  $V_j = \{V_{1j}, \dots, V_{Kj}\}$ . Sauerbrei and Royston suggest that either a fixed effect or random effects meta-analysis can be carried out; although as discussed in section 8.5 a random effects analysis is strongly preferred.

This is essentially a non-parametric approach as we only require fitted values over a grid of points and their associated standard errors for each study. Under this method we can combine studies where we have fitted models of differing functional form to each study. Combining fitted values pointwise is appealing because it uses the standard univariate meta-analysis methods which are commonly available in statistical software packages. The computation of DerSimonian and Laird random effects meta-analyses is very quick. Hence, the pooled relationship although requiring many calculations, can be produced quickly.

By choosing to meta-analyse the fitted functions pointwise we are reduced to a graphical display of the results as we lose the ability to write down the fitted function in a concise manner. However, one of the advantages noted in chapter 2 of fractional polynomials fitted to individual studies is that they can be written concisely. This is not an issue when pooling P-spline models because they cannot be written compactly. When combining the results across studies in a pointwise manner the degree of heterogeneity between studies is allowed to vary across the range of exposure, as we shall shortly demonstrate.

A major criticism of pointwise meta-analysis is that the relationship obtained for each study is given with respect to an arbitrarily chosen reference value. Sauerbrei and Royston note that the results depend on this choice and advise that it ‘should be chosen sensibly’ [67]. The choice of reference value can lead not just to a trivial vertical shift of the graph, but to a change in the shape of the estimated exposure–disease relationship as we shall see in the examples that follow.

One case where the choice of reference value is not important is under a fixed effects meta-analysis using pointwise pooling, where the predictor contains a single term e.g. an FPI model. Since then for all pairs of studies  $i$  and  $j$  and all choices of reference value  $x_0$ ,

$$\frac{y_i}{y_j} = \frac{\beta_i(x - x_0)}{\beta_j(x - x_0)} = \frac{\beta_i}{\beta_j}$$

and

$$\frac{V_i}{V_j} = \frac{V_i(x - x_0)^2}{V_j(x - x_0)^2} = \frac{V_i}{V_j}$$

which are both independent of  $x_0$ . Since the ratio of all pairs of effect sizes and their variances are independent of the choice of reference value, the estimated exposure–disease relationship is also independent of the reference value. If the predictor contains more than one term, or we are using random effects meta-analysis, then we do not get the cancellation observed above and it

seems unlikely that there will not be some difference in the point estimates and their variances dependent on the choice of reference value. It is unsatisfactory that the arbitrary choice of reference value should affect the shape of the pooled relationship, although in many practical situations the difference may be small.

- **Pointwise derivative pooling**

A solution to the reference value problem is to work with the derivative of the fitted function instead of the fitted function itself as this is independent of the choice of reference value. We can calculate the derivative of the fitted log hazard ratio and its standard deviation over a suitably fine grid of points either analytically, as discussed in section 5.3.1, or empirically. For the  $j$ th grid point in the  $i$ th study we have

$$y_{ij} = \hat{\beta}_i^T D_{xi}, \quad V_{y_{ij}} = D_{xi} \mathbf{V}_{\hat{\beta}_i} D_{xi}^T$$

where  $D_{xi}$  is the derivative of the design matrix for the  $i$ th study. At each point  $j$  on the grid, we can perform a univariate meta-analysis on  $\mathbf{y}_j = \{y_{1j}, \dots, y_{kj}\}$  and  $V_j = \{V_{1j}, \dots, V_{kj}\}$ . The pooled derivative can be used to find the implied relationship. We can then choose a suitable reference value e.g. the mean exposure; and whatever the choice we shall obtain the same shape for the exposure–disease relationship.

This approach has the same problem with lack of a concise way to write down the resultant mean function. From the pointwise meta-analysis of the derivative we only obtain the variance of the derivative of the fitted log hazard ratio, and not of the implied relationship. We have not found a method for analytically obtaining the pointwise standard error of the implied relationship; bootstrap resampling (within studies) can be used, however, this is computationally intensive.

- **Parameter pooling**

This approach is mentioned by Sauerbrei and Royston in their discussion but they do not apply this approach. It is the approach taken by Rota *et al.* [267] in their meta-regression of the relationship between alcohol and oesophageal squamous cell carcinoma. This approach can only be applied if the linear-predictor of the model fitted to each study is the same. We take the parameter estimates  $\mathbf{y}_i = \hat{\beta}_i$  and their variance-covariance matrices  $\mathbf{V}_i = \mathbf{V}_{\hat{\beta}_i}$  from each of the studies and combine them using multivariate meta-analysis to obtain the average parameter estimates and associated variance-covariance matrix. Again this approach is not dependent on the choice of reference value. Another advantage of this strategy is that we obtain models that can be written down in a concise manner as they are of the same parametric form as the constituent studies.

We now investigate some of the properties of the three pooling rules using three simulated examples. These are not intended to be serious examples of meta-analyses that might be carried

## 8. Meta-analysis

---

out in practice, but are aimed at illustrating some of the differences between the methods.

### Example 1

The aim of this example is to highlight the differences between the methods when we have studies with different exposure ranges. The following data were generated:

Study 1: 50,000 observations generated from a normal distribution with mean 3 and standard deviation of 0.6. The exposure–disease relationship was generated from a true linear exposure–disease relationship with a slope of 0.15 and an exponential baseline hazard  $h_0(t) = 0.01$ .

Study 2: 150,000 observations generated from a normal distribution with mean 3 and standard deviation of 0.2. The exposure–disease relationship was generated from a true linear exposure–disease relationship with a slope of 0.8 and an exponential baseline hazard  $h_0(t) = 0.01$ .

A P-spline model with 4 degrees of freedom and 17 knots was fitted to both of these studies. Each of the three pooling rules described above was used to combine the two studies. Fixed effect analyses were used as there are only two studies. The results from this example can be seen in figure 8.4.

The pointwise meta-analysis leads to Simpson’s paradox being exhibited. There are two regions, 2.3–2.5 units of exposure and 3.5–3.7 units of exposure, in which both of the studies give an increasing/decreasing hazard ratio but the average effect size is decreasing/increasing. This is a result of the confidence intervals for Study 1 ‘blowing up’ outside the range of the data and Study 2 receiving more weight; we would not see this effect with fractional polynomials. We get a similar, although less pronounced effect, with the parameter based meta-analysis. This is because each parameter corresponds to a basis function which has an influence over a localised range. The pointwise derivative pooling rule does not show this effect and increases across the range of exposure which is consistent with the individual studies.

### Example 2

The aim of this example is to highlight the different sized confidence intervals that can be obtained under each of the methods resulting from different assumptions about how the shape of the exposure–disease relationship varies between studies. The following data were generated:

Study 1: 70,000 observations generated from a normal distribution with mean 2 and standard deviation of 0.4. The exposure–disease relationship was generated from a true J-shaped relationship given by  $0.3(x - 2.5)^2$  with an exponential baseline hazard  $h_0(t) = 0.005$ .

Study 2: 70,000 observations generated from a normal distribution with mean 4 and standard deviation of 0.4. The exposure–disease relationship was generated from the same relationship as Study 1.

Here we use the overall model selection rule. The best fitting FP2 to the two studies combined (stratified by study) was chosen which gave powers of 0.5 and 2. This model was then fitted

to each study separately and each of the three pooling rules were applied using a fixed effect analysis in each case.

All three methods give similar pooled relationships (figure 8.5), however the confidence intervals under the parameter pooling rule are considerably smaller than for the other two methods. Under the parameter pooling rule we assume that the shape of the relationship is known and that we are interested only in estimating the parameters  $(\beta_1, \beta_2)$ . Each study, independent of the range of exposure values, gives us information about these parameters. The confidence intervals under the pointwise pooling rules are much larger because each study tells us very little about the hazard ratio at the reference value of 3 and this is reflected in the larger confidence intervals. Note how the choice of reference value obscures this.

### Example 1

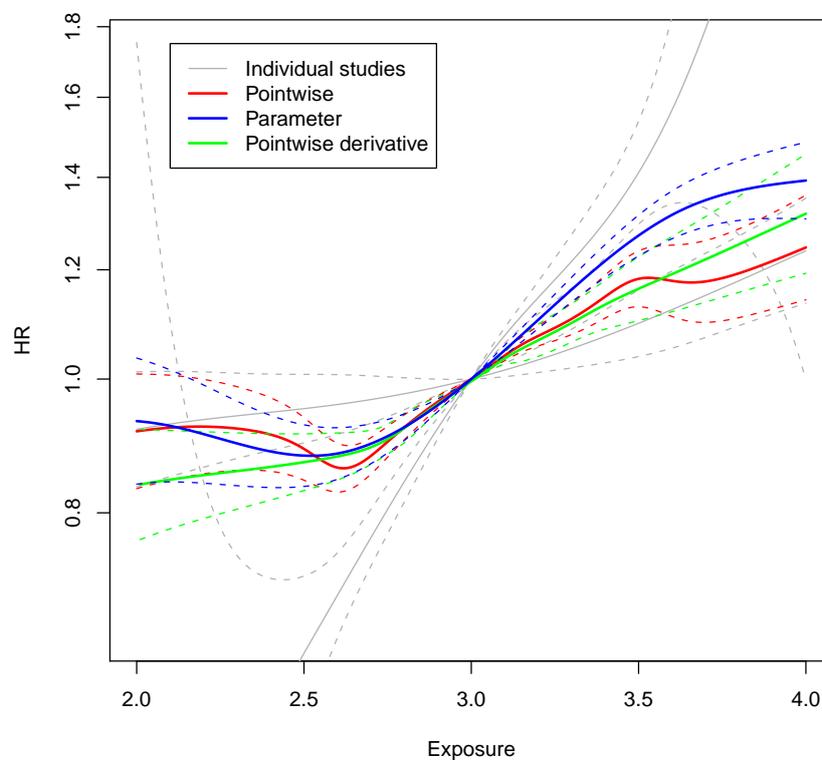


Figure 8.4: Example 1 — Pooled exposure–disease relationship under each of the three pooling rules for simulated data.

### Example 3

The aim of this example is to highlight that different choices of reference value can lead to different shapes for the exposure–disease relationship being obtained under the pointwise pooling approach. The following data were generated:

Study 1: 100,000 observations generated from a normal distribution with mean 2.5 and standard deviation of 0.3. The exposure–disease relationship was generated from a true threshold shaped relationship given by  $(X - 3)I(X > 3)$  with an exponential baseline hazard  $h_0(t) = 0.03$ .

## Example 2

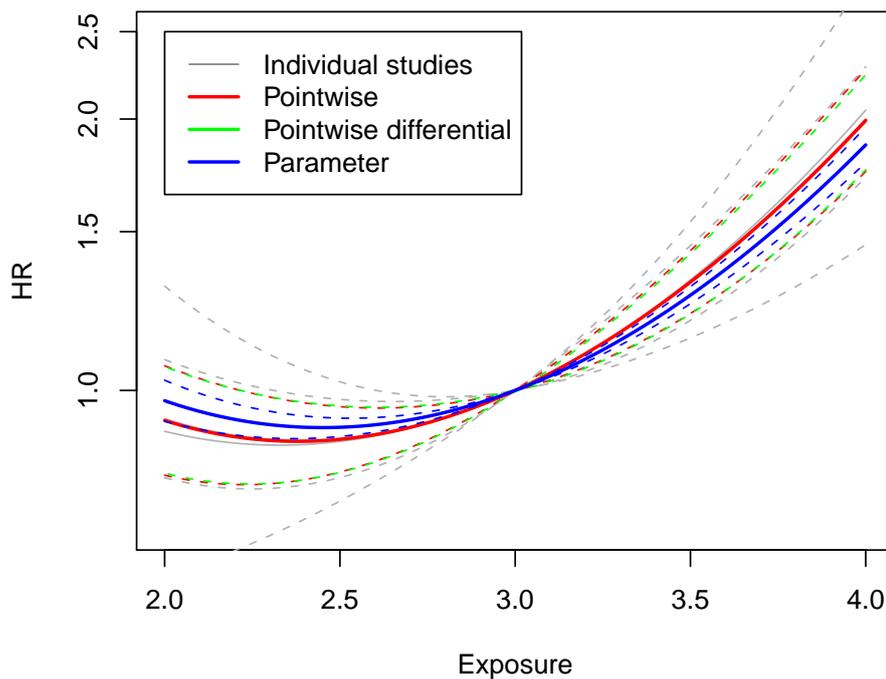


Figure 8.5: Example 2 — Pooled exposure–disease relationship under each of the three pooling rules for simulated data. Note that in this example the pointwise and pointwise differential methods give identical results.

Study 2: 100,000 observations generated from a normal distribution with mean 3.5 and standard deviation of 0.3. The exposure–disease relationship was generated from a true threshold shaped relationship given by  $(3 - X)I(X < 3)$  with an exponential baseline hazard  $h_0(t) = 0.03$ .

A P-spline model with 4 degrees of freedom and 17 knots was fitted to both of these studies. Each of the three pooling rules was used to combine the two studies using fixed effects meta-analysis. In figure 8.6 we see that the pooled relationship under the pointwise pooling rule is clearly dependent on the choice of the reference value; in fact the relationship obtained under the two choices of reference value considered are near mirror images of each other. As in example 1, the pooled relationship is going in the opposite direction to the relationships in the individual studies i.e. below exposure of 2.6 units when the reference value is 3.5 units, and above 3.5 units when the reference value is 2.5 units. Under the pointwise derivative and parameter pooling rules there is essentially no relationship between exposure and disease under both choices of reference value. Obviously this example is pretty extreme and in practice you would probably not try to combine studies that exhibited completely opposite trends. It does, however, serve to highlight the pointwise pooling rule’s dependence on the choice of reference value even if the reference value is ‘chosen sensibly’.

---

## Graphical presentation of continuous meta-analyses

Sauerbrei and Royston suggest plotting the fitted function for each of the studies on a single plot. This allows us to visually assess heterogeneity between studies in much the same way as with a forest plot. There are, however, two main downsides to this plot. Firstly, there is no way to distinguish between large cohorts whose model parameters will be well estimated and small cohorts which will be subject to a greater amount of uncertainty, and secondly as the number of cohorts increases so does the number of lines on our graph which can make it difficult to interpret. Sauerbrei and Royston also plot the weight each study receives under the pointwise pooling rule across the exposure range. Although we can see how the weights vary between studies, it is hard to deduce the effect of the change in weight on the pooled relationship.

We propose combining these two plots to aid interpretability, and save page space. We propose that the saturation of the colour of the line in the plot of the fitted function from each study depends on its weight in the meta-analysis at that point. This allows us to see which cohorts are contributing most to the overall shape of the exposure–disease relationship at each level of exposure. This will allow us to consider heterogeneity in the shape of the relationship between studies but also we may wish to investigate heterogeneity in regions of exposure e.g. Why does study ... have such a large weight in this range of exposure?

In the case that we use the parameter pooling rule with fractional polynomials then we can use the bubbleplot approach described in section 8.5 to display the results, although we found that if the variance in one parameter is much larger than the variance in the second then a bubbleplot does not give a good presentation of the data. When we use P-spline models we will have many coefficients making this approach impractical.

### Heterogeneity

Heterogeneity between studies can be investigated under the pointwise pooling rule by plotting how  $\hat{\tau}^2$  varies across the exposure range. This will be dependent on the choice of reference value.  $\tau^2 = 0$  at the reference value although this is purely artificial, and we would expect it to increase as we get further from the reference value since the pointwise variances generally increase as we move away from the reference value. This plot will be more meaningful under the pointwise derivative pooling rule because it does not depend on the choice of reference value. A better way of visualising heterogeneity across the exposure range may be to use Higgin's  $I^2$  statistic because it is independent of scale.

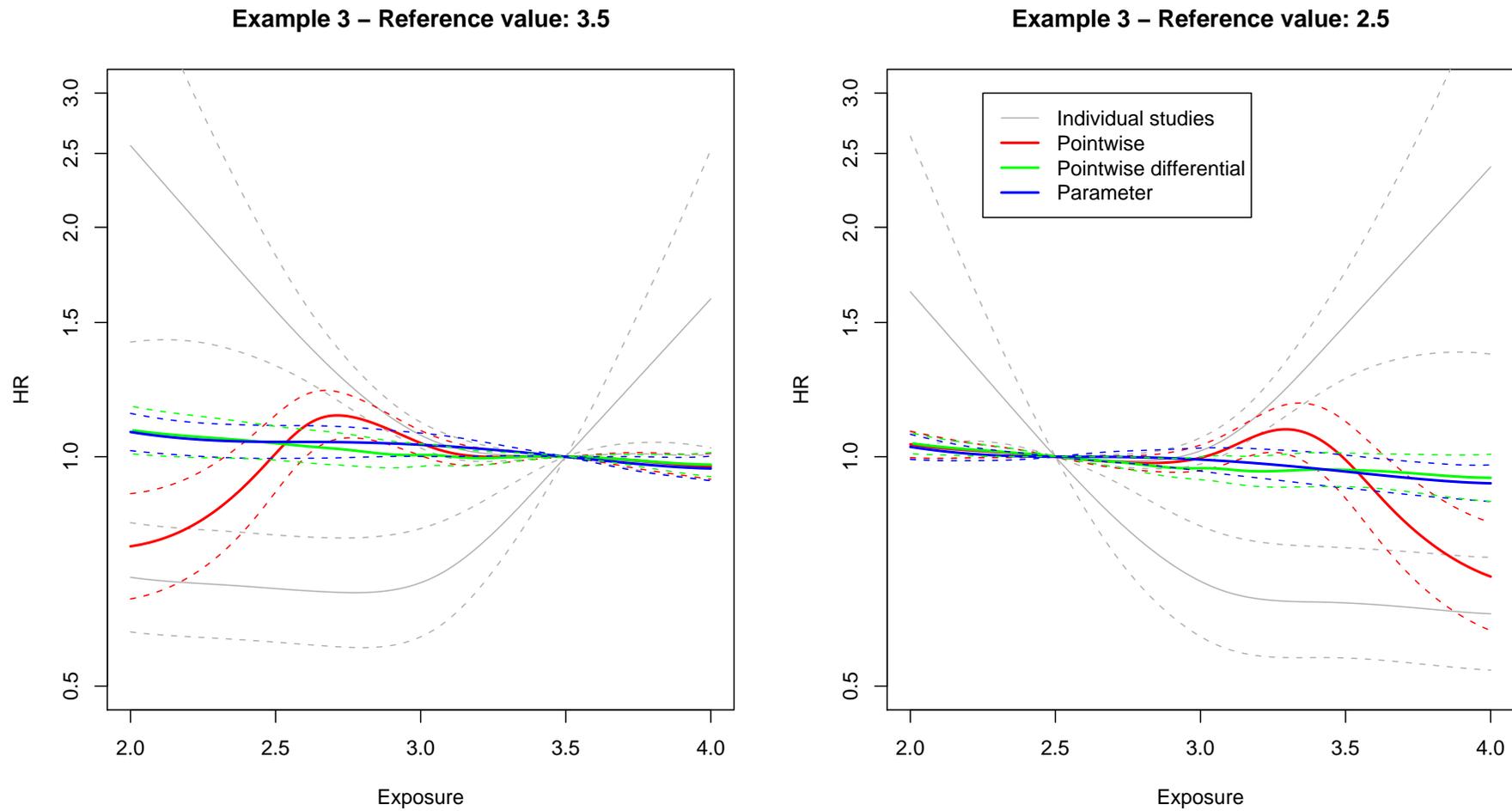


Figure 8.6: Example 3 - Pooled exposure–disease relationship under each of the three pooling rules for simulated data with the reference value at 3.5 units of exposure (left hand panel) and 2.5 units of exposure (right hand panel).

## 8.8 Application of methods to the ERFC FBG data

We now apply the methods introduced in section 8.7.2 to the FBG–CHD relationship using the ERFC data that were introduced in chapter 6. Each of the analyses that follow are adjusted for the confounding effects of BMI, smoking status (current smoker vs. non-current smoker), total cholesterol, and systolic blood pressure.

### 8.8.1 Relationship uncorrected for the effects of measurement error

Firstly, we demonstrate a meta-analysis of a grouped exposure analysis. Within each cohort we fit a grouped exposure analysis using the groups defined in section 6.4.1, and combined the studies using multivariate meta-analysis. The result is shown in figure 8.7. We have used the mean of all observations in each group across all studies to choose the  $x$ -axis plotting locations and we have selected the third group as the reference to prevent perfect prediction in the reference group within individual studies. Perfect prediction occurs when a group contains observations but no events. We observe essentially the same relationship as in the third panel of figure 6.2 although the middle group (FBG of 5.5–6 mmol/l) now lies below the 6th group (FBG of 6–6.5 mmol/l).

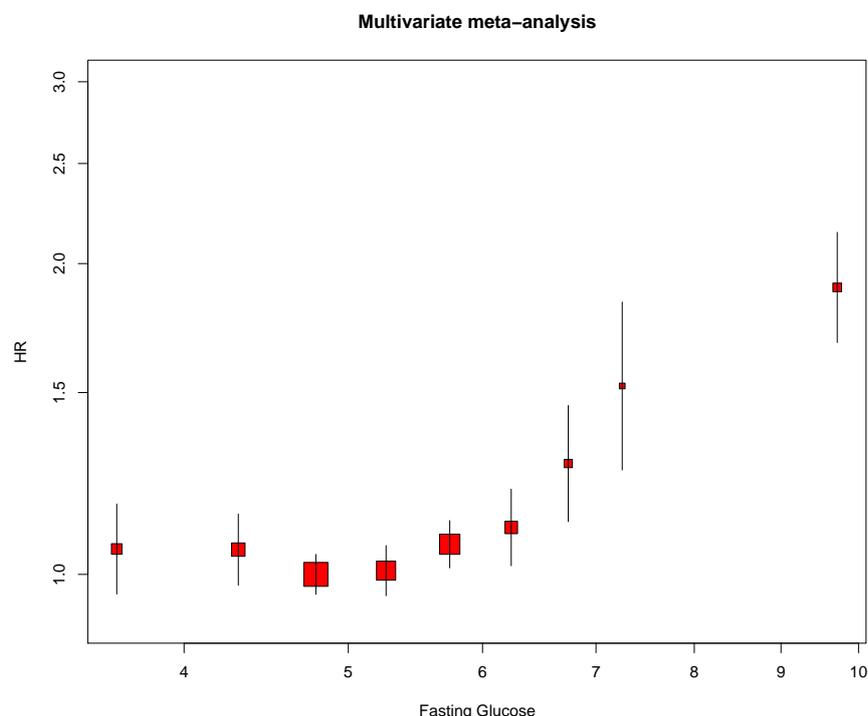


Figure 8.7: Multivariate meta-analysis of a grouped exposure analysis of the observed FBG–CHD relationship, with quasi-variance based confidence intervals.

In figure 8.8 we apply each of the three model selection rules for P-splines to the relationship between FBG and CHD using the pointwise pooling rule. When we use the overall model

## 8. Meta-analysis

---

selection rule, the shape of the relationship (1st row, 2nd column) is similar to that when we applied a model stratified by cohort to the whole dataset, although the hazard is slightly lower at higher levels of FBG. At the lowest levels of FBG, a small number of studies have a large amount of the weight (1st row, 1st column) whilst above the mean of FBG the weights are distributed over a number of studies. Almost all of these show increasing risk with increasing FBG although they differ substantially in how strong this relationship is.

Under the studywise model with  $4df$  (2nd row, 2nd column) we find that the shape of the relationship is broadly similar, although we get a ‘kink’ at an FBG of 9 mmol/l. We can see (2nd row, 1st column) that this is caused by one study gaining greater weight between 9 and 10 mmol/l. We see that the weights are concentrated on a much smaller number of studies.

In the last row of figure 8.8, where we have used the studywise (varying  $df$ ) pooling rule we see that although we get a shape of relationship that seems plausible in the middle of the range of FBG values we get a very linear relationship in the tails; most of the weight is concentrated on a very small number of studies. As noted above this method does not take advantage of the fact that we can borrow strength across studies.

Figure 8.9 displays the same information as the top left hand panel of figure 8.8 but using the plots of Sauerbrei and Royston. On the left hand side we can see the shape of the relationship from each of the studies but we have little idea as to which curves are the ones which tell us most about the shape of the relationship. If we consider the two plots concurrently then we can identify which curves have the most weight and where, but with a large number of studies, even when we use the same colour coding in the two plots, it is not easy to match the studies with the most weight in the right hand plot to the relationship in the left hand plot. In contrast in the left hand column of figure 8.8 our attention is immediately drawn to those studies which have most influence on the pooled relationship.

Figure 8.10 shows the result of applying each of the three pooling rules to a fractional polynomial analysis of the FBG–CHD relationship where we have used the overall model selection rule. Although in section 8.7.2 we showed that the different pooling rules could provide us with different shapes and confidence intervals for the relationship, we see in this example that there is little difference between the three rules. All three methods suggest that the relationship is slightly stronger above the mean of FBG than when we fitted a model stratified by cohort to the whole dataset.

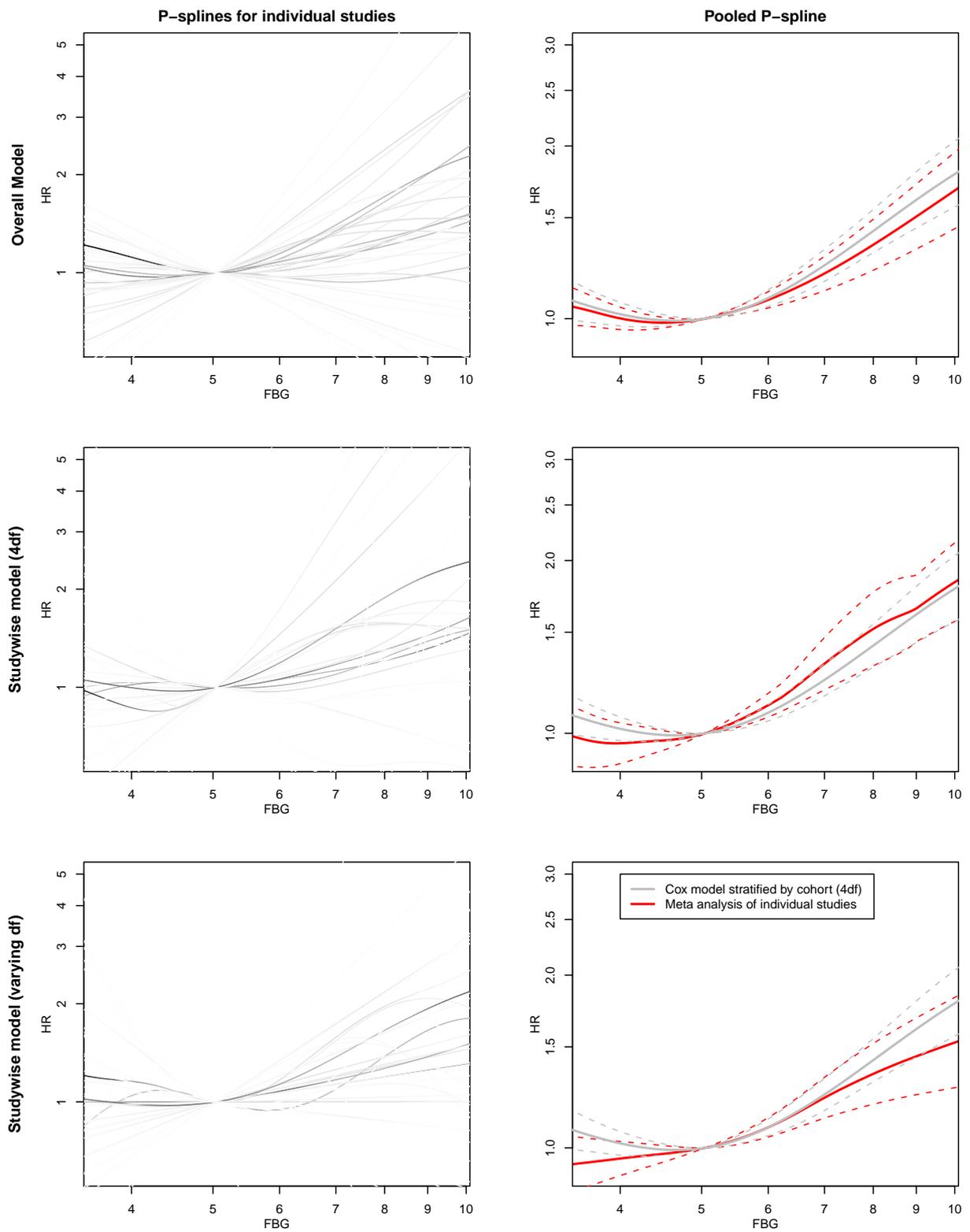


Figure 8.8: P-spline models of the FBG–CHD relationship under each of the three model selection rules. Left hand column: relationships in individual studies where saturation of line represents weight in meta-analysis, right hand column: pointwise meta-analysis of relationships, with stratified model ( $4df$ ) for comparison.

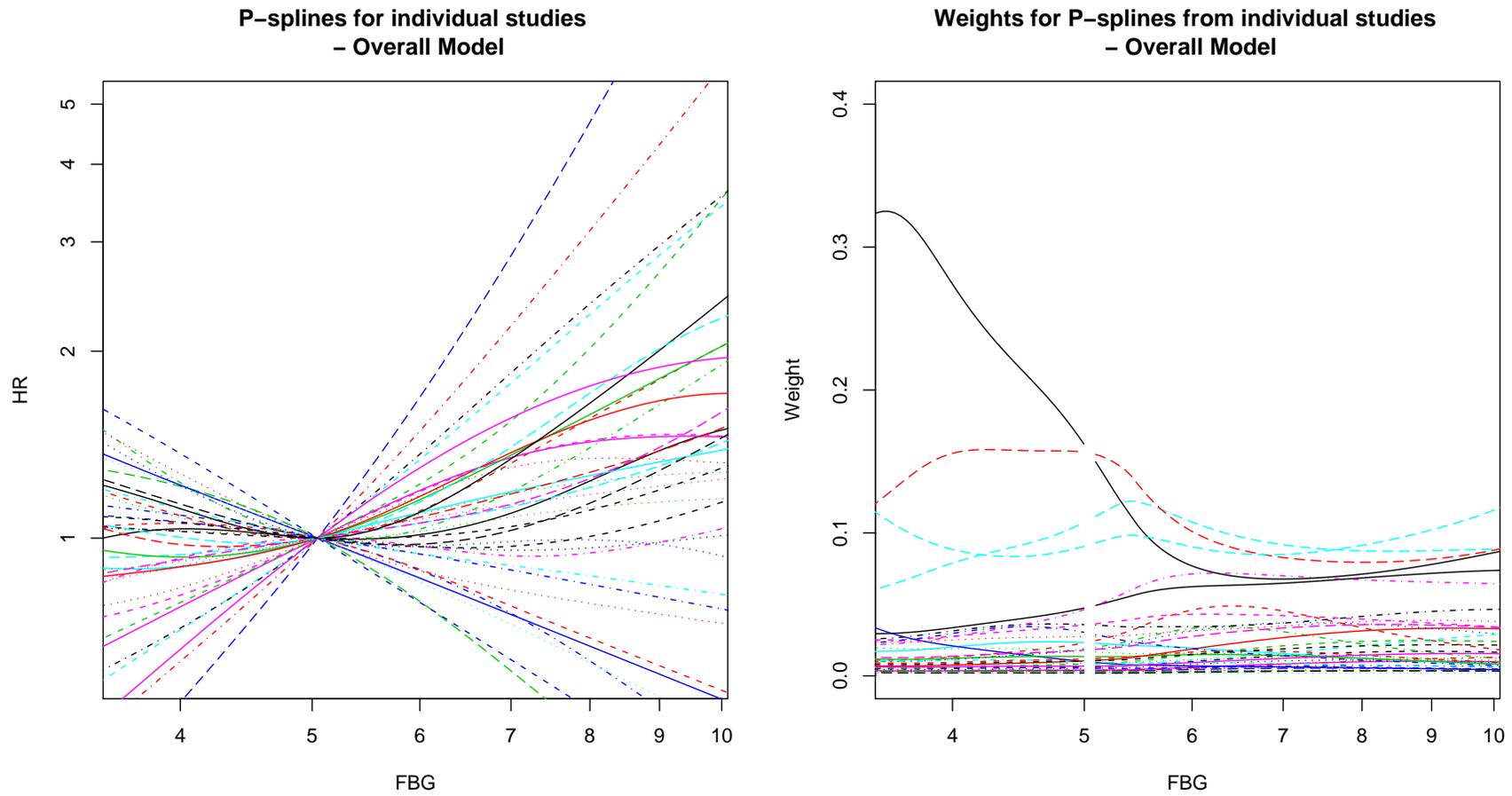


Figure 8.9: P-spline models for the FBG-CHD relationship in individual studies using the Overall model selection rule (left hand panel) and weights of these models in a pointwise random effects meta-analysis across the range of FBG (right hand panel).

### Fractional Polynomials – Overall model

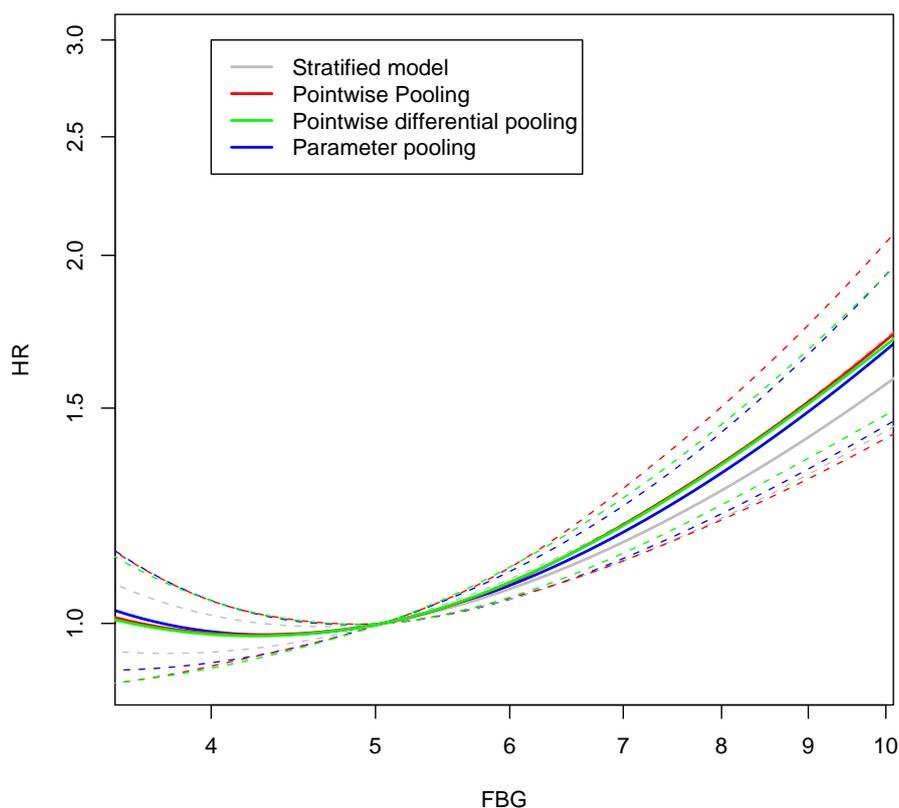


Figure 8.10: Random effects meta-analysis of the FBG–CHD relationship using fractional polynomial models for each study chosen by the overall model selection rule and pooled using each of the pooling rules, plus the relationship obtained by fitting a model to the combined dataset stratified by study.

In figure 8.11 we consider how the heterogeneity under the two pointwise methods varies across the range of FBG. Firstly, we note that the two curves in the left hand panel are not directly comparable. We see that under the pointwise pooling method  $\tau^2$  increases as we move away from the reference value, however as discussed previously this is partially an artefact of the choice of reference value. Under the pointwise derivative pooling  $\tau^2$  increases sharply in both directions as we move away from the centre of the distribution of FBG values of around 4.5 to 6 mmol/l. The right hand panel of 8.11 is much more useful because the measures are on a ratio scale meaning that they are directly comparable (except at the reference value). The pointwise pooling rule suggests there is less heterogeneity across much of the range of FBG compared with the pointwise derivative rule.

### 8.8.2 Relationship corrected for the effects of measurement error

This chapter has provided us with the final pieces to be able to complete the jigsaw, and to provide an estimate of the FBG–CHD relationship that both allows for the effects of exposure measurement error and for the heterogeneity in the shape of the relationship between studies. We can obtain the relationship using a 3 stage model:

## 8. Meta-analysis

---

Stage 1 Fit a measurement error model to each study using regression calibration for each study with repeat exposure measurements available and use the average of the parameters from these models as the measurement error model for studies without repeats.

Stage 2 Fit the analysis model (such as a structural fractional polynomial or P-spline) using one of the model selection rules from section 8.7.2 and the results from stage 1 for each study.

Stage 3 Perform meta-analysis on the output from stage 2 using one of the pooling methods described in section 8.7.2.

There are a number of possible choices and variations in each of these three stages. We choose the following combination of disease model, model selection rule, and pooling rule:

- Structural P-spline analysis — We choose to use a P-spline based analysis because they allow a good fit to local features of the data compared with fractional polynomials which give a more global fit.
- Overall model — We saw in figure 8.8 that studywise models gave very linear associations in the case of allowing the number of degrees of freedom to vary between studies and are variable in the case that we allow each study  $4df$ . As Sauerbrei and Royston found, the overall model gave the most consistent shape for the exposure–disease relationship.
- Pointwise derivative pooling — In example 2 of section 8.7.2 we saw that this approach is free of the choice of reference value and does not display Simpson’s paradox for P-splines.

In figure 8.12 we see our best estimate of the FBG–CHD relationship corrected for the effects of measurement error. Although FBG is subject to substantial measurement error this appears to result in only a small but relatively constant increase in the hazard ratio above the reference value. Below the reference value there is a much greater difference between the corrected and uncorrected relationship; FBG below the reference value appears to offer some protection against CHD with the minimum hazard occurring at an FBG of about 4 mmol/l. Firstly, we note that if we had chosen a reference value of 4 mmol/l then the difference between the corrected and uncorrected analyses would have appeared to be much greater. Secondly, there are individual studies that exhibit regions where the hazard ratio decreases with increasing blood glucose levels. Measurement error correction increases the magnitude of these decreases and this will partially offset increases in hazard ratios associated with measurement error correction in those studies with increasing relationships. This will lead to a pooled measurement error corrected relationship that exhibits a smaller difference when compared to the uncorrected relationship than might be expected.

---

## 8.9 Discussion

The FBG–CHD relationship corrected for measurement error and allowing for heterogeneity between studies is J-shaped with the nadir occurring at about 4 mmol/dl. As mentioned in chapter 6, low levels of exposure were also observed to confer protection from CHD by the APCSC. The pointwise pooling rule suggested that there was moderate heterogeneity between studies across the range of exposure. Each of the three pooling rules gave similar results for the FBG–CHD relationship in figure 8.10. An area for future research is to perform an empirical evaluation of the three pooling rules to identify which of the three methods performs best, and to investigate whether the limitations of the simple pointwise pooling rule highlighted in section 8.7.2 are likely to be exhibited in practice.

A key question regarding the model selection rules is how many degrees of freedom should each study have? Under the overall model selection procedure although we use 4 degrees of freedom in choosing the overall model, in the individual studies we are only fitting a model with approximately two degrees of freedom to the individual studies in the case of fractional polynomials, or between 1 and 4 for P-splines. Since the model for each study is fitted using a small number of degrees of freedom this does not penalise the smaller studies so severely. Under the studywise (fixed  $df$ ) model selection rule with  $4df$ , the meta-analysis gave most of the weight to a small number of large studies. We suggest that smaller studies may be better modelled with fewer degrees of freedom so that they are not completely dominated by the largest studies. This remains an area for future research. Another area for future investigation could be to investigate using different numbers of  $df$  for P-spline models. Although the studywise (varying  $df$ ) model selection rule initially sounds like a good idea, in practice it fails to borrow strength across studies and therefore we recommend that it should not be used.

The pointwise pooling rule can lead to Simpson’s paradox being displayed and is dependent on the choice of reference value. The confidence intervals for fractional polynomials do not become large outside the range of the data, hence large studies may still have considerable influence on the shape of the relationship under the pointwise pooling rule well beyond the range of the data. With P-spline analyses the confidence intervals do ‘blow-up’ and therefore studies only contribute significantly to the pooled relationship in the region where the exposure for that study lies. The parameter pooling rule is free from the choice of reference value but we require that the model for each study is of the same form. We saw that we can obtain much smaller confidence intervals from assuming that the same parametric form is suitable for the whole exposure range, even those areas where we might not have much information from the individual studies. The pointwise derivative pooling rule is also independent of the choice of reference and makes no assumption about the shape of the exposure–disease relationship, and was the only method of the three considered to give an increasing relationship when each of the individual studies gave an increasing relationship in example 1 of section 8.7.2. This

## 8. Meta-analysis

---

is why we prefer this method over the other two approaches. The downside to the pointwise derivative pooling rule is that pointwise confidence intervals can only be obtained using the bootstrap which is computationally intensive. Despite the differences between the methods, in practice the different pooling methods may produce similar shapes for the exposure–disease relationship, as was seen in section 8.8.1.

A related approach to the multivariate meta-analysis of grouped exposure analyses discussed in section 8.7.1, is to:

1. Fit a continuous model for the exposure–disease relationship to each study, for example a fractional polynomial or P-spline model.
2. Select a small set of exposure values  $x^* = \{x_1^*, \dots, x_m^*\}$ , and for each study  $i$  find the fitted values  $Y_i(x^*)$ , and the estimated variance-covariance matrix of the fitted values  $V_{Y_i}(x^*)$ .
3. Combine  $Y_i(x^*)$  across studies using multivariate meta-analysis.

This approach was not successful when we tried to implement it; the variance-covariance matrices for the individual studies were computationally singular even for small  $m$ , causing the meta-analysis to fail.

Although we have suggested model selection rules for P-splines, and have emphasised the importance of using pooling rules that are independent of the choice of reference value, we believe that further research into meta-analysis of continuous exposure-disease relationships is required. This could allow us to make more principled rather than pragmatic choices with regards to model selection and pooling rules.

### 8.10 Conclusion

Meta-analysis is an approach to combining data from multiple studies that address similar research hypotheses. IPD meta-analysis is the gold standard approach as we can analyse the data across studies in a consistent manner. Meta-analysis for continuous exposures without resorting to cutpoints is a relatively new area of research.

We have discussed Sauerbrei and Royston method for IPD meta-analysis of fractional polynomials allowing for heterogeneity in the shape of the exposure–disease relationship between studies. Sauerbrei and Royston [67] proposed three model selection rules for choosing a model for each study. We proposed similar model selection rules for P-splines.

We considered Sauerbrei and Royston’s pointwise pooling rule to combine the relationships across studies to obtain the mean exposure–disease relationship, as well as two alternative pooling rules: a parameter pooling rule and a novel pointwise derivative pooling rule. The alternative pooling rules are independent of the choice of reference value, whereas in general

---

the pointwise pooling rule is not. Although in our artificial examples the three pooling rules gave differing results, when we applied them to an analysis of the ERFC FBG data uncorrected for measurement error using the overall model selection rule they gave very similar results. The meta-analysis methods described in this chapter allowed us to reach the goal of this dissertation which was to be able to model non-linear exposure–disease relationships in a large individual participant meta-analysis allowing for the effects of exposure measurement error. We applied structural P-splines, using the overall model selection rule and pointwise derivative pooling to obtain our best estimate of the FBG–CHD relationship in the ERFC data.

In the next chapter we apply the methods we have developed throughout this thesis to the analysis of another risk factor for CHD in the ERFC, lipoprotein(a).

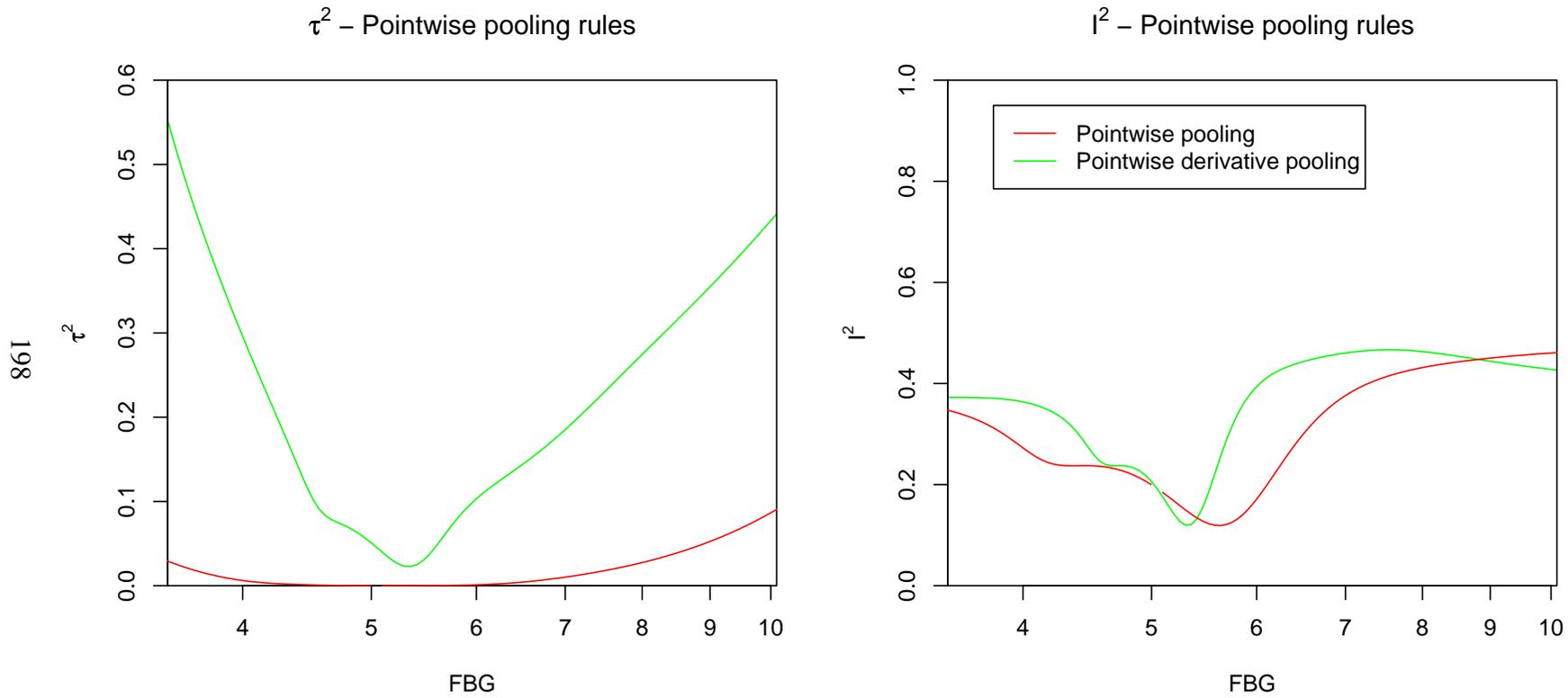


Figure 8.11: Exploration of how heterogeneity in the FBG–CHD relationship between studies varies with FBG. Left panel:  $\tau^2$  varies across the range of FBG values, Right panel: Higgin’s  $I^2$  varies across the range of FBG.

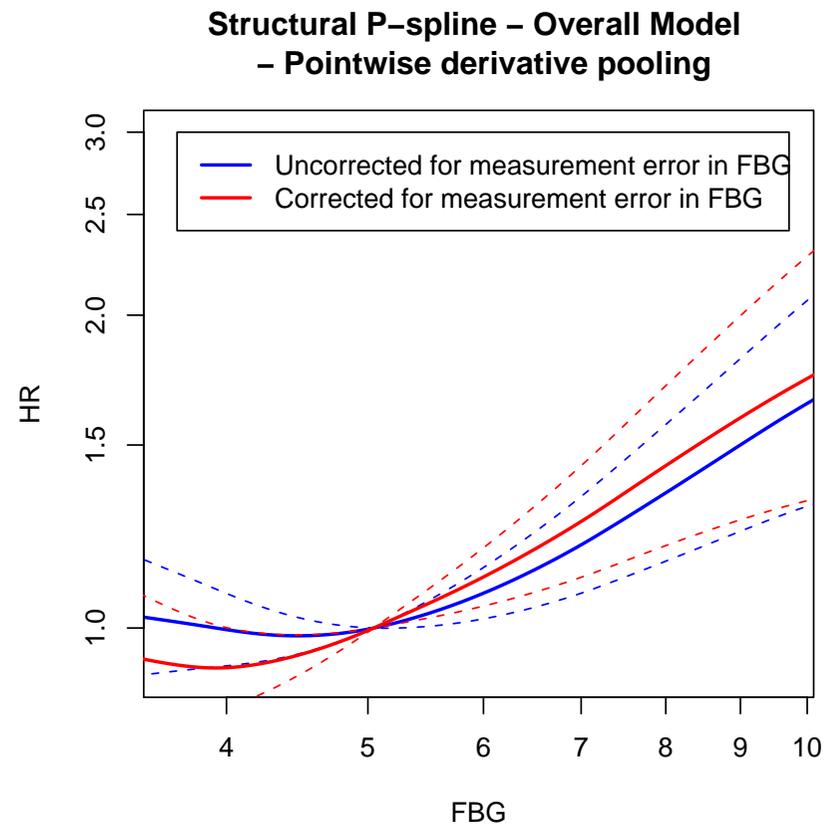
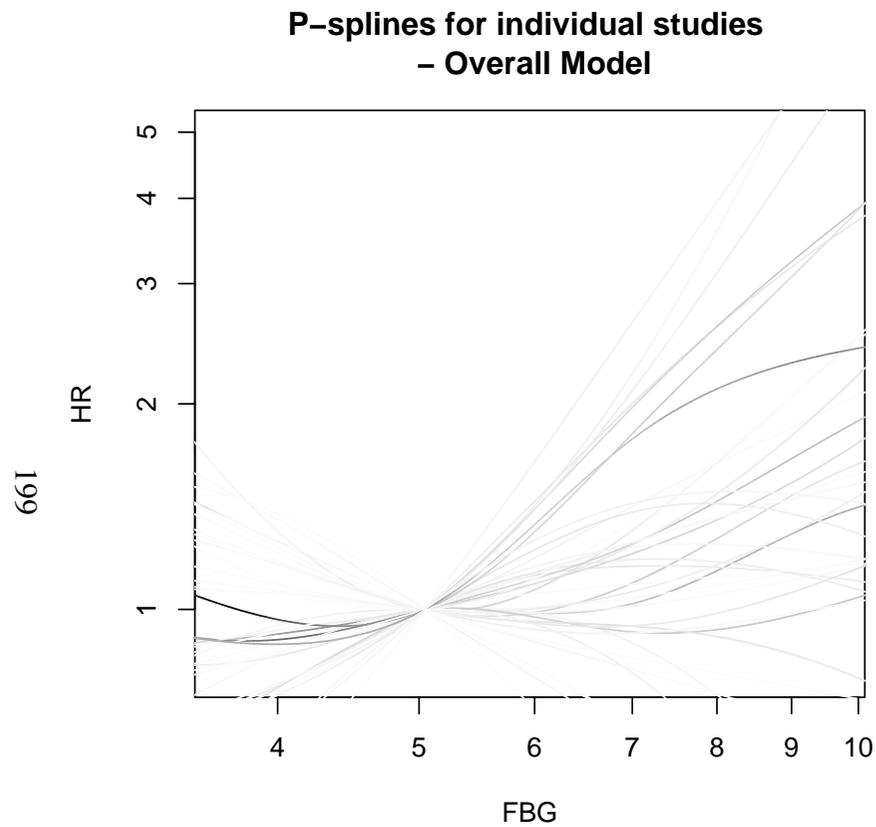


Figure 8.12: Meta-analysis of the relationship between FBG and CHD, uncorrected and corrected, for the effects of measurement error. Models for individual studies selected using the overall model selection rule (left hand panel), and pooled using the pointwise derivative pooling rule (right hand panel).



# Chapter 9

## Analysis of ERFC Lipoprotein(a) dataset

In this chapter we analyse the ERFC lipoprotein(a) (Lp(a)) dataset using methods that we have introduced and developed throughout this dissertation. We aim to illustrate that the methods provided in this dissertation are generally applicable to the analysis of large IPD meta-analyses. We consider the analysis of both frequency and individually matched nested case control studies, and discuss some of the differences in how these should be analysed. Firstly, we provide some background on what is known about the Lp(a)–CHD relationship. We then provide a summary of the ERFC Lp(a) data including an analysis of the measurement error structure. We then provide an analysis of the Lp(a)–CHD relationship that is both corrected for the effects of measurement error, and that takes account of heterogeneity in the shape of the Lp(a)–CHD relationship between studies using meta-analysis. We conclude this chapter with a discussion of the results.

### 9.1 Background — Lp(a) and CHD

Lp(a) is a plasma lipoprotein consisting of a low density lipoprotein-like particle and molecules of apolipoprotein B100 and apolipoprotein(a), and is mainly produced in the liver [268]. Lp(a) concentrations vary enormously within the population, with over a thousandfold difference between individuals often observed, from  $<0.2$  to  $>200$  mg/dl. The concentration of an individual's level of plasma Lp(a) can be measured in a blood sample using one of several different assay methods. Several isoforms of the apolipoprotein(a) component of Lp(a) are typically observed in an individual's blood. Isoforms vary in size and some older assay techniques are sensitive to between-person variation in the average size of these isoforms [269]. The role of Lp(a) within the body is not well understood, however it is believed to be an acute phase reactant [270].

High levels of serum Lp(a) have been observed as a risk factor for CHD, cerebrovascular disease, atherosclerosis, thrombosis, and stroke. Nordestgaard *et al.* [271] provide a comprehensive review of the Lp(a)–CHD relationship. They describe the relationship as being continuous

## 9. Analysis of ERFC Lipoprotein(a) dataset

---

without a threshold, and recommend that a desirable level for Lp(a) is less than about 50 mg/dl. It is thought that the link between Lp(a) and CHD is causal.

If Lp(a) is an independent cause of CHD then it may be beneficial to reduce Lp(a), for those with high levels. Lp(a) concentrations within individuals appear to be affected by kidney disease, but only slightly affected by diet, exercise, and other environmental factors. Moderate alcohol consumption is known to reduce the risk of CHD, and it has been suggested that the mechanism is through reducing levels of Lp(a) [272]. Niacin has been effectively used to reduce plasma Lp(a) [273], as well as reduce the risk of CHD. However, no randomised controlled trials has so far focused on reducing the risk of developing CHD through reducing plasma Lp(a) in those with high levels [271].

### 9.2 The ERFC Lp(a) dataset

The ERFC Lp(a) dataset contains IPD on 120,654 individuals from 35 prospective cohort studies who suffered nearly 7,500 CHD events during over 1.1 million years of follow up. Of these studies, 9 are nested case–control studies; 6 individually matched and 3 frequency matched, mainly due to costs of measuring Lp(a) in the full cohort. The event of interest was first CHD event. A list of the study abbreviations used in this chapter are given in appendix E.

There appears to be no discernible lower level of quantification of Lp(a) in the studies except for the in the GRIPS study where there was a lower limit of 4mg/dl, analyses were conducted that excluded this study were conducted to check that this lower level did not significantly impact on the results. Three studies (ATTICA, GOH, TARFS) were excluded from our analyses because they observed fewer than 10 CHD events each. 6,178 individuals had at least one repeat measurement of Lp(a). 657 unmatched individuals, mainly controls, from nested case–control studies were removed from the analysis. In each analysis we control for the confounding effects of conventional cardiovascular risk factors: age, sex, BMI, SBP, smoking status (current smoker vs. non-current smoker), history of diabetes and total cholesterol. These confounding variables were available on 101,530 individuals from 29 studies who suffered 6,526 CHD events during nearly 1 million years of follow up, and we use this set of individuals in all analyses that follow. A summary of these data is given in table 9.1.

Although we excluded TARFS from our main analyses due to lack of events, TARFS does have repeat measurements of Lp(a) along with information about confounding variables available for a subset of participants; we therefore use the data from this study in our measurement error modelling. The dataset used in this chapter does not include the Reykjavik study which was additionally included by the ERFC in their analyses. Further details about the ERFC Lp(a) dataset, and the analyses that they performed can be found in the ERFC's papers [7, 11].

We model the Lp(a)–CHD relationship in each study using a Cox proportional hazards model stratified by sex and trial arm (where appropriate), except for the individually nested case–

control studies where conditional logistic regression is used, and frequency matched case–control studies where logistic regression is used. These models give estimates of hazard ratios in the case of standard cohort studies and frequency matched case–control studies, and odds ratios in the case of individually matched case–control studies. If the disease is rare, odds ratios approximate hazard ratios and therefore it is reasonable to combine them across studies using meta-analysis [15].

The left hand panel of figure 9.1 shows the distribution of observed Lp(a), which is highly positively skewed (2.40). We therefore applied a log transform (right hand panel) to Lp(a) to achieve greater normality (skew -0.30, excess kurtosis -0.19). We shall use log-Lp(a) in all analyses that follow.

Study	No. of individuals	No. of ind. with repeats	Events	Male	Current smokers	Age mean (sd)	Lp(a) median (IQR)	Repeat Lp(a) median (IQR)
Cohort studies								
AFTCAPS	902	874	21	745	115	58.8 (7.1)	7.6 (3.3 - 17.9)	7.7 (3.3 - 20.0)
ARIC	13989		843	6065	3572	54.4 (5.7)	18.3 (6.9 - 43.8)	
BRUN	798		53	385	193	57.9 (11.4)	8.8 (4.4 - 21.6)	
CHS1	3837		588	1472	452	72.3 (5.2)	12.6 (4.8 - 22.2)	
COPEN	7484	3787	269	3168	3623	59.3 (13.4)	18.6 (6.8 - 41.9)	20.7 (4.8 - 64.3)
DUBBO	1995		272	840	312	68.4 (6.7)	11.0 (5.0 - 27.8)	
EAS	622		53	316	144	64.3 (5.6)	9.2 (3.8 - 25.7)	
FINRISK92	2190		92	1015	550	53.6 (6.2)	12.2 (4.5 - 31.6)	
GRIPS	5783		299	5783	2178	47.7 (5.1)	9.0 (4.0 - 25.0)	
KIHD	1981		383	1981	602	52.5 (5.3)	9.6 (3.9 - 22.1)	
NHANES3	2457		60	1033	591	54.6 (15.6)	22.0 (9.0 - 45.0)	
NPHSII	2367		157	2367	875	56.5 (3.4)	10.9 (4.3 - 29.2)	
PRIME	7431		114	7431	1972	54.8 (2.9)	10.0 (5.0 - 30.0)	
PROCAM	3185	453	94	2244	1140	43.0 (10.4)	4.0 (2.0 - 13.0)	7.0 (4.8 - 64.3)
QUEBEC	617		20	617	280	57.2 (7.0)	21.6 (8.9 - 49.0)	
SHS	3728		407	1477	1239	56.1 (8.0)	3.0 (1.2 - 6.7)	
TARFS <sup>1</sup>	1235	359	2	580	317	53.7 (10.5)	10.1 (4.1 - 21.8)	12.1 (4.3 - 25.9)
ULSAM	1420	320	386	1420	925	50.9 (5.1)	8.3 (3.5 - 22.7)	11.1 (5.3 - 32.8)
WHITE2	7720		168	5343	1484	49.5 (6.0)	20.0 (12.0 - 46.0)	
WHS	22667		206	0	2585	55.1 (7.2)	10.8 (4.4 - 32.9)	
WOSCOPS	4617		299	4617	2080	55.2 (5.6)	17.0 (7.0 - 50.0)	
ZUTE	304		42	304	103	75.5 (4.4)	12.2 (5.8 - 28.7)	
Nested case–control studies (Individually matched)								
BUPA	111		19	111	38	51.7 (7.3)	16.2 (7.2 - 45.5)	
FIA	1155		401	952	291	53.8 (7.4)	26.5 (11.9 - 44.9)	
FLETCHER	535	71	139	432	108	59.7 (14.0)	20.6 (7.3 - 57.6)	20.2 (5.0 - 62.6)
MRFIT	727		243	727	485	46.5 (5.6)	3.4 (1.2 - 9.3)	
NHS	627		214	0	189	60.2 (6.5)	9.5 (4.8 - 28.6)	
Nested case–control studies (Frequency matched)								
BRHS	1554		459	1554	706	52.2 (5.3)	6.5 (3.4 - 16.6)	
GOTO33	126		16	126	47	50.6 (0.2)	9.8 (4.2 - 31.8)	
USPHS	601		209	601	104	59.4 (9.0)	9.3 (3.7 - 25.3)	
Total <sup>2</sup>	101530	5505	6526	53126	26983	55.0 (9.3)	12.5 (4.9 - 32.1)	16.5 (8.2 - 38.6)

Table 9.1: Summary of the ERFC Lp(a) data by study.

<sup>1</sup> TARFS included as it provides information on repeats and contributes only to the regression calibration modelling.

<sup>2</sup> Only includes TARFS in columns pertaining to repeat measurements.

## 9. Analysis of ERFC Lipoprotein(a) dataset

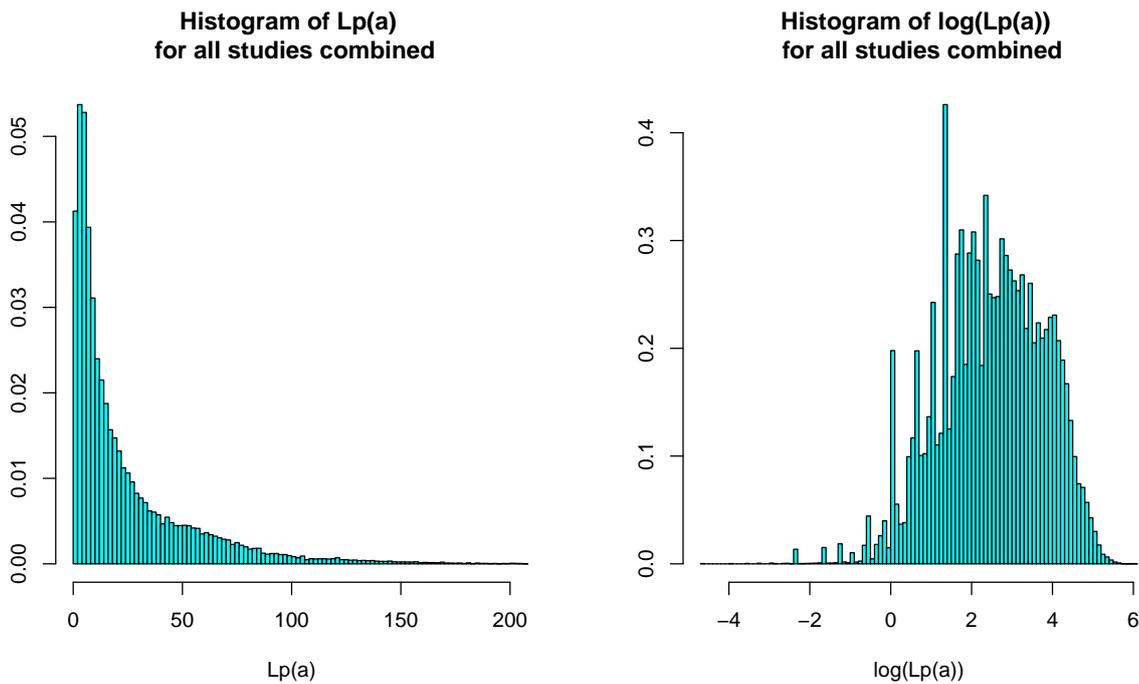


Figure 9.1: Histograms of Lp(a) and log-Lp(a) for all studies combined.

### 9.2.1 Measurement error

Unlike most of the risk factors considered by the ERFC, measurements of Lp(a) concentration are subject to relatively modest measurement error. For example, Bennet *et al.* [274] found that the RDR for log-Lp(a) in the Reykjavik study [275] was 0.92 (95% CI, 0.85-0.99). Much of the variation in observed Lp(a) levels between individuals can be explained by genetic variation [276], therefore we might expect levels to stay relatively constant over time. Although the amount of measurement error in observed Lp(a) may be small it should still be accounted for, firstly because variation in the degree of measurement error is likely to vary between studies (e.g. because of (non)sensitivity to isoforms of apolipoprotein(a)) making the observed levels of Lp(a) incomparable, and secondly so that we may obtain our best estimate of the true Lp(a)–CHD relationship. We can, however, only take account of differences in measurement error between studies where we have repeat Lp(a) measurements available.

Plots of within-person mean against within-person standard deviation for each of the six studies with repeats are given in figure 9.2. The COPEN study, which has the largest number of repeat measurements, suggests that the measurement error variance decreases with increasing levels of Lp(a). This trend is also seen in FLETCHER and PROCAM, and to a lesser extent in ULSAM, although this may be partially due to low numbers of individuals with repeats, with low levels of Lp(a). TARFS suggests a constant error variance. The banding that can be seen in the PROCAM study is because the Lp(a) measurements in this study only take integer values (except for the lowest level of Lp(a) which is 0.2 mg/dl). The most interesting result is in AFTCAPS, where the points seem to be clustered into three distinct groups. We do not know of any reason for this clustering.

---

Figure 9.4 shows the normal Q-Q plot of within-person differences for the ULSAM study, which is representative of what we observed for the other studies. It appears to show that the measurement error distribution has heavy tails. In chapter 6, we saw that structural fractional polynomials appear to be relatively robust to heteroscedastic/non-normal measurement error. Figure 9.3 shows the normal Q-Q plot of within-person means for each study. Most of the studies suggest that within-person mean  $Lp(a)$  shows some non-normality in the tails, but as we noted in chapter 6 this can be caused by non-normality of the measurement error. COPEN again is outlying, exhibiting a different trend to the other four studies with large numbers of individuals with a repeat  $Lp(a)$  measurement. The measurement error variance did not appear vary with the level of any of the confounders (plots not shown).

Table 9.2 gives the parameter estimates and their standard errors for each of the regression calibration models fitted to the studies where repeat observations were available in a subset of participants, along with an ‘overall model’ obtained via random effects meta-analysis. ULSAM does not contribute to the coefficient for sex since all participants in this study were male. There appears to be a difference between COPEN and the other studies with repeat measurements; the RDR (on the log-scale) for  $Lp(a)$  is significantly lower than in the other studies. We speculate that COPEN may differ from the other studies with repeats because it is the only study known to have used an isoform sensitive assay (the assay method for AFTCAPS is unknown). Ideally, either meta-regression on the RDR, or subgroup analyses, would be conducted to try and ascertain the cause of the substantial differences in the RDR between studies, however this is not possible when repeat measurements are only available for six studies.

## **9.3 Analysis of the $Lp(a)$ –CHD relationship**

### **9.3.1 Grouped exposure analysis**

As an initial exploration of the  $Lp(a)$ –CHD relationship we perform a grouped exposure analysis on deciles of the observed data; the results of which are shown in figure 9.5. The grouped exposure analysis suggests a threshold shaped relationship between  $Lp(a)$  and CHD, with the threshold occurring between the 7th and 8th groups i.e. at an  $Lp(a)$  of between about 22–32 mg/dl. This relationship does not appear to be greatly affected by adjustment for conventional cardiovascular risk factors (right hand panel), which suggests that  $Lp(a)$  is an independent risk factor.

## 9. Analysis of ERFC Lipoprotein(a) dataset

### Mean-variance association for log-Lp(a)

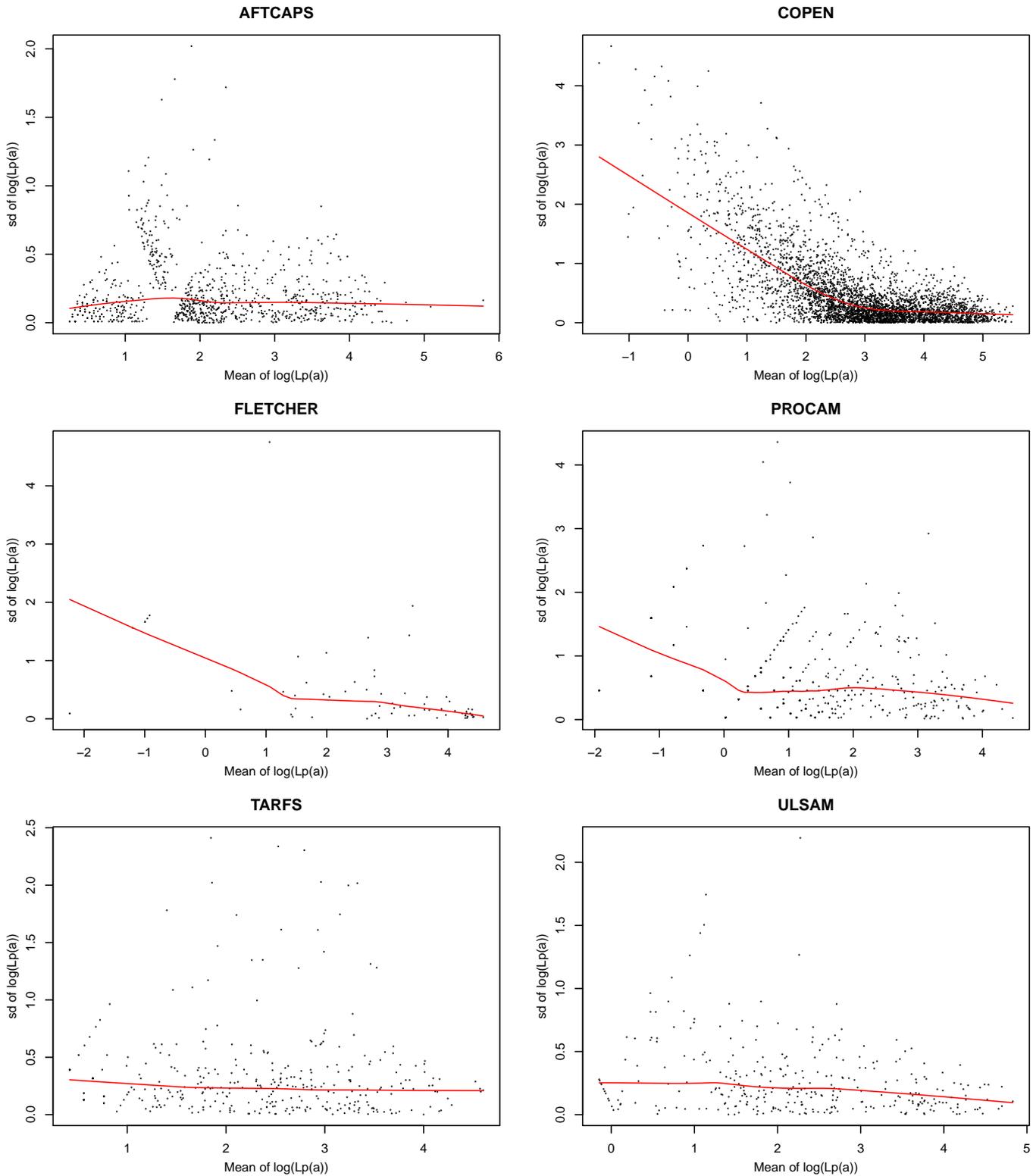


Figure 9.2: Mean-variance association of log-Lp(a) for each of the six studies with repeat Lp(a) measurements. Red lines give LOESS smooth of points to aid trend identification.

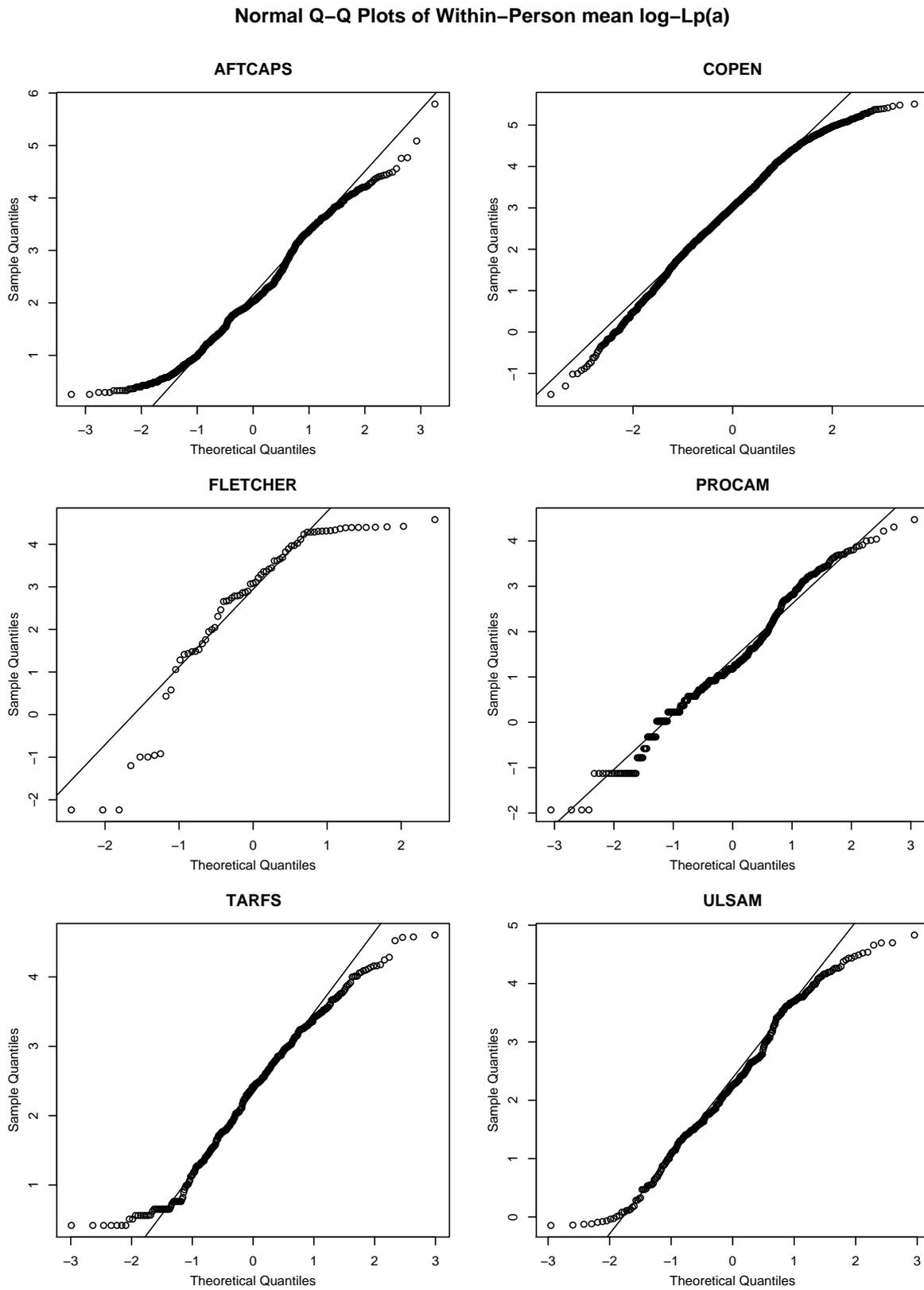


Figure 9.3: Normal Q-Q plots of within-person means of log-Lp(a) for each of the six studies with repeat Lp(a) measurements.

	(Intercept)	log-Lp(a)	Age	Sex: Female	SBP	BMI	Total cholesterol	Smoking status: Current	Diabetic status: Definite diabetic
AFTCAPS	0.178 (0.260)	0.947 (0.015)	-0.006 (0.002)	0.133 (0.043)	0.001 (0.001)	-0.001 (0.005)	0.019 (0.029)	-0.040 (0.047)	-0.105 (0.126)
COPEN	1.475 (0.088)	0.528 (0.007)	-0.004 (0.001)	0.078 (0.021)	0.002 (0.001)	0.000 (0.003)	0.035 (0.009)	0.015 (0.020)	0.039 (0.082)
FLETCHER	-0.567 (1.708)	0.852 (0.080)	-0.003 (0.014)	0.263 (0.420)	-0.003 (0.008)	0.037 (0.040)	0.095 (0.131)	-0.081 (0.387)	-0.559 (0.597)
PROCAM	-0.453 (0.632)	0.663 (0.042)	0.009 (0.007)	-0.206 (0.135)	0.003 (0.004)	-0.013 (0.020)	0.071 (0.061)	0.063 (0.118)	0.105 (0.494)
TARFS	0.323 (0.313)	0.833 (0.037)	-0.015 (0.004)	0.082 (0.084)	0.005 (0.002)	-0.002 (0.003)	0.045 (0.033)	0.037 (0.103)	0.073 (0.139)
ULSAM	-2.274 (4.251)	0.879 (0.023)	0.051 (0.085)	-	-0.001 (0.002)	0.007 (0.010)	0.002 (0.025)	0.017 (0.060)	-0.032 (0.149)
Overall	0.382 (0.450)	0.783 (0.105)	-0.005 (0.002)	0.081 (0.034)	0.002 (<0.001)	-0.001 (0.002)	0.032 (0.008)	0.009 (0.017)	0.001 (0.056)

Table 9.2: Regression calibration models for log-Lp(a) for each study with repeats, and overall model obtained via random effects meta-analysis.

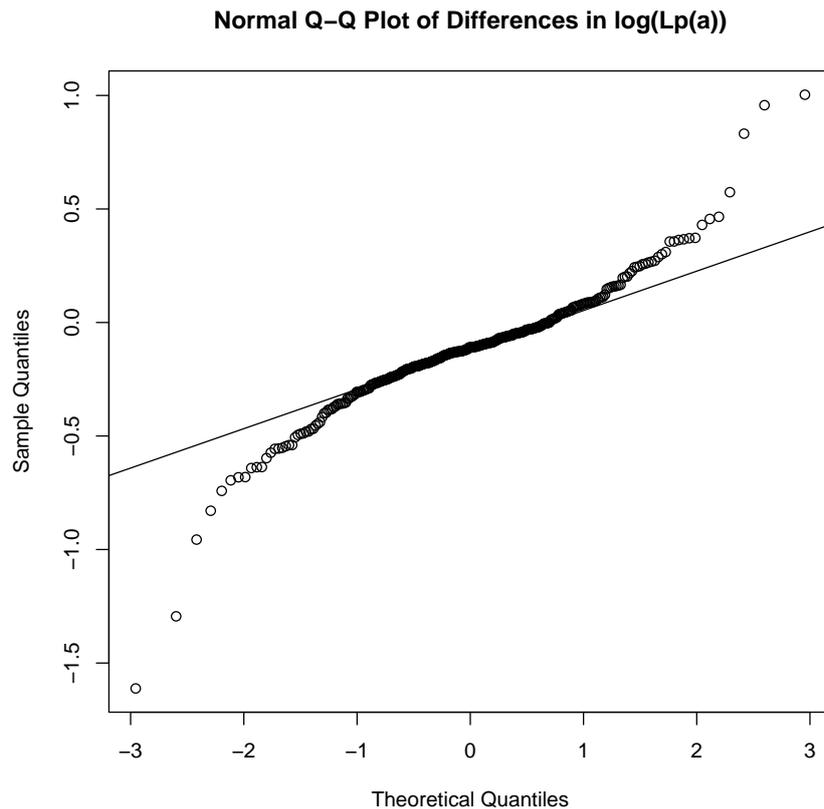


Figure 9.4: Normal Q-Q plot of differences in log-Lp(a) between baseline and repeat measurements for the ULSAM study.

### 9.3.2 Continuous exposure analysis

#### Uncorrected relationship

Figure 9.6 shows the estimated shape of the relationship from a random effects meta-analysis of fractional polynomial and P-spline models for the Lp(a)–CHD relationship, using the studywise model (fixed  $4df$ ) selection rule and pointwise pooling rule. We could not use the overall model selection rule for these data because we are using a combination of Cox and logistic regression models for the individual studies.

Under the fractional polynomial model the relationship appears J-shaped, with the nadir occurring at about 4 mg/dl. Under the P-spline model the relationship is flat below the geometric mean (11.89 mg/dl), and then increases steadily up to about 70 mg/dl where it appears to begin flattening off. If we compare the fractional polynomial model with the top right hand panel of figure 3.2, and the P-spline model with figure 3.3, then we observe that these shapes for the Lp(a)–CHD relationship are what we would expect to obtain under P-spline and fractional polynomial models if the true underlying relationship exhibited a threshold. In chapter 3 we saw that the nadir of the fractional polynomial model occurs significantly before the threshold, and that with P-spline models the risk begins to increase closer, although still before, the true threshold. These models therefore suggest that a threshold may occur somewhere in the range

## 9. Analysis of ERFC Lipoprotein(a) dataset

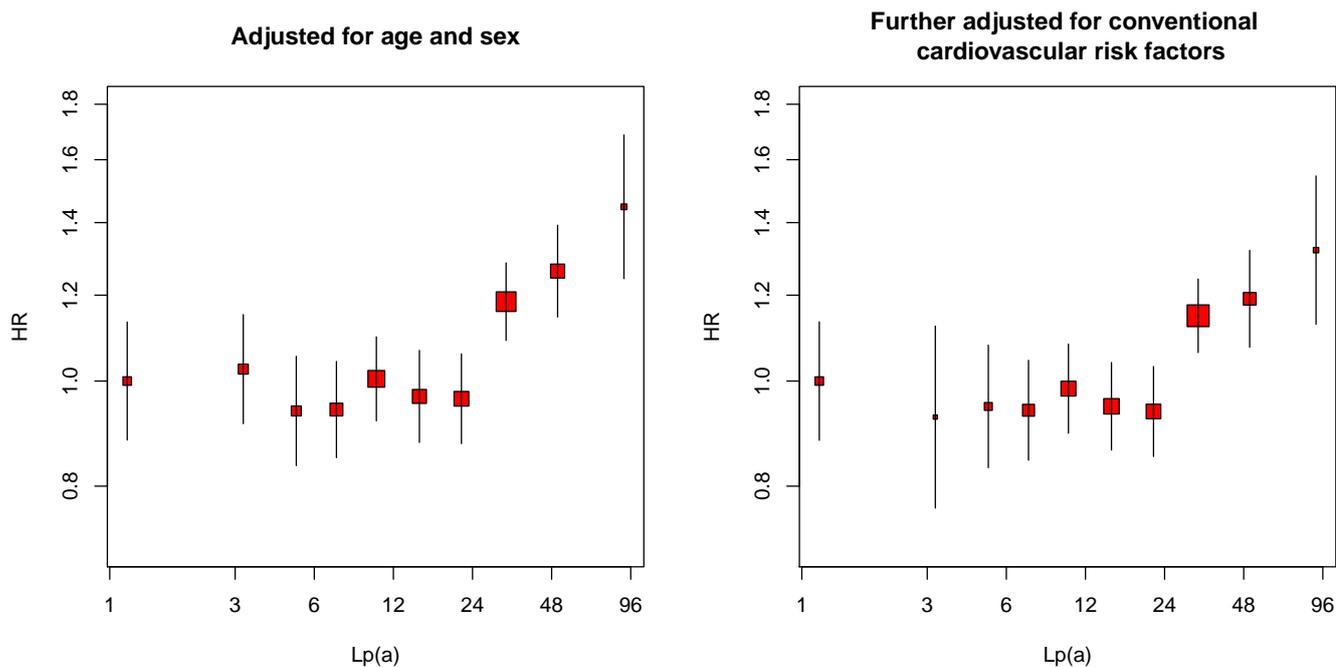


Figure 9.5: Group based analysis of observed Lp(a) adjusted for age and sex (and other conventional cardiovascular risk factors). The group containing those with the lowest Lp(a) measurements is taken as the reference..

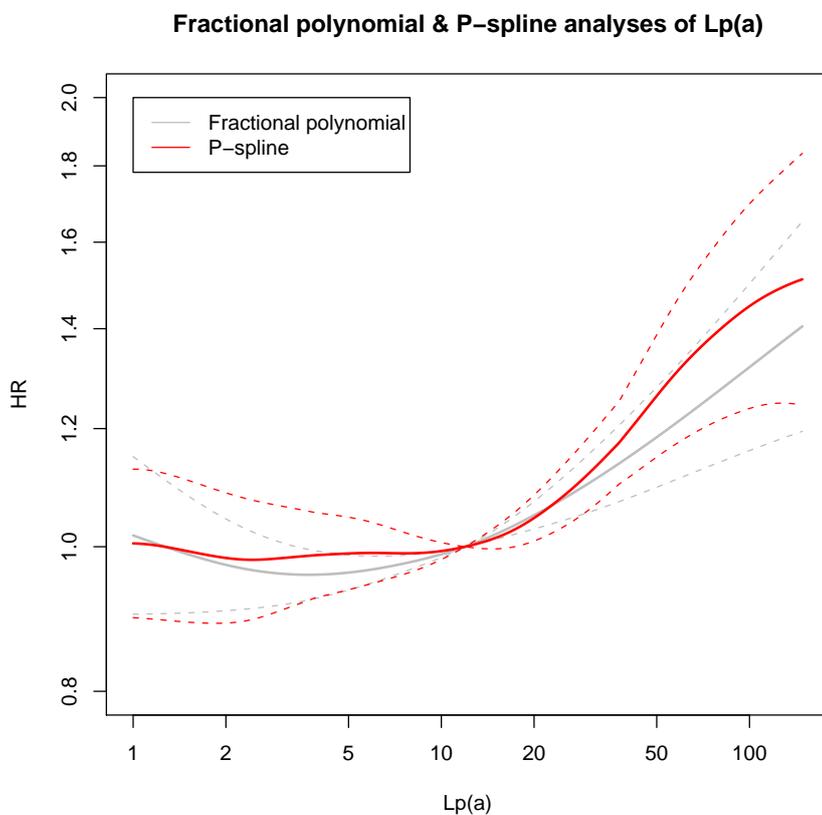


Figure 9.6: Fractional polynomial and P-spline analyses of observed Lp(a) adjusted for age, sex, and other conventional cardiovascular risk factors.

---

of 16–20 mg/dl.

In figure 9.7 we can see the Lp(a)–CHD relationship for each of the individual studies using fractional polynomial and P-spline analyses where, as in chapter 8, the saturation of the lines represents their weight in the pointwise meta-analyses resulting in figure 9.6. Under both plots we can see a definite increasing trend in the hazard ratio with increasing Lp(a). In the fractional polynomial models we can see that for a number of the studies the hazard ratio increases steeply beyond the geometric mean, whilst in others, it increases linearly. Under the P-spline models we see a far greater number of studies with large increases in the hazard ratio at higher levels which suggests that the fractional polynomials are failing to pick up this increase where a comparatively small amount of individuals will be located in most of the individual studies. We have observed previously that P-splines are better than fractional polynomials at estimating the shape of true threshold relationships; we shall use P-spline models in the rest of this chapter.

### **Measurement error corrected Lp(a)–CHD relationship**

We have used the same approach as in chapter 6 to obtain values for the expectation and variance of true Lp(a) given observed Lp(a), for each study. Namely, for those studies where we had repeat measurements available, we used the regression calibration model corresponding to that study from table 9.2; and for those studies for which there were no repeat measurements available, we used the ‘overall model’ in the final row of table 9.2. As we saw in section 9.2.1 there was significant heterogeneity in the RDR between studies ( $I^2=99.4\%$ , 95% CI [99.2%, 99.5%]), so there is some doubt about the transportability of our model. We note in that this is a significant uncertainty in our results.

In previous chapters we have only used structural P-splines with Cox models, although conditional logistic regression can be formulated as a special case of the Cox proportional hazards model. In order to fit structural P-spline models using logistic regression we created a custom smooth class for the `mgcv` package [277].

Figure 9.8 shows the results from a meta-analysis of structural P-spline models using the pointwise, and pointwise derivative pooling rules. Again, we see that the relationship appears to be threshold in shape, with the relationship beyond the mean level of Lp(a) stronger than in the uncorrected relationship. Below the mean level of Lp(a), measurement error correction makes little difference to the relationship, since essentially it is null under pointwise pooling; although there is a small increase in the hazard ratio for very low levels of Lp(a) under the pointwise derivative pooling rule. Under the pointwise pooling rule the relationship starts to flatten off at around 70 mg/dl, whereas under the pointwise derivative pooling rule the gradient is shallower

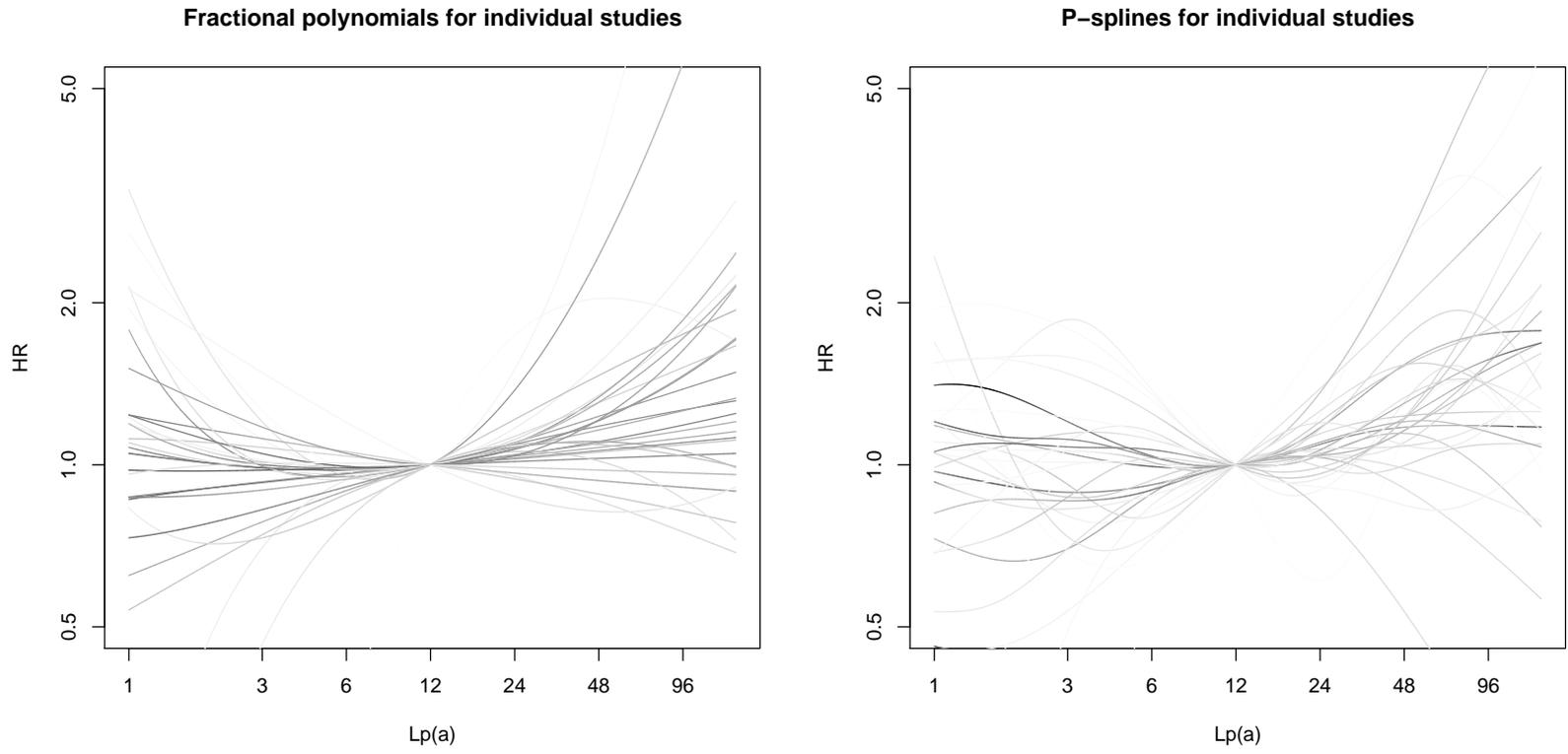


Figure 9.7: Fractional polynomial (left) and P-spline models (right) of the Lp(a)–CHD relationship for each study.

after about 30 mg/dl, but shows no sign of flattening off at higher levels. The corrected relationship still appears to be consistent with the relationship having a threshold between 16–20mg/dl.

The pointwise standard errors for the pointwise derivative pooling rule in figure 9.8 are obtained via the nonparametric bootstrap as in chapter 8. In this analysis, however, the resampling is complicated by the case–control studies. We used the following resampling scheme for each kind of study:

- Standard cohort study — sample with replacement within each study.
- Individually matched case–control studies — sample with replacement from the cases, and then sample with replacement from the matched controls for each sampled case.
- Frequency matched case–control studies — sample with replacement from the cases and the controls separately.

We found that we had model convergence problems for a large proportion of our resampled datasets for the two smallest studies, GOTO33 and BUPA. We therefore decided to include the original data without resampling in each bootstrap dataset for these two studies. Since these studies are small and therefore the parameter estimates from the model fit are poorly estimated, these studies have small weights in the meta-analysis and hence this approach should not unduly influence the results obtained.

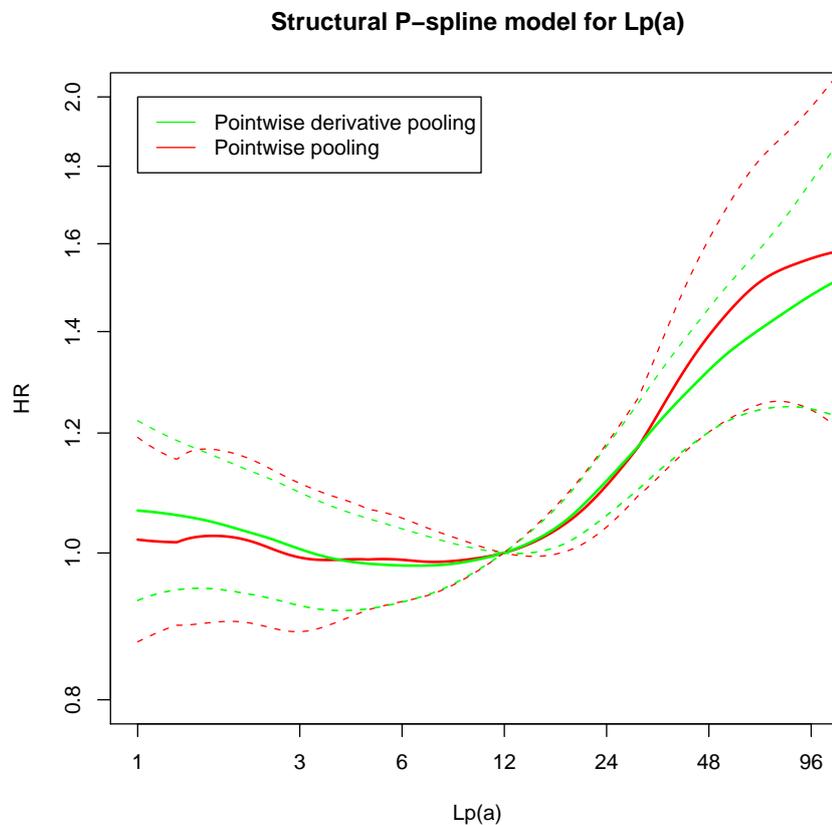


Figure 9.8: Structural P-spline analysis of the Lp(a)–CHD relationship adjusted for age, sex, and other conventional cardiovascular risk factors.

## 9. Analysis of ERFC Lipoprotein(a) dataset

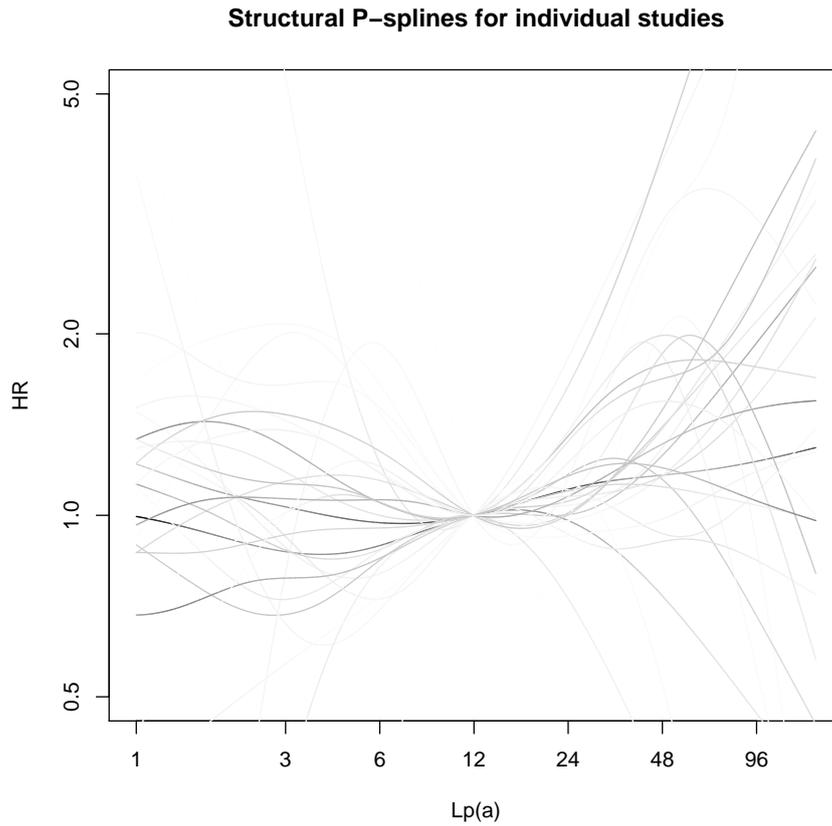


Figure 9.9: Structural P-spline analysis of the Lp(a)–CHD relationship for each study. Dark-ness of lines related to weights in a pointwise meta-analysis of the individual studies.

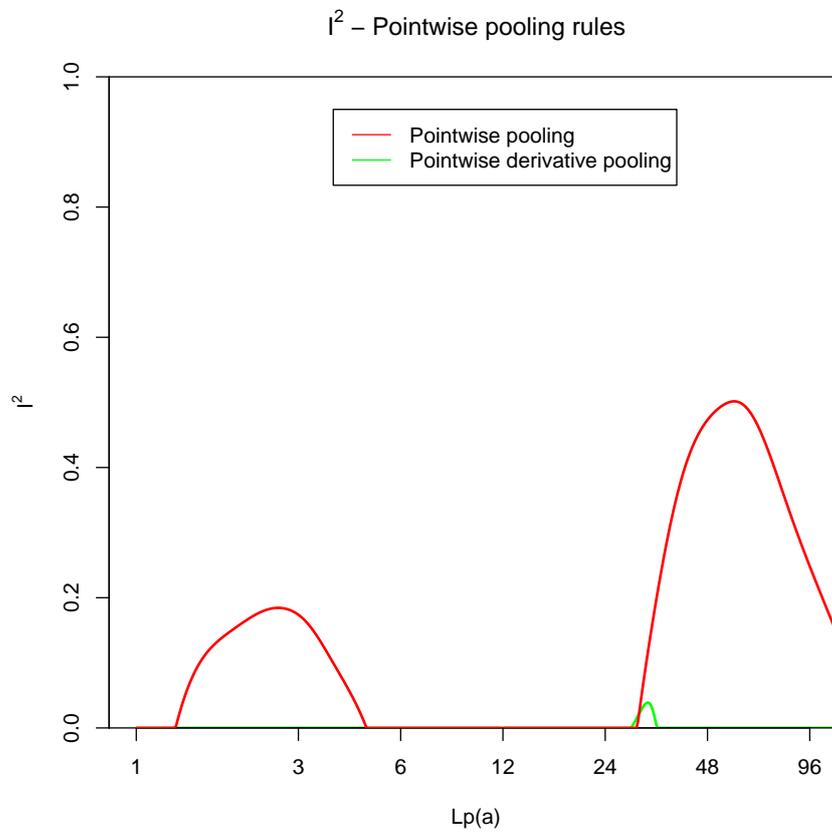


Figure 9.10: Higgin's  $I^2$  across the range of Lp(a) for pointwise and derivative pointwise pooled structural P-spline models of the Lp(a)–CHD relationship.

---

Figure 9.9 shows the measurement error corrected LP(a)–CHD relationship for each of the individual studies. Many of the studies do seem to suggest that there is a sharp increase in risk somewhere between 16–20 mg/dl, but there appears to be a lot of heterogeneity in the shape of the Lp(a)–CHD relationship beyond an Lp(a) of 24 mg/dl. At low levels of Lp(a) there appears to be considerable differences between studies, with no clear consensus on whether there is an increased or decreased risk of CHD associated with low levels of Lp(a).

Figure 9.10 investigates how heterogeneity varies across the range of Lp(a) using Higgin’s  $I^2$ . Under the pointwise derivative pooling rule there is no heterogeneity across the entire range, except for a blip between about 30–35 mg/dl. Under the pointwise pooling rule there is no heterogeneity between 5 and 30 mg/dl, and two peaks of heterogeneity, one at low levels of Lp(a) and the other after 30 mg/dl. In figure 9.8, the heterogeneity under the pointwise pooling rule can be seen to impact the confidence intervals in each of the two areas where the heterogeneity is non-zero. It may seem surprising that so little heterogeneity is observed in figure 9.10 given that there is such variation in the shape of the relationship in the individual studies (figure 9.9), this is because the within-study variances are generally large meaning that the relationships are not incompatible with each other given zero between-studies variance.

## 9.4 Discussion

In this analysis of the ERFC Lp(a) dataset we have seen a modest relationship between levels of Lp(a) and CHD. Lp(a) is subject to measurement error, but considerably less so than many other risk factors for CHD. Lp(a) appears to have little effect on the risk of CHD below 16mg/dl, it appears that there is a threshold somewhere between 16–20 mg/dl, and that the hazard increases considerably beyond this point. The hazard ratio at Lp(a) level of 96 mg/dl, relative to the geometric mean, is in excess of 1.5. In the ERFC’s analysis [11] they described the relationship between Lp(a) and CHD as being curvilinear. If the true relationship were to exhibit a threshold somewhere in the region of 16–20 mg/dl then, as we saw in section 5.1, this is exactly the shape we would expect to see if we had applied MacMahon’s method. As we have seen previously, sharp features in the exposure–disease relationship can be lost, even if the measurement error is relatively small.

In our analysis we have used P-splines to model the shape of the Lp(a)–CHD relationship. We have previously noted the difficulty P-splines, and to an even greater degree, fractional polynomials, have in picking up thresholds in exposure–disease relationships. A change point model may be better for modelling these data, especially if the threshold value were of particular interest. Even with this large dataset consisting of over a hundred thousand individuals, we still do not have the ability to obtain a precise estimate of the shape of the relationship beyond 85 mg/dl.

There was a lot of variation in the median level of Lp(a) between studies; this could simply

## 9. Analysis of ERFC Lipoprotein(a) dataset

---

be due to heterogeneity, or differing levels of systematic bias between studies. When we have data from an individual study then systematic bias that affects all individuals is not necessarily a problem as we are still able to compare individuals relative to each other. However, when we perform a meta-analysis of multiple studies then the location of the data from each study, relative to each of the other studies is important, and we could potentially get distortion of the exposure–disease relationship, due to systematic bias differing between studies.

COPEN appeared to differ from the other studies with repeats, however, it is hard to evaluate this when there are only a small number of studies with repeats. There is clearly much uncertainty about the transportability of the measurement error models from those studies with repeats to those without given the significant heterogeneity in the RDR; however, when further information is unavailable these type of assumptions are unavoidable and are a limitation to our analysis. A sensitivity analysis was carried out where we excluded the COPEN study from our meta-analysis in obtaining our regression calibration model and it was found that it did not significantly affect the results. The Lp(a) dataset has highlighted again the need for studies to be designed such that replicate or gold standard measurements are obtained for a subset of individuals. As in chapter 6 the analyses in this chapter have the limitations that they do not take account of the uncertainty from the estimation of the measurement error model, and do not take account of measurement error in confounders

In section 9.2.1 we saw that the measurement error appeared to be heteroscedastic. However, as we saw in chapter 7 this appears to have little effect on the shape of the exposure–disease relationship, or our correction methods. The logarithm of log-Lp(a) was relatively normal, and the plot of within-person means suggested that the distribution of log-Lp(a) may be relatively normal, although COPEN again showed a different trend to the other studies with repeats. Overall, we feel that the non-normality present is probably not large enough to have a significant effect on the Lp(a)–CHD relationship we have obtained.

## 9.5 Conclusion

In this chapter we have analysed the Lp(a)–CHD relationship using IPD from the ERFC, utilising the methods we have discussed and developed throughout this dissertation. We corrected our models for the modest amount of measurement error in the baseline Lp(a) measurements using structural P-splines. In previous chapters we had only used Cox models, but in this chapter we used structural P-splines to correct for measurement error in logistic, and conditional logistic, regression models. The Lp(a)–CHD relationship appears to exhibit a threshold in the region of 16–20 mg/dl, with constant hazard below the threshold, and then increasing to a hazard ratio of in excess of 1.5 at an Lp(a) level of 96 mg/dl.

In chapter 10 we conclude this dissertation with a discussion of the topics covered in this thesis, areas for future work, and recommendations for practice.

# Chapter 10

## Discussion

We start this final chapter by providing a summary of the key points from each of the preceding chapters. We then discuss the key themes considered in this dissertation, placing our contributions in context. We round off the chapter by describing some of the limitations of this dissertation, and possible areas for future research. Ultimately, we give a final conclusion.

### 10.1 Dissertation summary

In chapter 1 we gave the motivation for this dissertation: the characterisation of exposure–disease relationships within the ERFC; a collaboration of studies providing IPD on risk factors for CHD. The two main challenges in characterising these relationships were highlighted as the effects of exposure measurement error, and the combination of non-linear exposure–disease relationships across studies.

In chapter 2 we discussed survival analysis and introduced the Cox model. We looked at forms for the linear predictor including: grouped exposure analyses which are sensitive to the number and choice of cutpoints and are biased for non-linear exposure–disease relationships; fractional polynomials which are relatively simple yet flexible polynomial based functions; and splines which are generally more complex but can give a good local fit to the data.

In chapter 3 we showed the effects of measurement error on a range of non-linear exposure–disease relationships that are commonly observed in epidemiological studies using simulation. We found that random measurement error severely biases the exposure–disease relationship, usually, but not always, towards the null. We also saw that measurement error severely reduces our power to detect non-linear exposure–disease relationships.

In chapter 4 we discussed various methods for correcting for exposure measurement error in non-linear exposure–disease relationships and proposed two new methods: structural fractional polynomials, which are based on applying the method of regression calibration to fractional polynomials; and group-SIMEX, which was an extension of SIMEX to where we categorise a

## 10. Discussion

---

continuous exposure.

In chapter 5 we firstly showed that MacMahon's simple method for correcting for measurement error in a grouped exposure analysis does not, in general, work for non-linear exposure–disease relationships. We then compared Natarajan's proposed method, another regression calibration based method, multiple imputation, moment reconstruction, and group-SIMEX for correcting a grouped exposure analysis for the effects of measurement error when the exposure–disease relationship is linear. Natarajan's method was shown to be flawed, the other regression calibration based method did not work because it wrongly made the non-differential error assumption, group-SIMEX performed poorly for multiple reasons, but moment reconstruction and multiple imputation both showed little bias. We finished the chapter by comparing the performance of structural P-splines and structural fractional polynomials, and found that they performed similarly with neither method performing uniformly better.

In chapter 6 we described the ERFC FBG data. We then corrected the J-shaped FBG–CHD relationship for measurement error in FBG using SIMEX, structural fractional polynomials and P-splines; however, we did not take account of heterogeneity in the shape of the FBG–CHD relationship between studies. We found that the structural fractional polynomial and P-spline methods made a much larger correction for the effects of measurement error than SIMEX.

In chapter 7 we showed the effects of a range of non-classical measurement error structures on the observed exposure–disease relationship. This was motivated by the FBG data, where in chapter 6, we had observed that the measurement error appeared to be non-normal and heteroscedastic. We then considered the performance of structural fractional polynomial and P-spline models when we erroneously assumed normality of the true exposure given the observed, and found the method was relatively robust to non-normality and heteroscedasticity of the measurement error, but not to skewness of the true exposure distribution. We also considered robustification of structural P-spline and fractional polynomial analyses of the ERFC FBG data by considering a mixture of normals.

In chapter 8 we introduced meta-analysis and described the method of Sauerbrei and Royston [67] for IPD meta-analysis of non-linear exposure–disease relationships. We proposed model selection rules similar to those of Sauerbrei and Royston for P-splines, and a pointwise derivative pooling rule. We also showed the dependence of the pointwise pooling rule on the arbitrary choice of reference value, and that the exposure–disease relationships obtained using this method can exhibit Simpson's paradox. We then applied these methods to obtain our best estimate of the FBG–CHD relationship, that was both corrected for the effects of measurement error in the FBG measurements, and allowed for heterogeneity in the shape of the FBG–CHD relationship across studies.

In chapter 9 we applied the methods developed throughout this dissertation to the threshold shaped relationship between  $L_p(a)$  and CHD.

---

## 10.2 Modelling non-linear exposure–disease relationships

Grouped exposure analyses have traditionally been used in epidemiology and are simple to perform. In chapter 2 we argued that grouped exposure analyses are sensitive to the number and location of cutpoints, do not reflect the continuous nature of the underlying relationship, and we showed that they give a biased view of the shape of the exposure–disease relationship unless the relationship is linear. In chapter 3 we also saw that grouped exposure analyses can severely reduce our power to detect non-linearity compared with using continuous predictors such as P-splines and fractional polynomials. We therefore suggest that grouped-exposure analyses should not routinely be used in final analyses of exposure–disease relationships, and if they are then sensitivity analyses should be conducted to assess dependence on the number and choice of cutpoints. Statisticians have, and we hope will continue to, highlight the downsides of grouped exposure analyses [1, 118, 278] and promote the use of alternative, continuous forms for the linear predictor, such as splines and fractional polynomials [119, 121, 279].

Despite their drawbacks we feel that there are some features of a grouped exposure analysis that are still appealing. Firstly, when using Cox models, grouped exposure analyses can provide a quick and easy way to get an initial idea about the shape of the exposure–disease relationship. Secondly, quasi-variances [83] (or floating absolute risks [84]), that allow us to view the uncertainty within each group without the need for a reference category, are appealing and are without an equivalent when considering continuous forms for the linear predictor.

Continuous forms for the linear predictor have clear benefits over grouped exposure analyses—they provide a realistic, smoother, model for the exposure–disease relationship; they are not sensitive to the choice of arbitrary cutpoints; and use the data more efficiently. Hence, we strongly believe that continuous linear predictors should be used in practice. Modellers have to make a choice between different forms for the linear predictor and it is often not obvious which method should be used: fractional polynomials give a good global fit and can be written compactly; P-splines give a good local fit but cannot be written compactly. Neither fractional polynomials nor P-splines seem to be universally superior [123], and the different choices of splines and the associated terminology can be confusing for those not well versed in the literature. In Cox models where we cannot easily visualise the model’s fit to the data, it would seem sensible to follow the proposal of Royston [124] whereby we compare the model fit against a P-spline with a large number of degrees of freedom. Results are best displayed graphically, but if they are to be tabulated, then the advice of Royston [121] can be used.

A barrier to fitting models with a continuous predictor is the availability of software. Fractional polynomials are widely available [94], however the availability of spline based models is patchy. Different spline models are available in each software package, and even between disease models within the same package. Although different spline methods can give very similar results e.g. smoothing splines and P-splines, the results will still be dependent on the

## 10. Discussion

---

particular choice of spline. P-splines, which we have used extensively in this dissertation, are not readily available in many standard software packages because they require fitting routines that can handle penalisation. Simpler spline models such as quadratic and cubic splines are much easier to code for use with standard fitting routines, and several authors have provided code [279–281].

As we have seen throughout this dissertation, threshold shaped relationships are modelled poorly by all the modelling methods we have used. Although a true threshold relationship is usually epidemiologically implausible, it may be the case, as we saw with  $L_p(a)$ , that there is a relatively sharp change in the gradient of the linear predictor beyond some value. In this case it may be more appropriate to use a change point model, especially as the threshold will often be of interest in itself.

The choice of linear predictor has a significant impact on how we approach measurement error correction. In addition to the problems associated with grouped exposure analyses, we also noted that categorising continuous exposures introduces differential measurement error when the underlying exposure is subject to non-differential error. In general, differential measurement error is more difficult to correct for because the correction model will need to include both additional exposure data and the outcome. As we have shown that MacMahon's method does not work for non-linear relationships, and there are no simple proven alternatives, we recommend that grouped exposure analyses are not used when exposures are subject to measurement error.

### 10.3 Measurement error

#### 10.3.1 Effects of measurement error

Throughout this dissertation we have seen the profound effect that the '*triple whammy*' of measurement error can have on exposure–disease relationships: bias in parameter estimation in models; loss of power in detecting associations between variables; and masking of features of the data [33]. Despite this, Jurek *et al.* [282] took a selection of 57 articles from three major epidemiological journals published in 2000–2001 and found that nearly 40% said nothing about exposure measurement error, whilst of those that discussed exposure measurement error only one attempted to quantitatively evaluate the effect of measurement error using a sensitivity analysis. Although there have been many articles in the epidemiological literature regarding the effects of exposure measurement error [1, 31, 283–285], it is clear that there is a continuing need to increase awareness of the importance of the measurement error problem, especially with regards to non-linear relationships.

In chapter 3 we saw the effect that random measurement error has on the shape of commonly observed exposure–disease relationships, and how it significantly reduces our power to detect

---

non-linear relationships. Therefore, measurement error may explain heterogeneity in the observed shape of the exposure–disease relationship between studies. A lack of power to detect non-linear relationships can cause model-misspecification which could lead to inappropriate correction for measurement error. We feel that this could be a very real problem in practice. The results show the importance of study size in ascertaining whether a non-linear relationship is present, especially when the degree of measurement error is substantial. When studies are designed they should try to take account of this lack of power when deciding upon the cohort size.

In chapter 7 we saw the effects of non-classical exposure measurement error and how it can introduce non-linearity into linear relationships and distort the shape of non-linear relationships. The introduction of non-linearity by non-classical error into truly linear relationships has the opposite effect of classical measurement error on non-linear relationships, thus highlighting the importance of trying to discern some properties of the measurement error structure through the use of diagnostic plots, something that is not usually done in practice, but that we believe should be.

### **10.3.2 Correction for measurement error**

We have discussed how measurement error correction requires repeat, or gold standard, measurements to be available for a subset of individuals. Despite this, only 17 of the 38 FBG studies, and 6 of the 30 Lp(a) studies, had repeat measurements available. When studies have not obtained additional exposure data then there is no way we can quantify the measurement error in those studies. In this situation we are forced to make assumptions about the measurement error, and in this dissertation we took a weighted average from the other studies. This is far from ideal since we have observed that there can be considerable heterogeneity in the amount of measurement error between studies, thus making it difficult to assess whether our model is transportable. Studies, therefore, must be designed so that a validation substudy is included, because there is no point in conducting a study to ascertain the shape of an exposure–disease relationship if data are not collected that allow you to estimate it accurately. We note that this is an unfair comment for some of the ERFC’s studies since FBG and Lp(a) were not always exposures of primary interest in the individual studies.

For a linear exposure–disease relationship the use of regression calibration, or equivalently correction using the RDR, is a simple and widely applicable way to correct for the effects of exposure measurement error, and has therefore been widely used in practice [43, 44]. Regression calibration is exact for linear regression, and approximate for logistic and Cox regression [28, 136].

When the exposure–disease relationship is non-linear there is no correction method that is as simple, effective, and widely applicable as regression calibration is for the linear relationship. As we have seen, regression calibration is not so simple because we have to make additional

## 10. Discussion

---

assumptions about the relationship between the observed and true exposure; although it is still much simpler and more widely applicable than most other correction methods. Structural fractional polynomials and P-splines combine regression calibration with non-linear predictors for the disease model. Despite having to make additional assumptions when using regression calibration for non-linear relationships, we saw in chapter 7 that structural fractional polynomials and P-splines are relatively robust to both heteroscedastic and non-normal measurement error when we assume normality.

Of the many methods that have been described for correcting non-linear exposure–disease relationships for the effects of measurement error, both in this dissertation and elsewhere, none seems to have been widely used in practice; the majority of citations for correction methods appearing to come from other methodological papers. Many correction methods only perform well for a limited range of disease models, or place restrictions on the type of additional exposure data we must have available. There is a real need to identify those methods for non-linear exposure–disease relationships that are most promising, and well written advice for epidemiologists should be provided so that the methods may be adopted.

As argued above, continuous linear predictors are strongly preferred over grouped exposure analyses. If grouped exposure analyses are to persist in the analysis of epidemiological data then it is essential that effective methods for measurement error correction exist. In chapter 5 we showed that MacMahon’s simple method for correcting grouped exposure analyses, which has been used by large pooling studies [36, 189], does not give a corrected shape for the exposure–disease relationship in general, unless the relationship is linear and only allows us to view the relationship over a reduced range. Therefore we recommend that MacMahon’s method is not used for final analyses, but as we saw in chapter 3 it may give a good approximation to the shape of the relationship in many circumstances and therefore may be suitable for an initial ‘quick and dirty’ analysis of the data.

Since MacMahon’s method does not provide adequate measurement error correction, alternative methods are required. In chapter 5 we showed that if the exposure–disease relationship is linear then moment reconstruction or multiple imputation can be used. These two methods do not suffer from a reduction in the range over which the relationship can be viewed, unlike MacMahon’s method. Multiple imputation and moment reconstruction are therefore promising methods, to take forward to consider for non-linear exposure–disease relationships; we leave this for future work.

Most correction methods for non-linear relationships are complex to implement. Guolo [150] points out that the development of user-friendly software is key for the diffusion of promising measurement error correction methods; especially as correction for measurement error may form only one part of a complicated analysis. Although it is becoming more common for authors to publish code as supplementary material to papers, this code often is specific to the examples or simulations in the paper and not easy for others to amend for their own purposes.

---

This approach also relies on people being aware of the existence of the material. Authors should look to publish flexible code that performs measurement error correction as a contributed package for the programme in which it was produced.

There are currently three packages for R which deal with measurement error: `simex`, `gbevc` and `decon`. The `simex` package performs SIMEX and MCSIMEX, however, the SIMEX method only handles measurement error in terms included linearly (in which case simpler and more effective methods generally exist). The `gbevc` package [286] uses boosted regression trees to correct for measurement error and the `decon` [287] package uses deconvolution to estimate a measurement error corrected Nadaraya-Watson type kernel regression estimator [185, 288]. The latter two methods, although are able to correct for the effects of measurement error in non-linear exposure–disease relationships, do not appear to have been used in practice. Commands `regcal` and `eivreg`, and the package `meerror` all feature in STATA but none perform measurement error correction of non-linear relationships. We find it surprising that so few methods have been implemented given that so many analyses of non-linear exposure–disease relationships, subject to measurement error are carried out. Software for correcting non-linear relationships for the effects of measurement error is sorely needed because of the computational complexities of implementing methods.

Relationships that have been corrected for exposure measurement error should be reported making sure that assumptions are ‘clearly stated and evidence presented in their support, perhaps with a sensitivity analysis. Large corrections presented without such supporting evidence are unsound’ [48]. We believe it is also important that the observed relationship is made available to the reader either within the main report, or as supplementary material. We feel that this approach will allow the reader to have a better understanding of the correction for measurement error, and would help to generate debate on the topic.

### 10.3.3 Final remarks on measurement error

We have seen throughout this dissertation that it is essential to acknowledge the presence of measurement error, the magnitude and structure of the error, and its effect on the exposure–disease relationship. Measurement error needs to be corrected for and we have provided some methods for doing this. We believe that there needs to be a more unified approach to measurement error in practice, from diagnosis to explanation, to methods for correction, and for reporting of exposure–disease relationships that are subject to measurement error. Statisticians can help by following up methodological papers that introduce a new method with papers in applied journals, which give clear examples, and associated code. This will make it easier for epidemiologists to adopt new methods.

A recent development comes from the Public Population Project in Genomics (P<sup>3</sup>G) [289] who are currently performing a systematic review of the methodological literature on methods to correct for measurement error in epidemiological studies, and of studies that have produced

## 10. Discussion

---

measurement error corrected diet–disease relationships. They aim to publish a comparison chart on how to perform calibration studies and how to obtain correction factors, and ultimately develop an inventory of the typical correction factors used for various nutrients. Having a table of typical correction factors may make measurement error correction less controversial, however this is no substitution for a validation substudy. If such a table were used then it would be important to make sure the methods employed for measuring the exposure in those studies for which the factors were calculated, were the same as the study that was using the factor. We see this as a positive step forward, as it will hopefully provide greater awareness of the effects and correction methods for measurement error amongst the P<sup>3</sup>G collaborators and beyond. Methods that are adopted by high profile research groups such as EPIC, the ERFC and P<sup>3</sup>G have the potential to influence the use of measurement error correction methods amongst epidemiologists.

We would like to highlight again in this final section on measurement error that it is preferable to reduce the measurement error in the observed exposure as this has big advantages, such as greater power to detect non-linearity. Within-person variation will always exist therefore there will always be a need for measurement error correction.

### 10.4 Meta-analysis

As discussed in chapter 8, meta-analysis allows us to make better use of existing research. IPD meta-analysis is the gold standard approach because it allows us to treat each study consistently and it is theoretically a more principled approach. Meta-analysis gives us much greater power to detect non-linearity than in individual studies, and therefore when we correct for measurement error we can have greater certainty that the correction we make is appropriate.

Sauerbrei and Royston [67] provide an excellent starting place for answering the two key questions in 2-stage IPD meta-analysis of continuous exposure–disease relationships: how to choose a model for each study and how to combine non-linear exposure–disease relationships across studies. We have developed upon this foundation by considering how we may choose P-spline models for each study, how we may combine curves across studies using either multivariate meta-analysis of the model parameters or pointwise derivative pooling, and how we can show heterogeneity in the shape of the exposure–disease relationship between studies more clearly.

Although the pointwise pooling rule suggested by Sauerbrei and Royston [67] may be simple, we saw in chapter 8 that it can provide misleading pooled exposure–disease relationships, and therefore we believe it should not in general be the only pooling rule considered. With the availability of `mvmeta` in STATA and R for multivariate meta-analyses it is easy to implement the parameter pooling rule, if only as a sensitivity analysis. The pointwise derivative pooling rule does not rely on the choice of reference value, but suffers from the problem that the

---

pointwise standard errors are difficult to compute, especially when we are combining across different model types such as in chapter 9. We prefer P-spline models because under the pointwise pooling rules their confidence intervals become extremely large outside the range of the data, whereas this is not true for fractional polynomials meaning that large studies can still have a significant influence on the pooled relationship well beyond the range of the data.

With the growing number of epidemiological studies that are sufficiently powered to detect non-linearity, and the growth of large IPD meta-analyses, further theoretical development of methods that allow us to combine non-linear relationships across studies will be required.

## 10.5 Limitations

In this dissertation we have focused almost exclusively on a continuous exposure. Many exposures are not purely continuous, such as alcohol consumption where we have a group of individuals who are non-consumers. The models we have considered in this dissertation do not extend to these situations. For the case where we have non-consumers a good place to start may be the extension of fractional polynomial modelling to include a non-exposed group [290]. We would also need a different, more complicated, model for the measurement error process.

Although we have mainly focused on classical measurement error in this dissertation, there are many practical situations where the classical measurement error model does not hold. For example, Heitmann *et al.* [291] describes a situation in nutritional epidemiology where under-reporting of total energy intake increases with increasing obesity. These situations will require their own measurement error models.

In the application of our methods in chapters 6, 8 and 9 we have only considered correction for measurement error in the exposure of interest, although confounders in our models have also been subject to measurement error. Correcting for measurement error in multiple exposures can be complex and computer intensive, but should be carried out. In our analyses we only used the first repeat observation in our models, further repeats could have allowed us to reduce the variability in our measurement error model, however in many studies only one repeat observation may be available.

We have not taken into account the uncertainty in estimated parameters from the measurement error model in our disease model. As we have previously discussed, this can be done using methods such as bootstrapping, or sandwich estimators, but we believe the impact of this to be only minor when the size of the validation study is reasonably large in comparison to the overall study size. In practice there are usually greater sources of uncertainty in our models.

In chapter 7 we used a mixture model to estimate the distribution of true FBG given observed FBG. Apart from the uncertainty in the measurement error model which we did not take account of, there was large uncertainty in the measurement error model for those studies without

repeats. This is because there is no natural ordering of the components in the models for those studies with repeats.

In chapter 9 we saw that there was significant heterogeneity in the RDR between studies—the COPEN study had a particularly low RDR. We speculated that this was due to the assay method used. To test this hypothesis we could either use data from a study that had used multiple assay methods, however no study did this; or alternatively we could have performed subgroup analyses but with only 6 studies with repeats this was not possible.

In this dissertation we have only considered some of the many approaches that have been proposed for measurement error correction. We have tried to pick out what we believe are the most promising methods, in terms of the methods being relatively simple yet effective. Given that a plethora of methods have been proposed there may be other methods that we would have benefited from considering in this dissertation.

## 10.6 Areas for future work

### 10.6.1 Modelling non-linear exposure–disease relationships

Spline based models should be made more accessible for epidemiologists. Firstly, there needs to be a clearer exposition of the different splines available. Secondly we believe that code should be available so that modellers are able to fit the same spline with the same options, whatever their choice of model, and across different software packages. A second area we think could receive greater attention is the threshold shaped relationship, which was not well modelled by any of the methods we considered. It would be of interest to find a way of identifying when such a relationship is present, especially when the exposure is subject to measurement error.

### 10.6.2 Effects of measurement error

We believe that heteroscedastic and non-normal measurement error are probably more common than has been reported in the literature. We suggest that the diagnostic plots proposed by Carroll *et al.* [33] are used, and the findings reported. In chapter 7 we considered the robustness of our models to non-classical error structures and there seemed to be some evidence that moderately heteroscedastic measurement error does not introduce large bias into regression calibration based methods. More research into the identification of non-classical error, and its correction, is needed.

In chapter 7 we saw that significant bias can be caused if there is a minimum or maximum value that the true exposure can take, but there are observations that lie beyond this value due to measurement error. Although this minimum/maximum true exposure limit may exist, it may not be known, and an interesting problem would be to estimate it.

---

### 10.6.3 Measurement error correction

In chapter 5 we considered correction methods for categorised continuous exposures when the exposure–disease relationship was linear and found that multiple imputation and moment reconstruction performed well. An important area for further research is to investigate how these methods perform when the exposure–disease relationship is non-linear. It would also be of interest to find out how these methods perform for non-linear exposure–disease relationships where the exposure is not categorised, as considered by Freedman *et al.* [202] for linear relationships. We could investigate both the uncategorised and categorised situation by modifying the simulation studies of chapters 3 and 5 respectively. It may be the case that when the relationship is non-linear, moment reconstruction does not work because it only matches the first two moments of the data. It might therefore be wise to additionally consider the recently proposed method, moment adjusted imputation, of Thomas, Stefanski, and Davidian [55] which matches higher moments.

Correcting the structural fractional polynomial and P-spline models for the exposure–disease relationship when the true exposure has a natural minimum or maximum may be an interesting research avenue.

Carroll *et al.* [157] did not consider a single normal for the distribution of true exposure given observed in their simulation study when they introduced structural P-splines, choosing only to consider a normal mixture model instead. It would be interesting to see how much more robust using mixtures of normals, or other flexible distributions, for the distribution of true exposure given observed is, and whether the significant computational challenges in implementation are actually worthwhile. We saw in chapter 7 that normal mixture models for the distribution of true exposure given observed cannot be easily pooled across studies to obtain a model for studies without repeat measurements. We therefore suggest that future work could consider robust models for the distribution of true exposure given observed in greater detail, and approaches for estimating a model that is suitable for studies without repeat measurements.

### 10.6.4 Meta-analysis

The methods presented in chapter 8 are somewhat pragmatic and *ad hoc*. We believe that there is much fruitful research to be carried out in putting this topic on a firmer theoretical footing. Although it is clear that choosing a model for each study using the studywise (varying *df*) model selection rule is a bad approach to take, as it does not allow us to borrow strength across studies, it is not clear which of the other two selection rules we should take. A simulation study could be conducted to investigate which of the two approaches performs best. Following on from this, we suggest that the number of degrees of freedom in each of the models fitted to the individual studies should be investigated.

Simulation could be used to investigate which of the pooling rules provide the best estimate

of the true exposure–disease relationship. This would also allow us to explore the different strengths and weaknesses of the pooling rules. In chapter 8 we proposed some ways of attempting to visualise various aspects of our meta-analyses, however as Jackson *et al.* comment ‘how to attempt to display all aspects of high-dimensional meta-analyses, and produce multivariate funnel and forest plots for example, remains an open question’ [255].

### 10.7 Conclusion

In chapter 1 we said that this dissertation was motivated by the observation that ‘characterizing the shape of the underlying exposure–disease relationship, while taking into account possibly heterogeneous measurement error, is not well studied, especially in the context of IPD meta-analysis’ [15]. We have shown the effects of exposure measurement error, both classical and non-classical; shown how we may correct for its effects; and how we can combine exposure–disease relationships across studies. We therefore hope that this dissertation goes some way towards meeting the needs identified by Thompson *et al.* [15].

# Appendix A

## Appendix to chapter 4

### A.1 Calculation of basis functions for structural P-splines when $X|W$ is normally distributed

Under the structural P-spline approach of Carroll *et al.* [157] we need to be able to calculate expectations of the form  $E((X - k)_+^p | W)$ . In the case that we assume that the distribution of  $X|W$  is normal, we can calculate these expectations exactly given the mean and variance of the distribution.

Consider the general case where we have  $x \sim N(\mu_x, \sigma_x^2)$  and we wish to calculate  $E(x^p | x > k)$  for some constant  $k$ . We shall show that we can generate a recursive formula to calculate this expectation.

Let  $\epsilon = \frac{x-\mu}{\sigma}$  and  $h = \frac{k-\mu}{\sigma}$  then

$$\begin{aligned} E(x^p | x \geq k) &= \frac{1}{1 - \Phi(h)} \int_h^\infty x^p f(x) dx \\ &= \frac{1}{1 - \Phi(h)} \int_h^\infty (\mu + \sigma\epsilon)^p \phi(\epsilon) d\epsilon \end{aligned} \quad (\text{A.1})$$

$$(\mu + \sigma\epsilon)^p = \sum_{r=0}^p \binom{p}{r} \mu^{p-r} \sigma^r \epsilon^r \quad (\text{A.2})$$

and let  $I_r = \frac{1}{1-\Phi(h)} \int_h^\infty \epsilon^r \phi(\epsilon) d\epsilon$  so

$$E(x^p | x \geq k) = \sum_{r=0}^p \binom{p}{r} \mu^{p-r} \sigma^r I_r. \quad (\text{A.3})$$

## A. Appendix to chapter 4

---

Let  $s = \epsilon^{r-1}$  and  $dt = \epsilon\phi(\epsilon)$  and integrate by parts to obtain the recursive formula

$$\begin{aligned} I_r &= \frac{1}{1 - \Phi(h)} \int_h^\infty \epsilon^r \phi(\epsilon) d\epsilon \\ &= \frac{h^{r-1}}{1 - \Phi(h)} \phi(h) + (r-1)I_{r-2}. \end{aligned} \quad (\text{A.4})$$

We have the initial values

$$\begin{aligned} I_0 &= \frac{1}{1 - \Phi(h)} \int_h^\infty \phi(\epsilon) d\epsilon = 1 \\ I_1 &= \frac{1}{1 - \Phi(h)} \int_h^\infty \epsilon \phi(\epsilon) d\epsilon = \frac{\phi(h)}{1 - \Phi(h)}. \end{aligned}$$

Let  $\rho(h) = \frac{\phi(h)}{1 - \Phi(h)}$  then we have for  $p = 3$

$$\begin{aligned} I_0 &= 1 \\ I_1 &= \rho(h) \\ I_2 &= h\rho(h) + 1 \\ I_3 &= h^2\rho(h) + 2\rho(h). \end{aligned}$$

Plugging this into equation A.3 we obtain

$$\begin{aligned} E(x^3|x \geq k) &= \mu^3 I_0 + 3\mu^2\sigma I_1 + 3\mu\sigma^2 I_2 + \sigma^3 I_3 \\ &= \mu^3 + 3\mu^2\sigma\rho(h) + 3\mu\sigma^2(1 + h\rho(h)) + \sigma^3(2\rho(h) + h^2\rho(h)). \end{aligned} \quad (\text{A.5})$$

Instead of calculating  $E(x^p|x \geq k)$  we want to be able to calculate  $E((x-k)^p|x \geq k)$ . If we let  $y = x - k$  then we can see

$$E((x-k)^p|x \geq k) = E((x-k)^p|x-k > 0) = E(y^p|y > 0)$$

where  $\mu_y = \mu_x - k$  and  $\sigma_y^2 = \sigma_x^2$  with cutpoint for  $y$ ,  $k_y = 0$ . Therefore,  $h_y = \frac{0 - (\mu_x - k)}{\sigma_x}$  and putting this all in to equation A.5 we can calculate  $E((x-k)^p|x \geq k) = E((x-k)_+^p)$

$$\begin{aligned} E((x-k)_+^p) &= P(y < 0)E(y^p|y < 0) + P(y > 0)E(y^p|y > 0) \\ &= 0 + (1 - \Phi(h_y))E(y^p|y > 0). \end{aligned} \quad (\text{A.6})$$

For  $p = 3$ ,

$$E((x-k)_+^3) = (1 - \Phi(h_y)) \left[ \mu_y^3 + 3\mu_y^2\sigma_y\rho(h_y) + 3\mu_y\sigma_y^2(1 + h_y\rho(h_y)) + \sigma_y^3(2\rho(h_y) + h_y^2\rho(h_y)) \right].$$

This can easily be coded in R and code is given in appendix F.

## A.2 Calculation of $E(X^p \log X)$

For structural fractional polynomial models where the distribution of  $X|W$  is log-normally distributed we need to be able to calculate  $E(X^p \log X|W)$  for the case when our FP2 powers are equal and non-zero.

For integer  $p$  greater than one,

$$\begin{aligned}
 E(X^p \log X) &= \int_0^\infty \frac{x^p \log x}{x \sigma \sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right) dx \\
 &= \int_{-\infty}^\infty \frac{u}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(u - \mu)^2}{2\sigma^2} + pu\right) du \\
 &= \exp\left(\frac{p\sigma^2(p\sigma^2 + 2\mu)}{2\sigma^2}\right) \int_{-\infty}^\infty \frac{u}{\sigma \sqrt{2\pi}} \exp\left(-\frac{u - (\mu + p\sigma^2)^2}{2\sigma^2}\right) du \\
 &= (\mu + p\sigma^2) \exp\left(\frac{p^2\sigma^2}{2} + \mu p\right).
 \end{aligned}$$

Since  $X|W \sim N(\lambda \log W + (1 - \lambda)\mu_x, (1 - \lambda)\sigma_x^2)$ , it implies

$$E(X^p \log X|W) = k(p) [W^{\lambda p} \log(W^\lambda) + (1 - \lambda)(\mu_x + p\sigma_x^2)W^{\lambda p}]$$

where  $k(p) = \exp\left((1 - \lambda)\left(\frac{\sigma_x^2 p^2}{2} + p\mu_x\right)\right)$ .

## A.3 Proof that $\lim_{\zeta \rightarrow -1} \text{Var}(\hat{\beta}_b(\zeta)) = 0$

This section provides the proof that  $\lim_{\zeta \rightarrow -1} \text{Var}(\hat{\beta}_b(\zeta)) = 0$  which is used in deriving the jackknife variance estimate of the SIMEX estimator.

Firstly we show, using a different approach to Stefanski and Cook [166], that if  $Z_1$  and  $Z_2$  are standard normal random variables then  $E((Z_1 + iZ_2)^n) = 0$  for  $n = 1, 2, \dots$

Consider the moment generating function for  $Z_1 + iZ_2$ :

$$\begin{aligned}
 M_{Z_1+iZ_2}(t) &= E(e^{t(Z_1+iZ_2)}) \\
 &= E(e^{tZ_1+itZ_2}) \\
 &= E(e^{tZ_1}) E(e^{itZ_2}) \\
 &= e^{\frac{1}{2}t^2} e^{-\frac{1}{2}t^2} \\
 &= 1.
 \end{aligned}$$

## A. Appendix to chapter 4

---

Hence,  $\frac{dM_{Z_1+iZ_2}(t)}{dt} = 0$  for all  $n = 0, 1, 2, \dots$  i.e.  $E((Z_1 + iZ_2)^n) = 0$  for  $n = 1, 2, \dots$

We can use the result above to show that  $\lim_{\zeta \rightarrow -1} E(f(\mu + \sigma(Z_1 + \sqrt{\zeta}Z_2))) = f(\mu)$  as follows [166],

$$\begin{aligned} \lim_{\zeta \rightarrow -1} E(f(\mu + \sigma(Z_1 + \sqrt{\zeta}Z_2))) &= \lim_{\zeta \rightarrow -1} E \left( f(\mu) + \sum_{n=1}^{\infty} \frac{f^{(n)}(\mu)\sigma^n}{n!} (Z_1 + \sqrt{\zeta}Z_2)^n \right) \\ &= E \left( f(\mu) + \sum_{n=1}^{\infty} \frac{f^{(n)}(\mu)\sigma^n}{n!} (Z_1 + iZ_2)^n \right) \\ &= f(\mu) + \sum_{n=1}^{\infty} \frac{f^{(n)}(\mu)\sigma^n}{n!} E((Z_1 + iZ_2)^n) \\ &= f(\mu). \end{aligned}$$

So,

$$\begin{aligned} \lim_{\zeta \rightarrow -1} \text{Var}(\hat{\beta}_b(\zeta)) &= \lim_{\zeta \rightarrow -1} E((\hat{\beta}_b(\zeta))^2) - E(\hat{\beta}_b(\zeta))^2 \\ &= \hat{\beta}_b(-1)^2 - \hat{\beta}_b(-1)^2 \\ &= 0. \end{aligned}$$

# Appendix B

## Appendix to chapter 5

### B.1 Simulation study 2

#### B.1.1 Results for linear regression

We repeated the simulations of 5.2 for a linear regression model where the response  $Y_i$  for each individual was generated according to  $Y_i = X_i + \epsilon_i$  where  $\epsilon_i$  is normally distributed with mean 0 and variance  $\sigma_\epsilon^2 = 0.1^2, 1$ .

The results from this simulation are in tables B.1–B.4. They are very similar to those found for the logistic model in section 5.2

#### B.1.2 Results for multiple imputation and moment reconstruction when the exposure is grouped into quintiles

In chapter 5 we found that multiple imputation and moment reconstruction performed well in correcting for the effects of measurement error in a grouped exposure analysis of a linear exposure–disease relationship using logistic regression. In this section we investigate whether these two methods still perform well when the continuous exposure is split into quintile groups, as is often done in epidemiological investigations.

Suppose that we split the observed true exposure  $X$  into quintile groups  $\mathbf{Q} = \{Q_1, \dots, Q_5\}$  and the disease model we wish to estimate is

$$g(E(Y|\mathbf{Q})) = \beta_0 + \beta_1 Q_1 + \beta_2 Q_2 + \beta_3 Q_4 + \beta_4 Q_5$$

where the central quintile group  $Q_3$  is the reference category.

The observed relationship would then be obtained by fitting the model

$$g(E(Y|Q^W)) = \beta_0^* + \beta_1^*Q_1^W + \beta_2^*Q_2^W + \beta_3^*Q_4^W + \beta_4^*Q_5^W$$

where  $Q^W = \{Q_1^W, \dots, Q_5^W\}$  are quintile groups of  $W$ .

The only differences in the application of the two measurement error correction methods when considering quintiles of exposure are as follows:

**Moment Reconstruction (MR)**—Instead of forming  $X_C^{MR}$  we form quintile groups of  $X^{MR}$ ,  $\mathbf{X}_Q^{MR} = \{X_{Q_1}^{MR}, \dots, X_{Q_5}^{MR}\}$ , and fit the disease model

$$g(E(Y|\mathbf{X}_Q^{MR})) = \beta_0^{MR} + \beta_1^{MR}X_{Q_1}^{MR} + \beta_2^{MR} + X_{Q_2}^{MR} + \beta_3^{MR}X_{Q_4}^{MR} + \beta_4^{MR}X_{Q_5}^{MR}.$$

**Multiple imputation (MI)**—Instead of forming  $X_C^{MI(k)}$  for each imputed dataset we form quintile groups of  $X^{MI(k)}$ ,  $\mathbf{X}_Q^{MI(k)} = \{X_{Q_1}^{MI(k)}, \dots, X_{Q_5}^{MI(k)}\}$ , and fit the disease model

$$g(E(Y|\mathbf{X}_Q^{MI(k)})) = \beta_0^{MI(k)} + \beta_1^{MI(k)}X_{Q_1}^{MI(k)} + \beta_2^{MI(k)} + X_{Q_2}^{MI(k)} + \beta_3^{MI(k)}X_{Q_4}^{MI(k)} + \beta_4^{MI(k)}X_{Q_5}^{MI(k)}.$$

The data for the simulation were generated in exactly the same was as in chapter 5. The results from this simulation are in tables B.5–B.8. We can see that both multiple imputation and moment reconstruction continue to provide almost unbiased parameter estimates for the true quintiles.

## B.2 Simulation study 3: RFrMSE

The reference free root mean squared error (RFrMSE) is an attempt to give a global measure of of the performance of the structural fractional polynomial and P-spline models for the measurement error corrected exposure–disease relationship without dependence on the arbitrary choice of reference value.

$$\text{RFrMSE} = \sqrt{\frac{1}{n^2} \sum_{j=1}^n \sum_{i=1}^n \left( \log \hat{\text{HR}}(x_i; x_j) - \log \text{HR}(x_i; x_j) \right)^2}$$

The RFrMSE was evaluated at the percentiles of the distribution of true exposure given observed obtained from the regression calibration model in each simulation of section 5.3.

Figure B.1 gives boxplots of the RFrMSe for structural fractional polynomial and P-spline models for each shape of the exposure–disease relationship considered in the simulation. The results are very similar to what was observed for rMSE (figure 5.5), which suggests that the choice of reference value made in the simulation did not unduly influence the rMSE.

$\sigma_u^2$	$c$	True $\beta_X$	Using $X_C$	Naive	RC1		RC2		MI		MR		gSIMEX		
					10%	50%	10%	50%	10%	50%	10%	50%	10%	50%	
$\sigma_\epsilon^2 = 0.1^2, (\alpha_0 = 0, \alpha_1 = 1)$															
0.25	0	1.596	Mean	1.595	1.426	1.443	1.510	1.945	1.737	1.595	1.595	1.595	1.595	1.577	1.578
			SD	0.018	0.021	0.020	0.019	0.030	0.020	0.018	0.018	0.018	0.018	0.018	0.027
	1	1.813	Mean	1.812	1.581	1.684	1.743	2.383	2.031	1.812	1.812	1.812	1.812	1.782	1.782
			SD	0.020	0.024	0.027	0.023	0.061	0.026	0.020	0.020	0.020	0.020	0.020	0.036
1	0		Mean	1.595	1.127	1.173	1.361	2.051	1.727	1.595	1.595	1.595	1.595	1.413	1.414
			SD	0.018	0.024	0.023	0.022	0.051	0.023	0.018	0.018	0.018	0.018	0.018	0.039
	1		Mean	1.812	1.203	1.485	1.653	2.668	2.033	1.812	1.812	1.812	1.812	1.545	1.546
			SD	0.020	0.028	0.044	0.027	0.102	0.030	0.020	0.020	0.020	0.020	0.020	0.051
$\sigma_\epsilon^2 = 1, (\alpha_0 = 0, \alpha_1 = 1)$															
0.25	0		Mean	1.595	1.425	1.442	1.510	1.944	1.736	1.595	1.594	1.595	1.594	1.576	1.576
			SD	0.034	0.035	0.034	0.035	0.049	0.040	0.042	0.034	0.045	0.035	0.035	0.050
	1		Mean	1.812	1.581	1.683	1.743	2.383	2.030	1.813	1.812	1.812	1.812	1.782	1.782
			SD	0.041	0.042	0.048	0.045	0.081	0.051	0.049	0.041	0.053	0.044	0.044	0.068
1	0		Mean	1.595	1.126	1.172	1.360	2.050	1.726	1.595	1.594	1.594	1.594	1.411	1.412
			SD	0.034	0.037	0.036	0.036	0.071	0.044	0.047	0.035	0.054	0.037	0.037	0.063
	1		Mean	1.812	1.202	1.486	1.652	2.667	2.032	1.813	1.812	1.812	1.812	1.544	1.545
			SD	0.041	0.044	0.066	0.050	0.131	0.058	0.056	0.042	0.066	0.046	0.046	0.079

Table B.1: Mean and standard deviation (SD) of estimates of the slope parameter in a linear regression of  $Y$  on  $X_C$  across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when  $X$  is assumed to be observed in 10% or 50% of the study population. (Scenario (1a))

$\sigma_u^2$	$c$	True $\beta_X$	Using $X_C$	Naive	RC1		RC2		MI		MR		gSIMEX		
					10%	50%	10%	50%	10%	50%	10%	50%	10%	50%	
$\sigma_\epsilon^2 = 0.1^2, (\alpha_0 = 0, \alpha_1 = 0.5)$															
0.25	0	1.596	Mean	1.595	1.127	1.127	1.127	1.436	1.437	1.571	1.573	1.592	1.593	1.271	1.275
			SD	0.018	0.024	0.024	0.024	0.081	0.034	0.033	0.025	0.020	0.019	0.041	0.034
	1	1.813	Mean	1.812	1.431	1.434	1.431	2.709	2.689	1.779	1.781	1.851	1.857	1.376	1.382
			SD	0.020	0.042	0.055	0.042	0.400	0.169	0.045	0.030	0.029	0.023	0.045	0.044
1	0		Mean	1.595	0.712	0.714	0.712	1.326	1.329	1.585	1.591	1.582	1.590	0.858	0.860
			SD	0.018	0.027	0.027	0.028	0.075	0.049	0.028	0.022	0.033	0.023	0.049	0.048
	1		Mean	1.812	0.788	1.180	1.172	2.457	2.421	1.798	1.807	1.818	1.827	0.889	0.891
			SD	0.020	0.034	0.202	0.126	0.336	0.155	0.042	0.028	0.034	0.023	0.054	0.051
$\sigma_\epsilon^2 = 1, (\alpha_0 = 0, \alpha_1 = 0.5)$															
0.25	0		Mean	1.595	1.126	1.126	1.126	1.435	1.436	1.444	1.484	1.448	1.445	1.271	1.274
			SD	0.034	0.037	0.037	0.037	0.090	0.048	0.049	0.034	0.045	0.038	0.061	0.056
	1		Mean	1.812	1.430	1.433	1.430	2.707	2.687	1.605	1.656	1.908	1.904	1.374	1.380
			SD	0.041	0.065	0.076	0.066	0.409	0.194	0.054	0.039	0.076	0.069	0.073	0.070
1	0		Mean	1.595	0.711	0.713	0.711	1.324	1.327	1.715	1.706	1.435	1.429	0.856	0.858
			SD	0.034	0.040	0.039	0.039	0.093	0.072	0.067	0.040	0.084	0.050	0.070	0.069
	1		Mean	1.812	0.787	1.187	1.172	2.453	2.416	1.987	1.971	1.750	1.740	0.888	0.890
			SD	0.041	0.050	0.262	0.182	0.356	0.189	0.082	0.049	0.118	0.075	0.075	0.073

Table B.2: Mean and standard deviation (SD) of estimates of the slope parameter in a linear regression of  $Y$  on  $X_C$  across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when  $W_2$  is assumed to be observed in 10% or 50% of the study population. (Scenario (2a))

$\sigma_u^2$	$c$	True $\beta_X$	Using $X_C$	Naive	RC1		RC2		MI		MR		gSIMEX	
					10%	50%	10%	50%	10%	50%	10%	50%	10%	50%
$\sigma_\epsilon^2 = 0.1^2, (\alpha_0 = 0, \alpha_1 = 0.5)$														
0.25	0	Mean	1.595	1.127	1.173	1.361	2.051	1.727	1.595	1.595	1.595	1.595	1.413	1.414
		SD	0.018	0.024	0.023	0.022	0.051	0.023	0.018	0.018	0.018	0.018	0.039	0.038
	1	Mean	1.812	1.431	1.485	1.653	2.298	1.925	1.812	1.812	1.812	1.812	1.545	1.546
		SD	0.020	0.042	0.044	0.027	0.079	0.029	0.020	0.020	0.020	0.020	0.051	0.048
1	0	Mean	1.595	0.712	0.801	1.153	1.952	1.660	1.595	1.595	1.595	1.595	0.982	0.982
		SD	0.018	0.027	0.026	0.023	0.065	0.025	0.018	0.018	0.018	0.018	0.052	0.051
	1	Mean	1.812	0.788	1.405	1.637	2.372	1.903	1.812	1.812	1.812	1.812	1.020	1.021
		SD	0.020	0.034	0.085	0.028	0.092	0.030	0.020	0.020	0.020	0.020	0.056	0.054
$\sigma_\epsilon^2 = 1, (\alpha_0 = 0, \alpha_1 = 0.5)$														
0.25	0	Mean	1.595	1.126	1.172	1.360	2.050	1.726	1.595	1.594	1.594	1.594	1.411	1.412
		SD	0.034	0.037	0.036	0.036	0.071	0.044	0.047	0.035	0.054	0.037	0.063	0.062
	1	Mean	1.812	1.430	1.486	1.652	2.299	1.924	1.813	1.812	1.812	1.812	1.544	1.545
		SD	0.041	0.065	0.066	0.050	0.110	0.058	0.056	0.042	0.066	0.046	0.079	0.078
1	0	Mean	1.595	0.711	0.800	1.152	1.952	1.659	1.594	1.595	1.593	1.594	0.979	0.981
		SD	0.034	0.040	0.039	0.037	0.092	0.048	0.050	0.035	0.061	0.039	0.076	0.075
	1	Mean	1.812	0.787	1.409	1.637	2.372	1.901	1.812	1.811	1.813	1.812	1.018	1.020
		SD	0.041	0.050	0.124	0.058	0.138	0.061	0.058	0.042	0.075	0.047	0.082	0.080

Table B.3: Mean and standard deviation (SD) of estimates of the slope parameter in a linear regression of  $Y$  on  $X_C$  across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when  $X$  is assumed to be observed in 10% or 50% of the study population. (Scenario (2a))

$\sigma_u^2$	$c$	True $\beta_X$	Using $X_C$	Naive	RC1		RC2		MI		MR		gSIMEX		
					10%	50%	10%	50%	10%	50%	10%	50%	10%	50%	
$\sigma_\epsilon^2 = 0.1^2, (\alpha_0 = 0, \alpha_1 = 0.5)$															
0.25	0	1.596	Mean	1.595	1.127	1.127	1.127	1.436	1.437	1.594	1.595	1.594	1.595	1.409	1.413
			SD	0.018	0.024	0.024	0.024	0.081	0.034	0.019	0.019	0.019	0.018	0.067	0.039
	1	1.813	Mean	1.812	1.431	1.434	1.431	2.709	2.689	1.811	1.811	1.811	1.812	1.537	1.544
			SD	0.020	0.042	0.055	0.042	0.400	0.169	0.025	0.021	0.025	0.021	0.065	0.047
1	0		Mean	1.595	0.712	0.714	0.712	1.326	1.329	1.585	1.591	1.586	1.592	0.980	0.982
			SD	0.018	0.027	0.027	0.028	0.075	0.049	0.029	0.022	0.028	0.022	0.053	0.050
	1		Mean	1.812	0.788	1.180	1.172	2.457	2.421	1.798	1.806	1.799	1.807	1.019	1.021
			SD	0.020	0.034	0.202	0.126	0.336	0.155	0.043	0.028	0.042	0.027	0.054	0.053
$\sigma_\epsilon^2 = 1, (\alpha_0 = 0, \alpha_1 = 0.5)$															
0.25	0		Mean	1.595	1.126	1.126	1.126	1.435	1.436	1.596	1.594	1.595	1.593	1.408	1.413
			SD	0.034	0.037	0.037	0.037	0.090	0.048	0.053	0.035	0.047	0.037	0.084	0.062
	1		Mean	1.812	1.430	1.433	1.430	2.707	2.687	1.815	1.811	1.814	1.811	1.534	1.542
			SD	0.041	0.065	0.076	0.066	0.409	0.194	0.062	0.042	0.062	0.051	0.092	0.077
1	0		Mean	1.595	0.711	0.713	0.711	1.324	1.327	1.599	1.594	1.598	1.593	0.979	0.980
			SD	0.034	0.040	0.039	0.039	0.093	0.072	0.077	0.043	0.074	0.045	0.076	0.074
	1		Mean	1.812	0.787	1.187	1.172	2.453	2.416	1.819	1.811	1.820	1.812	1.017	1.019
			SD	0.041	0.050	0.262	0.182	0.356	0.189	0.094	0.053	0.100	0.062	0.080	0.079

Table B.4: Mean and standard deviation (SD) of estimates of the slope parameter in a linear regression of  $Y$  on  $X_C$  across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when  $W_2, W_3$  are assumed to have been observed in 10% or 50% of the study population. (Scenario (2b))

$\sigma_u^2$		Using $X_C$				Naive				MI				MR			
		$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$
$\beta = \log(1.5), (\alpha_0 = 0, \alpha_1 = 1)$																	
0.25	Mean	-0.55	-0.21	0.22	0.59	-0.50	-0.19	0.20	0.52	-0.55	-0.21	0.22	0.58	-0.55	-0.21	0.22	0.59
	SD	0.20	0.18	0.17	0.16	0.20	0.17	0.17	0.16	0.17	0.14	0.13	0.14	0.20	0.18	0.17	0.16
1	Mean	-0.55	-0.21	0.22	0.59	-0.39	-0.15	0.16	0.41	-0.55	-0.21	0.22	0.58	-0.55	-0.21	0.22	0.59
	SD	0.20	0.18	0.17	0.16	0.18	0.18	0.16	0.15	0.17	0.14	0.13	0.14	0.21	0.18	0.17	0.17
$\beta = \log(2), (\alpha_0 = 0, \alpha_1 = 1)$																	
0.25	Mean	-0.93	-0.36	0.38	1.01	-0.84	-0.32	0.34	0.89	-0.93	-0.36	0.37	1.00	-0.93	-0.36	0.38	1.00
	SD	0.22	0.19	0.17	0.15	0.22	0.18	0.16	0.15	0.20	0.15	0.13	0.14	0.23	0.19	0.16	0.16
1	Mean	-0.93	-0.36	0.38	1.01	-0.66	-0.25	0.26	0.69	-0.92	-0.36	0.37	1.00	-0.93	-0.36	0.38	1.00
	SD	0.22	0.19	0.17	0.15	0.19	0.18	0.16	0.15	0.18	0.15	0.12	0.14	0.23	0.19	0.16	0.16

Table B.5: Mean and standard deviation (SD) of estimates of the parameters for each group in a logistic regression of  $Y$  on  $Q = \{Q_1, Q_2, Q_3, Q_4, Q_5\}$  across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when  $X$  is assumed to be observed in 10% or 50% of the study population.  $Q_3$  is the reference category. (Scenario (1a))

$\sigma_u^2$		Using $X_C$				Naive				MI				MR			
		$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$
$\beta = \log(1.5), (\alpha_0 = 0, \alpha_1 = 1)$																	
0.25	Mean	-0.55	-0.21	0.22	0.59	-0.50	-0.19	0.20	0.52	-0.56	-0.22	0.22	0.59	-0.55	-0.21	0.22	0.58
	SD	0.20	0.18	0.17	0.16	0.20	0.17	0.17	0.16	0.14	0.06	0.06	0.12	0.14	0.07	0.07	0.12
1	Mean	-0.55	-0.21	0.22	0.59	-0.39	-0.15	0.16	0.41	-0.56	-0.22	0.22	0.58	-0.55	-0.21	0.22	0.58
	SD	0.20	0.18	0.17	0.16	0.18	0.18	0.16	0.15	0.17	0.07	0.07	0.18	0.12	0.05	0.05	0.12
$\beta = \log(2), (\alpha_0 = 0, \alpha_1 = 1)$																	
0.25	Mean	-0.93	-0.36	0.38	1.01	-0.84	-0.32	0.34	0.89	-0.94	-0.37	0.37	1.01	-0.93	-0.37	0.37	1.00
	SD	0.22	0.19	0.17	0.15	0.22	0.18	0.16	0.15	0.15	0.07	0.06	0.12	0.16	0.07	0.07	0.11
1	Mean	-0.93	-0.36	0.38	1.01	-0.66	-0.25	0.26	0.69	-0.94	-0.37	0.37	1.01	-0.93	-0.36	0.37	0.99
	SD	0.22	0.19	0.17	0.15	0.19	0.18	0.16	0.15	0.17	0.07	0.07	0.19	0.13	0.05	0.05	0.12

Table B.6: Mean and standard deviation (SD) of estimates of the parameters for each group in a logistic regression of  $Y$  on  $Q = \{Q_1, Q_2, Q_3, Q_4, Q_5\}$  across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when  $W_2$  is assumed to be observed in 10% or 50% of the study population.  $Q_3$  is the reference category. (Scenario (1b))

$\sigma_u^2$	Using $X_C$				Naive				MI				MR			
	$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$
$\beta = \log(1.5), (\alpha_0 = 0, \alpha_1 = 0.5)$																
Mean	-0.55	-0.21	0.22	0.59	-0.39	-0.15	0.16	0.41	-0.55	-0.21	0.22	0.58	-0.55	-0.21	0.22	0.59
SD	0.20	0.18	0.17	0.16	0.18	0.18	0.16	0.15	0.17	0.14	0.13	0.14	0.21	0.18	0.17	0.17
Mean	-0.55	-0.21	0.22	0.59	-0.25	-0.09	0.10	0.25	-0.54	-0.21	0.22	0.58	-0.55	-0.21	0.22	0.58
SD	0.20	0.18	0.17	0.16	0.17	0.17	0.16	0.15	0.17	0.14	0.13	0.15	0.22	0.19	0.17	0.18
$\beta = \log(2), (\alpha_0 = 0, \alpha_1 = 0.5)$																
1 Mean	-0.93	-0.36	0.38	1.01	-0.66	-0.25	0.26	0.69	-0.92	-0.36	0.37	1.00	-0.93	-0.36	0.38	1.00
SD	0.22	0.19	0.17	0.15	0.19	0.18	0.16	0.15	0.18	0.15	0.12	0.14	0.23	0.19	0.16	0.16
Mean	-0.93	-0.36	0.38	1.01	-0.42	-0.16	0.16	0.43	-0.92	-0.36	0.37	1.00	-0.92	-0.36	0.37	1.00
SD	0.22	0.19	0.17	0.15	0.17	0.16	0.15	0.14	0.18	0.14	0.13	0.14	0.24	0.20	0.16	0.17

Table B.7: Mean and standard deviation (SD) of estimates of the parameters for each group in a logistic regression of  $Y$  on  $Q = \{Q_1, Q_2, Q_3, Q_4, Q_5\}$  across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when  $X$  is assumed to be observed in 10% or 50% of the study population.  $Q_3$  is the reference category. (Scenario (2a))

$\sigma_u^2$		Using $X_C$				Naive				MI				MR			
		$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$
$\beta = \log(1.5), (\alpha_0 = 0, \alpha_1 = 0.5)$																	
0.25	Mean	-0.55	-0.21	0.22	0.59	-0.39	-0.15	0.16	0.41	-0.56	-0.22	0.22	0.58	-0.55	-0.21	0.22	0.58
	SD	0.20	0.18	0.17	0.16	0.18	0.18	0.16	0.15	0.20	0.08	0.08	0.21	0.14	0.07	0.07	0.13
1	Mean	-0.55	-0.21	0.22	0.59	-0.25	-0.09	0.10	0.25	-0.56	-0.22	0.22	0.59	-0.56	-0.22	0.22	0.58
	SD	0.20	0.18	0.17	0.16	0.17	0.17	0.16	0.15	0.27	0.11	0.11	0.29	0.14	0.06	0.06	0.14
$\beta = \log(2), (\alpha_0 = 0, \alpha_1 = 0.5)$																	
0	Mean	-0.93	-0.36	0.38	1.01	-0.66	-0.25	0.26	0.69	-0.94	-0.37	0.37	1.00	-0.93	-0.36	0.37	1.00
	SD	0.22	0.19	0.17	0.15	0.19	0.18	0.16	0.15	0.20	0.08	0.08	0.21	0.15	0.07	0.07	0.12
1.000	Mean	-0.93	-0.36	0.38	1.01	-0.42	-0.16	0.16	0.43	-0.93	-0.37	0.37	1.01	-0.93	-0.37	0.37	1.00
	SD	0.22	0.19	0.17	0.15	0.17	0.16	0.15	0.14	0.26	0.11	0.11	0.30	0.14	0.06	0.06	0.14

Table B.8: Mean and standard deviation (SD) of estimates of the parameters for each group in a logistic regression of  $Y$  on  $Q = \{Q_1, Q_2, Q_3, Q_4, Q_5\}$  across 1000 simulated data sets using the true exposure, the naive method, and different correction methods when  $W_2, W_3$  are assumed to be observed in 10% or 50% of the study population.  $Q_3$  is the reference category. (Scenario (2b))

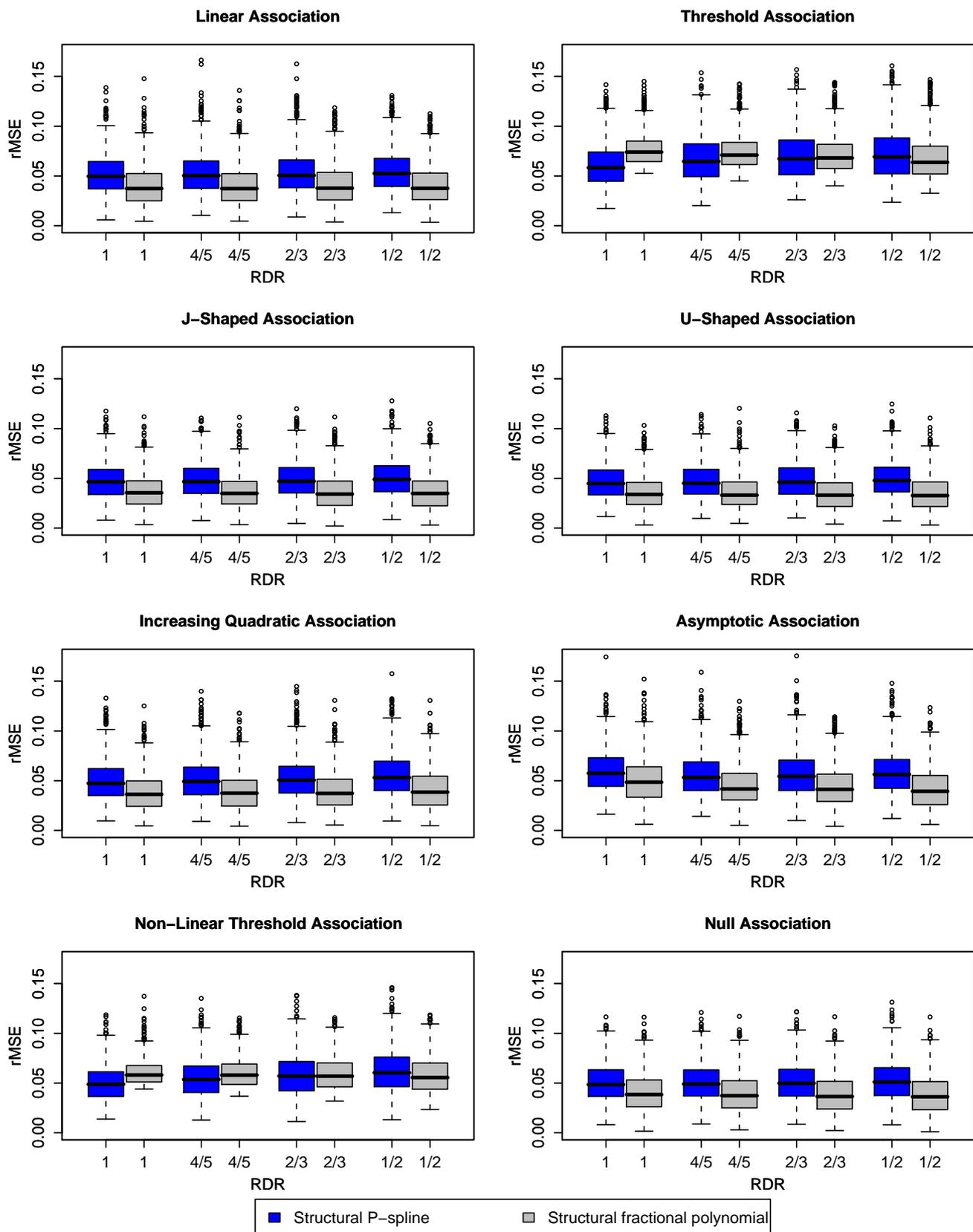


Figure B.1: Boxplots of RFrMSE for structural fractional polynomial and P-spline models for the exposure–disease relationship.



# Appendix C

## Appendix to chapter 6

### C.1 ERFC FBG data — study names

ALLHAT	Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial
ARIC	Atherosclerosis Risk in Communities Study
BHS	Busselton Health Study
BRUN	Bruneck Study
BUPA	BUPA Study
BWHHS	British Women's Heart and Health Study
CASTEL	Cardiovascular Study in the Elderly
CHARL	Charleston Heart Study
CHS1	Original cohort of the Cardiovascular Health Study
CHS2	Supplemental African-American cohort of the Cardiovascular Health Study
DUBBO	Dubbo Study of the Elderly
FINE_FIN	Finland, Italy and Netherlands Elderly Study - Finland cohort
GOH	The Glucose Intolerance Obesity and Hypertension Study
GOTO43	Göteborg 1943 Study
GOTOW	Population Study of Women in Gothenburg, Sweden
HELSINAG	Helsinki Aging Study
HOORN	Hoorn Study
KIHD	Kuopio Ischaemic Heart Disease Study
MALMO	Malmö Study
MATISS-83	Cohort of Progetto CUORE
MATISS-87	Cohort of Progetto CUORE
MATISS-93	Cohort of Progetto CUORE
MRFIT	Multiple Risk Factors Intervention Trial
NCS3	Cohort of The Cardiovascular Disease Study in Norwegian Counties
NHANES III	Third National Health and Nutrition Examination Survey
OSLO	Oslo Study
PARIS1	Paris Prospective Study I
PRHHP	Puerto Rico Heart Health Program
RANCHO	Rancho Bernardo Study
REYK	Reykjavik Study
RIFLE	Risk Factors and Life Expectancy Pooling Project
SHS	Strong Heart Study
TARFS	Turkish Adult Risk Factor Study
ULSAM	Uppsala Longitudinal Study of Adult Men
VHMPP	Vorarlberg Health Monitoring and Promotion Programme
VITA	Vicenza Thrombophilia and Atherosclerosis Project
WHITE II	Whitehall II Study
ZARAGOZA	Zaragoza study



# Appendix D

## Appendix to chapter 7

### D.1 Example measurement error plots for scenarios 2,4,5,8-10 using log-exposure

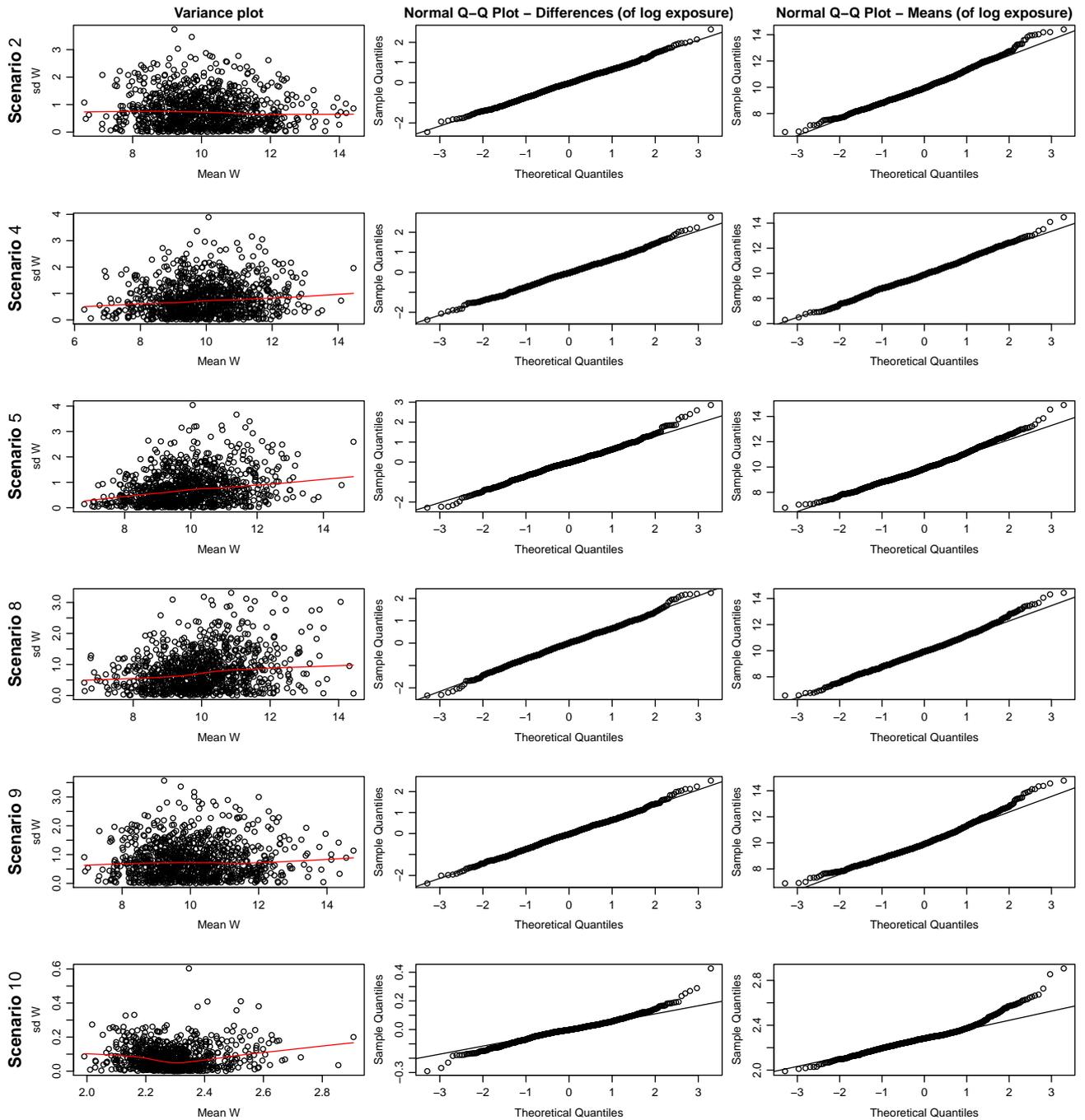


Figure D.1: Example measurement error plots under scenarios 2, 4, 5, 8-10,  $\sigma_u^2 = 1$  using log-exposure.



# Appendix E

## Appendix to chapter 9

### E.1 ERFC Lp(a) data — study names

AFTCAPS	Air Force/Texas Coronary Atherosclerosis Prevention Study
ARIC	Atherosclerosis Risk in Communities Study
ATTICA	Attica Study
BRHS	British Regional Heart Study
BRUN	Bruneck Study
BUPA	British Union Provident Association
CHARL	Charleston Heart Study
CHS	Cardiovascular Health Study
COPEN	Copenhagen City Heart Study
DUBBO	Dubbo Study of the Elderly
EAS	Edinburgh Artery Study
FIA	First Myocardial Infarction in Northern Sweden
FINRISK 92	Finrisk Cohort 1992
FLETCHER	Fletcher Challenge Blood Study
FRAMOFF	Framingham Offspring Cohort
GOH	The Glucose Intolerance, Obesity and Hypertension Study
GOTO33	Göteborg Study 1933
GRIPS	Göttingen Risk Incidence and Prevalence Study
HPFS	Health Professionals Follow-up Study
KIHD	Kuopio Ischaemic Heart Disease Study
MRFIT	Multiple Risk Factor Intervention Trial
NHANES III	National Health and Nutrition Examination Survey III
NHS	Nurses Health Study
NPHS II	Northwick Park Heart Study II
PRIME	Prospective Epidemiological Study of Myocardial Infarction
PROCAM	Prospective Cardiovascular Münster Study
QUEBEC	Quebec Cardiovascular Study
REYK	Reykjavik Study
SHS	Strong Heart Study
TARFS	Turkish Adult Risk Factor Study
ULSAM	Uppsala Longitudinal Study of Adult Men
USPHS	U.S. Physicians Health Study



# Appendix F

## R code

In this appendix we give R code to accompany many of the methods described in this dissertation.

### F.1 Code to accompany chapter 2

#### F.1.1 P-spline plotting function

`pspline.plot` will plot P-splines from Cox proportional hazards models.

`pspline.plot` takes the following arguments

<code>x</code>	Exposure to which the P-spline was fit
<code>coef</code>	Spline coefficients from a <code>coxph</code> fit
<code>v</code>	Variance-covariance matrix for <code>coef</code>
<code>add</code>	If <code>TRUE</code> then will plot curve on current graph
<code>se.plot</code>	If <code>TRUE</code> will plot standard errors
<code>df</code>	Degrees of freedom of spline model
<code>nterm</code>	The number of spline basis terms
<code>degree</code>	Degree of P-spline
<code>base</code>	Reference value, if <code>NULL</code> then the mean of <code>x</code> will be used as the reference value
<code>...</code>	Options to be passed on to plot functions

```
pspline.plot <- function(x, coef, v=NULL, add=FALSE, se.plot=FALSE, df=4,  
                          nterm=2.5*df, degree=3, base=NULL, ...){
```

```
  require(survival)
```

```
  #Recreate basis - uses code from pspline
```

```
  x <- sort(x)
```

```
  keepx <- !is.na(x)
```

## F. R code

---

```
base      <- ifelse(is.null(base), mean(x[keepx]), base)
rx        <- range(x[keepx])
dx        <- (rx[2] - rx[1])/nterm
knots     <- c(rx[1] +
              dx * ((-degree):(nterm - 1)), rx[2] + dx * (0:degree))
temp      <- spline.des(knots, x[keepx], degree+1)$design[,-1]

#Calculate log-HRs and subtract log-HR at reference
plot.vals <- temp%*%coef
pred.base <- spline.des(knots, base, degree+1)$design[,-1] %*%
            coef
plot.vals <- as.vector(plot.vals) - pred.base

#Calculate standard errors if required
if(se.plot){

base.vals <- spline.des(knots, base, degree+1)$design[,-1]
for(i in 1:ncol(temp))
temp[,i] <- temp[,i] - base.vals[i]
            se <- rep(NA, length(x[keepx]))
            for (i in (1:NROW(temp))) {
                se[i] <- sqrt(sum(temp[i,] %*% v %*% temp[i,]))
            }
}

#Plot spline
if(add==FALSE)
plot(x[keepx], exp(plot.vals), type="l" , ylab="Hazard Ratio"
     , xlab="Exposure level", log="y", ...)
else lines(x[keepx], exp(plot.vals), ...)
#Plot standard errors if required
if(se.plot){
lines(x[keepx], exp(plot.vals+1.96*se), col="red", lty="dashed")
lines(x[keepx], exp(plot.vals-1.96*se), col="red", lty="dashed")
}
}
```

**An example using the cancer dataset that is supplied with the survival package**

```
library(survival)
fit1 <- coxph(Surv(time, status) ~ ph.ecog + pspline(age), cancer)
pspline.plot(cancer$age, fit1$coef[-1], fit1$var[-1,-1], se.plot=TRUE)
```

---

## F.2 Code to accompany chapter 4

### F.2.1 Calculation of $E((x - k)_+^p)$ for structural P-splines

For the mathematical details and notation used we refer the reader back to appendix A.

First we create a function that calculates  $I_r$  for  $r = 0, \dots, p$ .

The function `Ir` takes the arguments:

- `p` The degree of the truncated power function
- `h` The standard score at the threshold  $k$

```
Ir <- function(p, h) {
  I.temp <- rep(NA, length(p))
  I.temp[1] <- 1
  I.temp[2] <- dnorm(h) / (1-pnorm(h))
  i <- 3
  while(i < (length(p)+1)) {
    I.temp[i] <- ((h^(i-2) * dnorm(h)) / (1-pnorm(h))) + (i-2) * I.temp[i-2]
    i <- i+1}
  I.temp}
```

The function `spline.trunc` which uses `Ir` calculates  $E((x - k)_+^p)$ .

`spline.trunc` takes the arguments

- `mu` The mean of the distribution of  $x$
- `sigma` The standard deviation of the distribution of  $x$
- `k` The threshold
- `p` The degree of the truncated power function

```
spline.trunc <- function(mu, sigma, k, p) {
  h <- (k-mu)/sigma
  r <- 0:p
  if((1-pnorm(h)) < .Machine$double.eps) 0
  else (1-pnorm(h)) * sum(choose(p, r) * mu^(p-r) * (sigma^r) * Ir(r, h))
}
```

Note that the fourth line catches the case that  $h$  is very large which can lead to (numerical) division by zero within `Ir`.

## F.3 Code to accompany chapter 5

### F.3.1 Multiple imputation with true exposure observed in a subset

`mi.mod` will perform multiple imputation, where we have a continuous exposure subject to classical measurement error but wish to perform our analysis on groups of this exposure; we have observed the true exposure for a subset of individuals; and where the exposure–outcome relationship is modelled using a generalised linear model. Note that we require the `mice` [182] package to be installed.

`mi.mod` takes the arguments:

<code>y</code>	Outcome vector
<code>w</code>	Vector containing the variable subject to measurement error
<code>int.val</code>	Vector containing the true exposure for each individual where this is observed and NA otherwise
<code>cutpoints</code>	A vector of cutpoints to be used to group the continuous exposure
<code>m</code>	The number of imputations, if NULL then defaults to percentage of missing data
<code>z</code>	A matrix or data frame of other covariates (assumed to be measured without error) to be included in the model
<code>family</code>	A description of the error distribution and link function to be used in the model. See <code>help</code> for <code>glm</code> for more details.

```
mi.mod <- function(y,w,int.val,cutpoints,m=NULL,z=NULL,family="gaussian"){  
  
  #Loads mice if not already loaded  
  require(mice)  
  
  #Sets the number of imputations to percentage of missing data if not  
  #specified  
  m <- ifelse(is.null(m),  
              ceiling((sum(is.na(int.val))/length(int.val))*100),m)  
  
  #Creates data frame of all variables  
  if(is.null(z)) mydata <- data.frame(y,int.val,w)  
  else mydata <- data.frame(y,int.val,w,z)  
  
  #Creates m multiply imputed datasets  
  imp.data <- mice(mydata,im="norm",m=m,printFlag=FALSE)  
  
  #Create formula for regression  
  fmla <- as.formula(ifelse(is.null(z),  
                            "y ~ cut(int.val,c(-Inf,cutpoints,Inf))",
```

---

```

        paste("y ~ cut(int.val, c(-Inf, cutpoints, Inf)) +",
              paste(colnames(z), collapse="+"))))

#Fit models using each of the imputed datasets
fit.imp <- with(data=imp.data, glm(fmla, family=family))

#Pool results using Rubin's rules
pool(fit.imp)
}

```

### Example:

```

#True exposure
x <- rnorm(1000)
#Mismeasured exposure
w <- x+rnorm(1000)
#Outcome
y <- 2+x+rnorm(1000, 0, sqrt(0.25))
#Validation substudy
int.val <- x
int.val[sample(1000,100)] <- NA
#Compare results from MI and using the true exposure
mi.mod(y, w, int.val, 2)
lm(y~(x>2))

```

## F.3.2 Moment reconstruction with true exposure observed in a subset

`mr.mod` will perform moment reconstruction, where we have a continuous exposure subject to classical measurement error but wish to perform our analysis on groups of this exposure; we have observed the true exposure for a subset of individuals; and where the exposure–outcome relationship is modelled using a generalised linear model.

`mr.mod` takes the arguments:

<code>y</code>	Outcome vector
<code>w</code>	Vector containing the variable subject to measurement error
<code>int.val</code>	Vector containing the true exposure measurement for each individual where this is observed and NA otherwise
<code>cutpoints</code>	A vector of cutpoints to be used to group the continuous exposure
<code>m</code>	The number of imputations, if NULL then defaults to percentage of missing data
<code>z</code>	A matrix or data frame of other covariates (assumed to be measured without error) to be included in the model
<code>family</code>	A description of the error distribution and link function to be used in the model. See help for <code>glm</code> for more details.

## F. R code

---

```
mr.mod <- function(y,w,int.val,cutpoints,m=NULL,z=NULL,family="gaussian"){

#Create formula for imputation model
fmla <- as.formula(iffelse(is.null(z),
  "int.val ~ y",
  paste("int.val ~ y +",paste(colnames(z),collapse="+"))))

fmla2 <- as.formula(iffelse(is.null(z),
  "w ~ y",
  paste("w ~ y +",paste(colnames(z),collapse="+"))))

#Create matrix of data
if(is.null(z)) mydata <- cbind(1,y)
else mydata <- cbind(1,y,z)

#Find conditional expectations and variances
mod.valgivyz <- lm(fmla)
mod.wgivyz <- lm(fmla2)
exp.x.given.yz <- mod.valgivyz$coef%*%t(mydata)
exp.w.given.yz <- mod.wgivyz$coef%*%t(mydata)
var.x.given.yz <- var(residuals(mod.valgivyz))
var.w1.given.yz <- var(residuals(mod.wgivyz))

g<-sqrt(var.x.given.yz/var.w1.given.yz)

#Find moment reconstruction values
mr.vals <- exp.x.given.yz+g*(w-exp.w.given.yz)

x.mr<- iffelse(!is.na(int.val),int.val,mr.vals)

#Fit disease model using moment reconstruction values
fmla3 <- as.formula(iffelse(is.null(z),
  "y ~ cut(int.val,c(-Inf,cutpoints,Inf))",
  paste("y ~ cut(int.val,c(-Inf,cutpoints,Inf))+",
  paste(colnames(z),collapse="+"))))
glm(fmla3, family=family)$coef
}
```

### Example:

```
#True exposure
x <- rnorm(1000)
#Mismeasured exposure
```

---

```

w          <- x+rnorm(1000)
#Outcome
y          <- 2+x+rnorm(1000,0,sqrt(0.25))
#Validation substudy
int.val <-x
int.val[sample(1000,100)] <- NA
#Compare results from MI and using the true exposure
mr.mod(y,w,int.val,2)
lm(y~(x>2))

```

### F.3.3 group-SIMEX

`gsimex.mod` will fit a group-SIMEX model, where we have a continuous exposure subject to classical measurement error but wish to perform our analysis on groups of this exposure; we have observed a replicate measure of exposure for a subset of individuals; and where the exposure–outcome relationship is modelled using a generalised linear model.

`gsimex.mod` takes the arguments:

<code>y</code>	Outcome vector
<code>w</code>	Vector containing the variable subject to measurement error
<code>meas.err</code>	Estimate of measurement error variance
<code>cutpoints</code>	Vector of cutpoints to be used to group the continuous exposure
<code>z</code>	Matrix or data frame of other covariates (assumed to be measured without error) to be included in the model
<code>family</code>	A description of the error distribution and link function to be used in the model. See <code>help for glm</code> for more details.
<code>lambda</code>	Vector of lambda values for which the simulation step should be performed
<code>B</code>	Number of pseudo datasets to be created for each value of lambda

```

gsimex.mod <- function(y,w,meas.err,cutpoints, z=NULL, family="gaussian",
                      lambda=c(0.5,1,1.5,2), B=100){

#Define lambda
lambda      <- c(0,lambda)

#Create formula for regression
fmla <- as.formula(iffelse(is.null(z),
                          "y ~ cut(wstar,c(-Inf,cutpoints,Inf))",
                          paste("y ~ cut(wstar,c(-Inf,cutpoints,Inf))+",
                                paste(colnames(z),collapse="+"))))

#Calculate number of coefficients

```

## F. R code

---

```
ncoef <- 1+length(cutpoints)+ifelse(is.null(z),0,ncol(z))

#Define coefficient vector and input value from naive model
wstar <- w
temp.mod <- glm(fmla, family=family)
coef.labs <- names(temp.mod$coef)
coef.vals <- matrix(NA,length(lambda),ncoef,
                    dimnames=list(lambda,coef.labs))
coef.var.vals <- matrix(NA,length(lambda),ncoef^2)
coef.vals[1,] <- temp.mod$coef
coef.var.vals[1,] <- as.vector(summary(temp.mod)$cov.unscaled)

#For each value of lambda create B replicate datasets and take average
#parameter and calculate covariance matrix
for(j in 2:length(lambda)){
temp <- matrix(NA,B,ncoef)
temp.var <- matrix(0,ncoef,
                  ncoef)

for(i in 1:B){
wstar <- w + sqrt(lambda[j])*sqrt(meas.err)*rnorm(length(w))
temp.mod <- glm(fmla, family=family)
temp[i,] <- temp.mod$coef
temp.var <- temp.var+summary(temp.mod)$cov.unscaled}

coef.vals[j,] <- apply(temp,2,mean)
coef.var.vals[j,] <- (1/100)*as.vector(temp.var)-as.vector(cov(temp))

cat("\n lambda=",lambda[j],"complete \n")
}

#Regress coefficient values on lambda
out.coef <- rep(NA,ncoef)
names(out.coef) <- coef.labs
for(i in 1:length(out.coef)){
temp.mod <- lm(coef.vals[,i] ~ lambda + I(lambda^2))
out.coef[i] <- temp.mod$coef[1] - temp.mod$coef[2] + temp.mod$coef[3]}

#Regress var values on lambda
out.var <- rep(NA,ncoef^2)
for(i in 1:length(out.var)){
temp.mod <- lm(coef.var.vals[,i] ~ lambda + I(lambda^2))
```

---

```

out.var[i] <- temp.mod$coef[1] - temp.mod$coef[2] + temp.mod$coef[3]
}
out.var      <- matrix(out.var,ncoef,ncoef,byrow=TRUE,
                      dimnames=list(coef.labs,coef.labs))

#Output
list(B=B, lambda=lambda,out.coef=out.coef, cutpoints=cutpoints,
      out.var=out.var, SIMEX.coef=coef.vals)
}

```

`gsimex.mod` returns:

<code>B</code>	Number of pseudo datasets to be created for each value of lambda
<code>lambda</code>	Vector of lambda values for which the simulation step was performed
<code>out.coef</code>	The group-SIMEX corrected parameter estimates
<code>out.var</code>	The jackknife estimated variance-covariance matrix for the group-SIMEX corrected parameter estimates
<code>cutpoints</code>	Vector of cutpoints that were used to group the continuous exposure
<code>SIMEX.coef</code>	A matrix of parameter values for each value of lambda

### Example

```

#Generate true exposure, mismeasured exposure and outcome
x <- rnorm(1000)
w <- x+rnorm(1000)
y <- 2 + x + rnorm(1000,0,sqrt(0.25))
#Fit true model with cutpoint c=0
lm(y~I(x>0))
#Fit group-simex model to observed data
gsimex.mod(y,w,meas.err,cutpoints=0)

```

## F.3.4 Plots of SIMEX extrapolation curves

`simexgraph` will produce plots of the quadratic extrapolation function for both `gsimex.mod` as well as for models fitted using the `simex` and `mcsimex` functions with the `simex` package.

`simexgraph` takes the arguments:

<code>mod</code>	Object fitted using <code>gsimex.mod</code> described above or <code>simex</code> and <code>mcsimex</code> functions from the <code>simex</code> package
<code>variables</code>	A vector of variable names or coefficient numbers to be plotted. If <code>NULL</code> all variables are plotted.
<code>title</code>	A vector of titles. If <code>NULL</code> titles are generated.

```

simexgraph <- function(mod, variables=NULL,title=NULL){

```

## F. R code

---

```
#Select all variables if none are chosen
if(is.null(variables))
variables <- colnames(mod$SIMEX.estimate)[-1]

#Create plot window
par(mfrow=c(ceiling(length(variables)/2),2))

for(i in variables){

#Create blank plot
plot(mod$SIMEX.estimate[,1],mod$SIMEX.estimate[,i],
xlab=expression(lambda), ylab=expression(beta(lambda)), type="n")

#Create title or use user defined
if(!is.null(title)) title(title[i])
else(title(paste("Extrapolation plot for variable",i)))

#Recreate quadratic extrapolation
bob <- lm(mod$SIMEX.estimate[-1,i]~mod$lambda+I(mod$lambda^2))

#Plot points and curves
curve(coefficients(bob)[1]+coefficients(bob)[2]*x+
coefficients(bob)[3]*x^2,
add=TRUE,lwd=1,from=0, to=2, col="blue")
curve(coefficients(bob)[1]+coefficients(bob)[2]*x+
coefficients(bob)[3]*x^2,
add=TRUE,lwd=1,from=-1, to=0, col="blue", lty=2)
points(mod$lambda, mod$SIMEX.estimate[-1,i],
pch=rep(16,length(mod$lambda)), cex=1.5)
points(-1,mod$coefficients[i], cex=1.5)
}

}
```

### Example

```
library(simex)
#This is the example from the SIMEX package
x <- rnorm(200, 0, 100)
u <- rnorm(200, 0, 25)
w <- x + u
y <- x + rnorm(200, 0, 9)
```

---

```

true.model <- lm(y ~ x)
naive.model <- lm(y ~ w, x = TRUE)
simex.model <- simex(model = naive.model, SIMEXvariable = "w",
  measurement.error = 25)
#Plot SIMEX graph
simexgraph(simex.model)

```

## F.4 Code to accompany chapter 8

### F.4.1 DerSimonian and Laird based multivariate meta-analysis

`multi.meta.dl` will perform a fixed- or random-effects multivariate meta-analysis using a DerSimonian and Laird based approach.

`multi.meta.dl` takes the arguments:

TE            Matrix with  $n$  rows and  $n_p$  columns  
varTE        List of  $n$   $n_p \times n_p$  variance-covariance matrices  
subset        Vector of which studies are to be included  
fixed        Whether a fixed effects analysis is to be performed (random effects is default)

```

multi.meta.dl <- function (TE, varTE, subset=NULL, fixed=FALSE){

#If subset then select data
if(!is.null(subset)){
TE <- TE[subset,]
temp <- list(varTE[[subset[1]]])
for(i in 2:length(subset))
temp[[i]] <- varTE[[subset[i]]]
varTE <- temp}

TE            <- as.matrix(TE)

#No. of outcomes
no            <- ncol(TE)

#No. of studies
n            <- nrow(TE)

#Fill in missing values - weight out using large variance
TE[is.na(TE)] <- 0
for(i in 1:n)
diag(varTE[[i]])[diag(varTE[[i]])==0] <- 1e12

```

## F. R code

---

```
#Calculate standard errors
seTE      <- t(sapply(varTE, function(x) sqrt(unlist(diag(x)))))
#Calculate correlation matrices
corTE     <- lapply(varTE, function(x) cov2cor(matrix(unlist(x),no,no)))

#Initialise variables
Sigma     <- Sigma_trun <- Q <- matrix(0,no,no)

#If fixed then set Sigma to zero
if(fixed)
Sigma_trun <- matrix(0,no,no)
else{
for(i in 1:no){
for(j in 1:no){
#Do pairwise (by outcome) calculations
#Select outcomes
X      <- TE[,i]
Y      <- TE[,j]

#Calculate weights
W      <- 1/(seTE[,i]*seTE[,j])

#Calculate weighted means
X.bar  <- weighted.mean(X,W)
Y.bar  <- weighted.mean(Y,W)

#Calculate Q matrix value
Q[i,j] <- sum( W*((X-X.bar)*(Y-Y.bar)))

temp.cor <- rep(NA,n)
for(k in 1:n)
temp.cor[k] <- as.numeric(corTE[[k]][i,j])
#Calculate coefficients in equation
a      <- sum(temp.cor) - weighted.mean(temp.cor, W)
b      <- sum(W) - weighted.mean(W, W)

#Calculate Sigma matrix value
Sigma[i,j] <- (Q[i,j]-a)/b
}}

#Perform truncation
```

---

```

eig <- eigen(Sigma)
for(i in 1:no)
Sigma_trun<-Sigma_trun+max(0, eig$values[i])*
eig$vectors[,i]%*%t(eig$vectors[,i])
}

#Calculate components of treatment effect
temp <- matrix(0,no,no)
temp2 <- rep(0,no)
for(i in 1:n){
temp.inv <- solve(Sigma_trun+matrix(unlist(varTE[[i]]),no,no))
temp <- temp + temp.inv
temp2 <- temp2 + temp.inv %*% TE[i,]
}

#Calculate overall TE and variance
TE.overall.var <- solve(temp)
TE.overall <- TE.overall.var%*%(temp2)

#Name rows and columns
colnames(Sigma) <- rownames(Sigma) <- colnames(Sigma_trun)
<- rownames(Sigma_trun) <- colnames(Q) <- rownames(Q) <- colnames(TE)

#Output list of results
if(fixed)
list(TE.overall=TE.overall,TE.var=TE.overall.var,subset=subset)
else list(TE.overall=TE.overall,Sigma_trun=Sigma_trun,Sigma=Sigma,
Q=Q,TE.var=TE.overall.var,
eigen=eig, subset=subset)
}

```

`multi.meta.dl` returns

<code>TE.overall</code>	The overall treatment effect
<code>TE.var</code>	The variance-covariance matrix for the overall treatment effect
<code>subset</code>	Vector of which studies were included (NULL if all were included)

and additionally for random effects analyses

<code>Sigma</code>	The estimated between study variance-covariance matrix
<code>Sigma_trun</code>	The estimated truncated between study variance-covariance matrix
<code>eigen</code>	The eigenvalues of Sigma
<code>Q</code>	The $Q$ matrix

**F.4.2 REML based multivariate meta-analysis**

`multi.meta.dl` will perform a fixed- or random-effects multivariate meta-analysis using a DerSimonian and Laird based approach. Note that this function requires that the package `meta` is installed as it is used to obtain starting values for maximising the restricted maximum likelihood.

`multi.meta.reml` takes the arguments:

`TE`            Matrix with  $n$  rows and  $n_p$  columns  
`varTE`        List of  $n$ ,  $n_p \times n_p$  variance-covariance matrices  
`subset`       Vector of which studies are to be included

```
multi.meta.reml <- function(TE, varTE, subset=NULL) {

#If subset then select data
if(!is.null(subset)){
TE <- TE[subset,]
temp <- list(varTE[[subset[1]]])
for(i in 2:length(subset))
temp[[i]] <- varTE[[subset[i]]]
varTE <- temp}

# Number of outcomes
m <- ncol(TE)

#Number of studies
k <- nrow(TE)

#Perform univariate meta-analysis to obtain start values for ML
uni.meta <- rep(0,2*m)

require(meta)

mu <- sig <- rep(0,m)
for(i in 1:m){
temp.se <- rep(0,k)
for(j in 1:k) temp.se[j] <- sqrt(unlist(varTE[[j]][i,i]))
mu[i] <- metagen(TE[,i], temp.se)$TE.random
sig[i] <- metagen(TE[,i], temp.se)$seTE.random
}

#Create function that calculates the log-likelihood
```

---

```

ll <- function(musig){

mu    <- musig[1:m]
Sigma <- matrix(0,m,m)
diag(Sigma) <- musig[(m+1):(2*m)]^2
Sigma[upper.tri(Sigma)] <- musig[((2*m)+1):(m+0.5*m*(m+1))]
Sigma[lower.tri(Sigma)] <- t(Sigma)[lower.tri(t(Sigma))]

#Make sure that Sigma is positive semi-definite
Sigma_trun <- matrix(0,m,m)
eig <- eigen(Sigma)
for(i in 1:m) Sigma_trun<-Sigma_trun+
                max(0, eig$values[i])*eig$vectors[,i]%*%t(eig$vectors[,i])
Sigma <- Sigma_trun

#Calculate the three components of the likelihood
term1 <- 0
for(i in 1:k)
term1 <- term1 + log(det(Sigma+matrix(unlist(varTE[[i]]),m,m,byrow=TRUE)))

temp <- matrix(0,m,m)
for(j in 1:k)
temp <- temp +
        solve(Sigma+matrix(unlist(matrix(unlist(varTE[[j]]),
                m,m,byrow=TRUE)),m,m,byrow=TRUE))
term2 <- log(det(temp))

term3 <- 0
for(l in 1:k)
term3 <- term3 + t(as.numeric(TE[l,])-mu)%*%solve(
                Sigma+matrix(unlist(varTE[[l]]),
                m,m,byrow=TRUE))%*%(as.numeric(TE[l,])-mu)

#Want to minimize -loglik
reml <- term1+term2 +term3
}

#Maximise the restricted likelihood
max.vals <- optim( c(mu,c(sig, rep(0,0.5*(m-1)*m))),ll, method="BFGS")

mu <- max.vals$par[1:m]
Sigma <- matrix(0,m,m)

```

## F. R code

---

```
diag(Sigma) <- max.vals$par[(m+1):(2*m)]^2
Sigma[upper.tri(Sigma)] <- max.vals$par[((2*m)+1):(m+0.5*m*(m+1))]
Sigma[lower.tri(Sigma)] <- t(Sigma)[lower.tri(t(Sigma))]

#Calculate the truncated matrix which is positive semi-definite
Sigma_trun <- matrix(0,m,m)
eig <- eigen(Sigma)
for(i in 1:m) Sigma_trun<-Sigma_trun+max(0, eig$values[i])*
               eig$vectors[,i]%*%t(eig$vectors[,i])

list(TE.overall=mu, Sigma_trun=Sigma_trun, Sigma=Sigma, eigen=eig, subset=subset)
}
```

`multi.meta.reml` returns

<code>TE.overall</code>	<b>The overall treatment effect</b>
<code>TE.var</code>	<b>The variance-covariance matrix for the overall treatment effect</b>
<code>subset</code>	<b>Vector of which studies were included (NULL if all were included)</b>

# References

- [1] R.H. Keogh, A.D. Strawbridge, and I.R. White. Effects of classical exposure measurement error on the shape of exposure-disease associations. *Epidemiologic Methods*, in press, 2012.
- [2] R.H. Keogh, A.D. Strawbridge, and I.R. White. Correcting for bias due to misclassification when error-prone continuous exposures are categorized. *Epidemiologic Methods*, in press, 2012.
- [3] R.O. Bonow, L.A. Smaha, S.C. Smith, G.A. Mensah, and C. Lenfant. World heart day 2002. *Circulation*, 106(13):1602–1605, 2002.
- [4] D. Lloyd-Jones, R.J. Adams, T.M. Brown, M. Carnethon, S. Dai, G. De Simone, T.B. Ferguson, E. Ford, K. Furie, C. Gillespie, A. Go, K. Greenlund, N. Haase, S. Hailpern, P.M. Ho, V. Howard, B. Kissela, S. Kittner, D. Lackland, L. Lisabeth, A. Marelli, Mary M. McDermott, J. Meigs, D. Mozaffarian, M. Mussolino, G. Nichol, V.L. Roger, W. Rosamond, R. Sacco, P. Sorlie, R. Stafford, T. Thom, S. Wasserthiel-Smoller, N.D. Wong, and J. Wylie-Rosett on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics - 2010 update. *Circulation*, 121(7):e46–e215, 2010.
- [5] S. Capewell, S. Allender, J. Critchley, F. Lloyd-Williams, M. O’Flaherty, M. Rayner, and P. Scarborough. Modelling the UK burden of cardiovascular disease to 2020. *Cardio & Vascular Coalition and the British Heart Foundation*, 2011.
- [6] P.A. Heidenreich, J.G. Trogon, O.A. Khavjou, J. Butler, K. Dracup, M.D. Ezekowitz, E.A. Finkelstein, Y. Hong, S.C. Johnston, A. Khera, D.M. Lloyd-Jones, S.A. Nelson, G. Nichol, D. Orenstein, P.W.F. Wilson, and Y.J. Woo. Forecasting the future of cardiovascular disease in the United States. *Circulation*, 123(8):933–944, 2011.
- [7] J. Danesh, S. Erqou, M. Walker, S.G. Thompson, R. Tipping, C. Ford, S. Pressel, G. Walldius, I. Jungner, A.R. Folsom, et al. The Emerging Risk Factors Collaboration: analysis of individual data on lipid, inflammatory and other markers in over 1.1 million participants in 104 prospective studies of cardiovascular diseases. *European Journal of Epidemiology*, 22(12):839–869, 2007.

## References

---

- [8] S. Kaptoge, E. Di Angelantonio, G. Lowe, M.B. Pepys, S.G. Thompson, R. Collins, and J. Danesh. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *Lancet*, 375(9709):132–140, 2010.
- [9] The Emerging Risk Factors Collaboration. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet*, 375(9733):2215–2222, 2010.
- [10] The Emerging Risk Factors Collaboration. Diabetes mellitus, fasting glucose, and risk of cause-specific death. *The New England Journal of Medicine*, 364:829–41, 2011.
- [11] The Emerging Risk Factors Collaboration. Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. *JAMA: The Journal of the American Medical Association*, 302(4):412–423, 2009.
- [12] E. Di Angelantonio, N. Sarwar, P. Perry, S. Kaptoge, K.K. Ray, A. Thompson, A.M. Wood, S. Lewington, N. Sattar, C.J. Packard, et al. Major lipids, apolipoproteins, and risk of vascular disease. *JAMA: The Journal of the American Medical Association*, 302(18):1993–2000, 2009.
- [13] M. Woodward, F. Barzi, A. Martiniuk, X. Fang, D.F. Gu, Y. Imai, T.H. Lam, W.H. Pan, A. Rodgers, I. Suh, S.H. Jee, H. Ueshima, and R. Huxley. Cohort profile: the Asia Pacific Cohort Studies Collaboration. *International Journal of Epidemiology*, 35(6):1412–1416, 2006.
- [14] E.E. Calle, C.W. Heath, H.L. MiracleMcMahill, R.J. Coates, and P.A. Van Den Brandt. Breast cancer and hormonal contraceptives: Collaborative reanalysis of individual data on 53,297 women with breast cancer and 100,239 women without breast cancer from 54 epidemiological studies. *Lancet*, 347(9017):1713–27, 1996.
- [15] S. Thompson, S. Kaptoge, I. White, A. Wood, P. Perry, and J. Danesh. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *International Journal of Epidemiology*, 39(5):1345–1359, 2010.
- [16] V. Curtis and S. Cairncross. Water, sanitation, and hygiene at Kyoto. *BMJ*, 327(7405):3–4, 2003.
- [17] J. Snow. *On the mode of communication of cholera*. John Churchill, 1855.
- [18] R. Doll and A.B. Hill. Smoking and carcinoma of the lung. *BMJ*, 2(4682):739–748, 1950.
- [19] M.L. Levin, H. Goldstein, and P.R. Gerhardt. Cancer and tobacco smoking. *JAMA: The Journal of the American Medical Association*, 143(4):336–338, 1950.

- [20] E.L. Wynder and E.A. Graham. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma. *JAMA: The Journal of the American Medical Association*, 143(4):329–336, 1950.
- [21] T.R. Dawber, G.F. Meadors, and F.E. Moore Jr. Epidemiological approaches to heart disease: the Framingham study. *American Journal of Public Health*, 41(3):279–286, 1951.
- [22] T.R. Dawber, W.B. Kannel, N. Revotskie, J. Stokes III, A. Kagan, and T. Gordon. Some factors associated with the development of coronary heart disease—six years' follow-up experience in the Framingham study. *American Journal of Public Health*, 49(10):1349–1356, 1959.
- [23] W.B. Kannel, T.R. Dawber, A. Kagan, N. Revotskie, and J. Stokes. Factors of risk in the development of coronary heart disease - six-year follow-up experience. *Annals of Internal Medicine*, 55(1):33–50, 1961.
- [24] R.F. Heller, S. Chinn, H.D. Pedoe, and G. Rose. How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. *BMJ*, 288(6428):1409–1411, 1984.
- [25] G. Rose. *The strategy of preventive medicine*. Oxford University Press, 1992.
- [26] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–220, 1972.
- [27] G.L. Myers, W.G. Miller, J. Coresh, J. Fleming, N. Greenberg, T. Greene, T. Hostetter, A.S. Levey, M. Panteghini, M. Welch, and J.H. Eckfeldt for the National Kidney Disease Education Program Laboratory Working Group. Recommendations for improving serum creatinine measurement: A report from the laboratory working group of the National Kidney Disease Education Program. *Clinical Chemistry*, 52(1):5–18, 2006.
- [28] B. Rosner, D. Spiegelman, and W.C. Willett. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *American Journal of Epidemiology*, 136(11):1400–1413, 1992.
- [29] I.M. Heid, H. Küchenhoff, J. Miles, L. Kreienbrock, and H.E. Wichmann. Two dimensions of measurement error: Classical and Berkson error in residential radon exposure assessment. *Journal of Exposure Analysis and Environmental Epidemiology*, 14(5):365–377, 2004.
- [30] S.L. Zeger, D. Thomas, F. Dominici, J.M. Samet, J. Schwartz, D. Dockery, and A. Cohen. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environmental Health Perspectives*, 108(5):419–426, 2000.
- [31] V. Kipnis, L.S. Freedman, C.C. Brown, A.M. Hartman, A. Schatzkin, and S. Wacholder.

## References

---

- Effect of measurement error on energy-adjustment models in nutritional epidemiology. *American Journal of Epidemiology*, 146:842–855, 1997.
- [32] I. R. White. The level of alcohol consumption at which all-cause mortality is least. *Journal of Clinical Epidemiology*, 52(10):967–975, 1999.
- [33] R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall, 2nd edition, 2006.
- [34] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904.
- [35] S. MacMahon, R. Peto, J. Cutler, R. Collins, P. Sorlie, J. Neaton, R. Abbott, J. Godwin, A. Dyer, and J. Stamler. Blood pressure, stroke, and coronary heart disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet*, 335(8692):765–774, 1990.
- [36] Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet*, 360(9349):1903–1913, 2002.
- [37] H.C. Boshuizen, M. Lanti, A. Menotti, J. Moschandreas, H. Tolonen, A. Nissinen, S. Nedeljkovic, A. Kafatos, and D. Kromhout. Effects of past and recent blood pressure and cholesterol level on coronary heart disease and stroke mortality, accounting for measurement error. *American Journal of Epidemiology*, 165(4):398–409, 2007.
- [38] C. Iribarren, D. Sharp, C.M. Burchfiel, P. Sun, and J.H. Dwyer. Association of serum total cholesterol with coronary disease and all-cause mortality: Multivariate correction for bias due to measurement error. *American Journal of Epidemiology*, 143(5):463–471, 1996.
- [39] J.R. Emberson, P.H. Whincup, R.W. Morris, and M. Walker. Re-assessing the contribution of serum total cholesterol, blood pressure and cigarette smoking to the aetiology of coronary heart disease: impact of regression dilution bias. *European Heart Journal*, 24(19):1719–1726, 2003.
- [40] Asia Pacific Cohort Studies Collaboration. Blood Glucose and Risk of Cardiovascular Disease in the Asia Pacific Region. *Diabetes Care*, 27(12):2836–2842, 2004.
- [41] J.R. Emberson, P.H. Whincup, R.W. Morris, M. Walker, G.D.O. Lowe, and A. Rumley. Extent of regression dilution for established and novel coronary risk factors: results from the British Regional Heart Study. *European Journal of Cardiovascular Prevention & Rehabilitation*, 11(2):125–134, 2004.
- [42] R.L. Prentice. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342, 1982.

- [43] P. Knekt, J. Ritz, M.A. Pereira, E.J. O'Reilly, K. Augustsson, G.E. Fraser, U. Goldbourt, B.L. Heitmann, G. Hallmans, S. Liu, P. Pietinen, D. Spiegelman, J. Stevens, J. Virtamo, W.C. Willett, E.B. Rimm, and A. Ascherio. Antioxidant vitamins and coronary heart disease risk: a pooled analysis of 9 cohorts. *The American Journal of Clinical Nutrition*, 80(6):1508–1520, 2004.
- [44] P. Koh-Banerjee, M. Franz, L. Sampson, S. Liu, D.R. Jacobs, D. Spiegelman, W. Willett, and E. Rimm. Changes in whole-grain, bran, and cereal fiber consumption in relation to 8-y weight gain among men. *The American Journal of Clinical Nutrition*, 80(5):1237–1245, 2004.
- [45] G. Davey Smith and A.N. Phillips. Inflation in epidemiology: “The proof and measurement of association between two things” revisited. *BMJ*, 312:1659–1661, 1996.
- [46] P. Elliott, J. Stamler, R. Nichols, A.R. Dyer, R. Stamler, H. Kesteloot, and M. Marmot. Intersalt revisited: further analyses of 24 hour sodium excretion and blood pressure within and across populations. *BMJ*, 312(7041):1249–1253, 1996.
- [47] A.R. Dyer, P. Elliott, M. Marmot, H. Kesteloot, R. Stamler, and J. Stamler. Commentary: Strength and importance of the relation of dietary salt to blood pressure. *BMJ*, 312(7047):1661–1664, 1996.
- [48] N.E. Day. Intersalt data. Epidemiological studies should be designed to reduce correction needed for measurement error to a minimum. *BMJ*, 315(7106):484, 1997.
- [49] G. Davey Smith and A.N. Phillips. Correction for regression dilution bias in Intersalt study was misleading. *BMJ*, 315:485–486, 1997.
- [50] W.A. Fuller. *Measurement Error Models*. John Wiley, New York, 1987.
- [51] P. Gustafson. *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman and Hall, 2003.
- [52] J.P. Buonaccorsi. *Measurement error: models, methods, and applications*. Chapman & Hall/CRC, 2009.
- [53] Y. Guo and R.J. Little. Regression analysis with covariates that have heteroscedastic measurement error. *Statistics in Medicine*, 30(18):2278–2294, 2011.
- [54] D. Spiegelman, R. Logan, and D. Grove. Regression calibration with heteroscedastic error variance. *The International Journal of Biostatistics*, 7(1):4, 2011.
- [55] L. Thomas, L. Stefanski, and M. Davidian. A moment-adjusted imputation method for measurement error models. *Biometrics*, in press, 2011.
- [56] G.V. Glass. *Fertilizers, pills, and magnetic strips: The fate of public education in America*. Information Age Publishing Inc, 2008.

## References

---

- [57] R.J.S. Simpson and K. Pearson. Report on certain enteric fever inoculation statistics. *BMJ*, 2(2288):1243–1246, 1904.
- [58] B.J. Guzzetti, T.E. Snyder, G.V. Glass, and W.S. Gamas. Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Reading Research Quarterly*, 28(2):117–159, 1993.
- [59] M.R. Barrick and M.K. Mount. The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44:1–26, 1991.
- [60] J.M. Phillips and E.P. Goss. The effect of state and local taxes on economic development: A meta-analysis. *Southern Economic Journal*, 62(2):320–333, 1995.
- [61] L.A. Stewart and M.K. Parmar. Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet*, 341(8842):418–22, 1993.
- [62] R.D. Riley, P.C. Lambert, and G. Abo-Zaid. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*, 340(7745):521–525, 2010.
- [63] H.J. Eysenck. An exercise in mega-silliness. *American Psychologist*, 33(5):517, 1978.
- [64] M.L. Smith and G.V. Glass. Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32(9):752–760, 1977.
- [65] M. Egger, M. Schneider, and G.D. Smith. Spurious precision? Meta-analysis of observational studies. *BMJ*, 316(7125):140–144, 1998.
- [66] M. Blettner, W. Sauerbrei, B. Schlehofer, T. Scheuchenpflug, and C. Friedenreich. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology*, 28(1):1–9, 1999.
- [67] W. Sauerbrei and P. Royston. A new strategy for meta-analysis of continuous covariates in observational studies. *Statistics in Medicine*, 2011.
- [68] E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.
- [69] M. Greenwood. A report on the natural duration of cancer. Technical report, HMSO, 1926.
- [70] T.P. Ryan and W.H. Woodall. The most-cited statistical papers. *Journal of Applied Statistics*, 32(5):461–474, 2005.
- [71] D.R. Cox and D. Oakes. *Analysis of survival data*. Chapman and Hall, London, 1984.
- [72] T. Therneau and original R port by T. Lumley. *survival: Survival analysis, including penalised likelihood*, 2009. R package version 2.35-8.

- [73] R. Peto. Discussion on professor Cox's paper—regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):205–207, 1972.
- [74] N.E. Breslow. Discussion on professor Cox's paper—regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):216–217, 1972.
- [75] D.G. Clayton. A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*, 47(2):467–485, 1991.
- [76] D.J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, 2000.
- [77] D. Schoenfeld. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67(1):145–153, 1980.
- [78] P.M. Grambsch and T.M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.
- [79] T. Therneau and P. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000.
- [80] T. Therneau, P. Grambsch, and T. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.
- [81] O.O. Aalen. A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8):907–925, 1989.
- [82] E. Turner, J. Dobson, and S. Pocock. Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. *Epidemiologic Perspectives & Innovations*, 7(1):9, 2010.
- [83] D. Firth and R.X. De Menezes. Quasi-variances. *Biometrika*, 91(1):65–80, 2004.
- [84] D.F. Easton, J. Peto, and A. Babiker. Floating absolute risk: an alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. *Statistics in Medicine*, 10(7):1025–1035, 1991.
- [85] S. Greenland, K.B. Michels, J.M. Robins, C. Poole, and W.C. Willett. Presenting statistical uncertainty in trends and dose-response relations. *American Journal of Epidemiology*, 149(12):1077–1086, 1999.
- [86] M. Plummer. Improved estimates of floating absolute risk. *Statistics in Medicine*, 23(1):93–104, 2004.
- [87] M.S. Ridout. Summarizing the results of fitting generalized linear models to data from designed experiments. In R. Gilchrist A. Decarli, B. Francis and G. Seeber, editors,

## References

---

- Statistical Modelling: Proceedings of GLIM89 and the 4th International Workshop on Statistical Modelling*, pages 262–269. Springer-Verlag, 1989.
- [88] J.W. Tukey. On the comparative anatomy of transformations. *The Annals of Mathematical Statistics*, 28(3):602–632, 1957.
- [89] R.A. Durazo-Arvizu, D.L. McGee, R.S. Cooper, Y. Liao, and A. Luke. Mortality and optimal body mass index in a sample of the US population. *American Journal of Epidemiology*, 147(8):739–749, 1998.
- [90] G.E.P. Box and P.W. Tidwell. Transformation of the independent variables. *Technometrics*, 4(4):531–550, 1962.
- [91] P. Royston and D.G. Altman. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Journal of the Royal Statistical Society. Series C*, 43:429–467, 1994.
- [92] SAS Institute Inc. *SAS/STAT 9.1 User's Guide*. SAS Institute Inc., Cary, NC, 2004.
- [93] StataCorp. *Stata Statistical Software: Release 10*. Stata Corporation, College Station, TX, 2007.
- [94] W. Sauerbrei, C. Meier-Hirmer, A. Benner, and P. Royston. Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. *Computational statistics & data analysis*, 50(12):3464–3485, 2006.
- [95] S. Abbas, J. Linseisen, T. Slinger, S. Kropp, E.J. Mutschelknauss, D. Flesch-Janys, and J. Chang-Claude. Serum 25-hydroxyvitamin D and risk of post-menopausal breast cancer — results of a large case–control study. *Carcinogenesis*, 29(1):93–99, 2008.
- [96] G. Ambler and P. Royston. Fractional polynomial model selection procedures: investigation of type I error rate. *Journal of Statistical Simulation and Computation*, 69:89–108, 2001.
- [97] P. Royston and W. Sauerbrei. Improving the robustness of fractional polynomial models by preliminary covariate transformation: A pragmatic approach. *Computational Statistics and Data Analysis*, 51(9):4240–4253, 2007.
- [98] C. De Boor. *A Practical Guide to Splines*. Springer, 1978.
- [99] P. Eilers and B. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.
- [100] H. Heinzl and A. Kaider. Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Computer Methods and Programs in Biomedicine*, 54(3):201–208, 1997.

- 
- [101] L. Desquilbet and F. Mariotti. Dose-response analyses using restricted cubic spline functions in public health research. *Statistics in Medicine*, 29(9):1037–1057, 2010.
- [102] E. Suli and D. Mayers. *An Introduction to Numerical Analysis*. Cambridge University Press, 2003.
- [103] C. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183, 1967.
- [104] E.A. Eisen, I. Agalliu, S.W. Thurston, B.A. Coull, and H. Checkoway. Smoothing in occupational cohort studies: an illustration based on penalised splines. *Occupational and Environmental Medicine*, 61(10):854–860, 2004.
- [105] H.L. Hillege, V. Fidler, G.F.H. Diercks, W.H. van Gilst, D. de Zeeuw, D.J. van Veldhuisen, R.O.B. Gans, W.M.T. Janssen, D.E. Grobbee, and P.E. de Jong for the Prevention of Renal and Vascular End Stage Disease (PREVEND) Study Group. Urinary albumin excretion predicts cardiovascular and noncardiovascular mortality in general population. *Circulation*, 106(14):1777–1782, 2002.
- [106] R.J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951, 1992.
- [107] E.J. Malloy, D. Spiegelman, and E.A. Eisen. Comparing measures of model selection for penalized splines in Cox models. *Computational Statistics & Data Analysis*, 53(7):2605–2616, 2009.
- [108] C.M. Hurvich, J.S. Simonoff, and C. Tsai. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B*, 60(2):271–293, 1998.
- [109] U.S. Govindarajulu, D. Spiegelman, S.W. Thurston, B. Ganguli, and E.A. Eisen. Comparing smoothing techniques in Cox models for exposure–response relationships. *Statistics in Medicine*, 26(20):3735–3752, 2007.
- [110] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.
- [111] D. Ruppert. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757, 2002.
- [112] J.H. Friedman and B.W. Silverman. Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1):3–39, 1989.
- [113] L.A. Sleeper and D.P. Harrington. Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, 85(412):941–949, 1990.

## References

---

- [114] H. Binder and W. Sauerbrei. Adding local components to global functions for continuous covariates in multivariable regression modeling. *Statistics in Medicine*, 29(7-8):808–817, 2010.
- [115] P. Royston and W. Sauerbrei. *Multivariable Model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley, Chichester, 2008.
- [116] P. Royston. A useful monotonic non-linear model with applications in medicine and epidemiology. *Statistics in Medicine*, 19(15):2053–2066, 2000.
- [117] S. Greenland. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology*, 6(4):450–454, 1995.
- [118] S. Greenland. Problems in the average-risk interpretation of categorical dose-response analyses. *Epidemiology*, 6(5):563–565, 1995.
- [119] S. Greenland. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology*, 6(4):356–365, 1995.
- [120] C.R. Weinberg. How bad is categorization? *Epidemiology*, 6(4):345–347, 1995.
- [121] P. Royston, G. Ambler, and W. Sauerbrei. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, 28(5):964–974, 1999.
- [122] C. Faes, M. Aerts, H. Geys, and G. Molenberghs. Model averaging using fractional polynomials to estimate a safe level of exposure. *Risk Analysis*, 27(1):111–123, 2007.
- [123] U.S. Govindarajulu, E.J. Malloy, B. Ganguli, D. Spiegelman, and E.A. Eisen. The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with Cox models in a simulation study. *The International Journal of Biostatistics*, 5(1):2, 2009.
- [124] P. Royston. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. *Statistics in Medicine*, 19(14):1831–1847, 2000.
- [125] E.A. Zang and E.L. Wynder. Reevaluation of the confounding effect of cigarette smoking on the relationship between alcohol use and lung cancer risk, with larynx cancer used as a positive control. *Preventive medicine*, 32(4):359–370, 2001.
- [126] S.S. Franklin, S.A. Khan, N.D. Wong, M.G. Larson, and D. Levy. Is pulse pressure useful in predicting risk for coronary heart disease?: The Framingham Heart Study. *Circulation*, 100(4):354–360, 1999.
- [127] S. Wacholder, B. Armstrong, and P. Hartge. Validation studies using an alloyed gold standard. *American Journal of Epidemiology*, 137(11):1251–1258, 1993.

- [128] J. Berkson. Are there two regressions? *Journal of the American Statistical Society*, 45(250):164–180, 1950.
- [129] B. G. Armstrong. The effects of measurement errors on relative risk regressions. *American Journal of Epidemiology*, 132(6):1176–1184, 1990.
- [130] M. Goldberg, H. Kromhout, P. Gunel, A.C. Fletcher, M. Grin, D.C. Glass, D. Heederik, T. Kauppinen, and A. Ponti. Job exposure matrices in industry. *International Journal of Epidemiology*, 22(Supplement 2):S10–S15, 1993.
- [131] H. Küchenhoff, R. Bender, and I. Langner. Effect of Berkson measurement error on parameter estimates in Cox regression models. *Lifetime Data Analysis*, 13(2):261–272, 2007.
- [132] I.M. Heid, H. Küchenhoff, J. Wellmann, M. Gerken, L. Kreienbrock, and H.E. Wichmann. On the potential of measurement error to induce differential bias on odds ratio estimates: an example from radon epidemiology. *Statistics in Medicine*, 21(21):3261–3278, 2002.
- [133] S.J. Iturria, R.J. Carroll, and D. Firth. Polynomial regression and estimating functions in the presence of multiplicative measurement error. *Journal of the Royal Statistical Society: Series B*, 61(3):547–561, 1999.
- [134] E. Biewen, S. Nolte, and M. Rosemann. Perturbation by multiplicative noise and the Simulation Extrapolation method. *AStA Advances in Statistical Analysis*, 92(4):375–389, 2008.
- [135] K.B. Michels. A renaissance for measurement error. *International Journal of Epidemiology*, 30(3):421–422, 2001.
- [136] M.D. Hughes. Regression dilution in the proportional hazards model. *Biometrics*, 49(4):1056–1066, 1993.
- [137] J. Kuha and J. Temple. Covariate measurement error in quadratic regression. *International Statistical Review*, 71(1):131–150, 2003.
- [138] R. Doll, R. Peto, E. Hall, K. Wheatley, and R. Gray. Mortality in relation to consumption of alcohol: 13 years observations on male British doctors. *BMJ*, 309(6959):911–918, 1994.
- [139] K. Poikolainen. Alcohol and mortality: a review. *Journal of Clinical Epidemiology*, 48(4):455–65, 1995.
- [140] A.G. Shaper, M. Walker, and G. Wannamethee. Alcohol and mortality in British men: explaining the U-shaped curve. *Lancet*, 2(8623):1267–73, 1988.
- [141] F. Boutitie, F. Gueyffier, S. Pocock, R. Fagard, and J.P. Boissel. J-shaped relationship

## References

---

- between blood pressure and mortality in hypertensive patients: New insights from a meta-analysis of individual-patient data. *Annals of Internal Medicine*, 136(6):438–448, 2002.
- [142] W. Cates. Contraception, unintended pregnancies, and sexually transmitted diseases: why isn't a simple solution possible? *American Journal of Epidemiology*, 143(4):311–318, 1996.
- [143] J. Polesel, L. Dal Maso, V. Bagnardi, A. Zucchetto, A. Zambon, F. Levi, C. La Vecchia, and S. Franceschi. Estimating dose-response relationship between ethanol and risk of cancer using regression spline models. *International Journal of Cancer*, 114(5):836–841, 2005.
- [144] S. Port, L. Demer, R. Jennrich, D. Walter, and A. Garfinkel. Systolic blood pressure and mortality. *Lancet*, 355(9199):175–180, 2000.
- [145] R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.
- [146] S. Lewington, T. Thomsen, M. Davidsen, P. Sherliker, and R. Clarke. Regression dilution bias in blood total and high-density lipoprotein cholesterol and blood pressure in the Glostrup and Framingham prospective studies. *European Journal of Cardiovascular Prevention & Rehabilitation*, 10(2):143–148, 2003.
- [147] V. Bagnardi, A. Zambon, P. Quatto, and G. Corrao. Flexible meta-regression functions for modeling aggregate dose-response data, with an application to alcohol and mortality. *American Journal of Epidemiology*, 159(11):1077–1086, 2004.
- [148] H. Schneeweiß and T. Augustin. Some recent advances in measurement error models and methods. *Allgemeines Statistisches Archiv*, 90(1):183–197, 2006.
- [149] T. Augustin and R. Schwarz. Cox's proportional hazards model under covariate measurement error - A review and comparison of methods. Technical report, Ludwig-Maximilians-Universität, München, 2001.
- [150] A. Guolo. Robust techniques for measurement error correction: a review. *Statistical Methods in Medical Research*, 17(6):555–580, 2008.
- [151] N.E. Day, N. McKeown, M.Y. Wong, A. Welch, and S. Bingham. Epidemiological assessment of diet: a comparison of a 7-day diary with a food frequency questionnaire using urinary markers of nitrogen, potassium and sodium. *International Journal of Epidemiology*, 30(2):309–317, 2001.
- [152] K.M. Flegal, P.M. Keyl, and F.J. Nieto. Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology*, 134(10):1233–1244, 1991.

- [153] D. Spiegelman, R.J. Carroll, and V. Kipnis. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statistics in Medicine*, 20(1):139–160, 2001.
- [154] C. Frost and S.G. Thompson. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A*, 163(2):173–189, 2000.
- [155] K.B. Michels, S.A. Bingham, R. Luben, A.A. Welch, and N.E. Day. The effect of correlated measurement error in multivariate models of diet. *American Journal of Epidemiology*, 160(1):59–67, 2004.
- [156] D.G. Clayton. *Models for the analysis of cohort and case-control studies with inaccurately measured exposures*, pages 301–331. Oxford University Press, 1991.
- [157] R.J. Carroll, J.D. Maca, and D. Ruppert. Nonparametric regression in the presence of measurement error. *Biometrika*, 86(3):541–554, 1999.
- [158] G.H. Golub and J.H. Welsch. Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23(106):221–230, 1969.
- [159] L.K. Chan and T.K. Mak. On the polynomial functional relationship. *Journal of the Royal Statistical Society: Series B*, 47(3):510–518, 1985.
- [160] C.L. Cheng and H. Schneeweiß. Polynomial regression with errors in the variables. *Journal of the Royal Statistical Society: Series B*, 60(1):189–199, 1998.
- [161] L.A. Stefanski. Unbiased estimation of a nonlinear function a normal mean with application to measurement error models. *Communications in Statistics — Theory and Methods*, 18(12):4335–4358, 1989.
- [162] T. Nakamura. Proportional hazards model with covariates subject to measurement error. *Biometrics*, 48(3):829–838, 1992.
- [163] T. Augustin. An exact corrected log-likelihood function for Cox’s proportional hazards model under measurement error and some extensions. *Scandinavian Journal of Statistics*, 31(1):43–50, 2004.
- [164] T. Augustin, A. Doring, and D. Rummel. Regression calibration for Cox regression under heteroscedastic measurement error: Determining risk factors of cardiovascular diseases from error-prone nutritional replication data. *Recent Advances in Linear Models and Related Areas: Essays in Honour of Helge Toutenburg*, pages 253–278, 2008.
- [165] S.J. Novick and L.A. Stefanski. Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, 97(458):472–481, 2002.

## References

---

- [166] L.A. Stefanski and J.R. Cook. Simulation-Extrapolation: The measurement error jack-knife. *Journal of the American Statistical Association*, 90(432):1247–1256, 1995.
- [167] J.R. Cook and L.A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994.
- [168] W. Lederer and H. Küchenhoff. *simex: SIMEX- and MCSIMEX-Algorithm for measurement error models*. R package version 1.2.
- [169] J.W. Hardin, H. Schmiediche, and R.J. Carroll. The simulation extrapolation method for fitting generalized linear models with additive measurement error. *Stata Journal*, 3(4):373–385, 2003.
- [170] C.M. Crainiceanu, D. Ruppert, and J. Coresh. Cox models with nonlinear effect of covariates measured with error: a case study of chronic kidney disease incidence. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, 116, 2006.
- [171] R.J. Carroll, H. Küchenhoff, F. Lombard, and L.A. Stefanski. Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models. *Journal of the American Statistical Association*, 91(433):242–250, 1996.
- [172] J. Avorn, S. Schneeweiss, L.R. Sudarsky, J. Benner, Y. Kiyota, R. Levin, and R.J. Glynn. Sudden uncontrollable somnolence and medication use in Parkinson disease. *Archives of Neurology*, 62(8):1242–1248, 2005.
- [173] A. de Gramont, M. Buyse, J.C. Abrahantes, T. Burzykowski, E. Quinaux, A. Cervantes, A. Figer, G. Lledo, M. Flesch, L. Mineur, E. Carola, P. Etienne, F. Rivera, N. Chirivella, I. Land Perez-Staub, C. Louvet, T. Andre, I. Tabah-Fisch, and C. Tournigand. Reintroduction of oxaliplatin is associated with improved survival in advanced colorectal cancer. *Journal of Clinical Oncology*, 25(22):3224–3229, 2007.
- [174] T.J. Webb and R.P. Freckleton. Only half right: Species with female-biased sexual size dimorphism consistently break Rensch’s rule. *PLoS One*, 2(9):897, 2007.
- [175] V. Devanarayan and L.A. Stefanski. Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics & Probability Letters*, 59(3):219–225, 2002.
- [176] S. Nolte. The Multiplicative Simulation-Extrapolation Approach. *SSRN eLibrary*, 2007.
- [177] G. Ronning and M. Rosemann. SIMEX estimation in case of correlated measurement errors. *AStA Advances in Statistical Analysis*, 92(4):391–404, 2008.
- [178] J. Staudenmayer and D. Ruppert. Local polynomial regression and simulation-extrapolation. *Journal of the Royal Statistical Society: Series B*, 66(1):17–30, 2004.

- [179] R. Alpizar-Jara, L.A. Stefanski, K.H. Pollock, and J.L. Laake. Assessing the effects of measurement errors in line transect sampling. *North Carolina State University, Institute of Statistics Mimeograph Series*, 2508.
- [180] S.R. Cole, H. Chu, and S. Greenland. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology*, 35(4):1074–1081, 2006.
- [181] D.B. Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley and Sons, New York, 1987.
- [182] S. van Buuren and K. Groothuis-Oudshoorn. *mice: Multivariate Imputation by Chained Equations*, 2009. R package version 1.21.
- [183] I.R. White and P. Royston. Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28(15):1982–1998, 2009.
- [184] L.S. Freedman, V. Fainberg, V. Kipnis, D. Midthune, and R.J. Carroll. A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics*, 60(1):172–18, 2004.
- [185] J. Fan and Y.K. Truong. Nonparametric regression with errors in variables. *The Annals of Statistics*, 21(4):1900–1925, 1993.
- [186] A. Delaigle and P. Hall. Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *Journal of the American Statistical Association*, 103(481):280–287, 2008.
- [187] S.M. Berry, R.J. Carroll, and D. Ruppert. Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Society*, 97(457):160–169, 2002.
- [188] Y.J. Cheng and C.M. Crainiceanu. Cox models with smooth functional effect of covariates measured with error. *Journal of the American Statistical Association*, 104(487):1144–1154, 2009.
- [189] The Fibrinogen Studies Collaboration. Regression dilution methods for meta-analysis: assessing long-term variability in plasma fibrinogen among 27 247 adults in 15 prospective studies. *International Journal of Epidemiology*, 35(6):1570–1578, 2006.
- [190] S.A. Bashir and S.W. Duffy. Correction of risk estimates for measurement error in epidemiology. *Methods of Information in Medicine*, 34(5):503–510, 1995.
- [191] B. Liu, A. Balkwill, A. Roddam, A. Brown, and V. Beral. Separate and joint effects of alcohol and smoking on the risks of cirrhosis and gallbladder disease in middle-aged women. *American Journal of Epidemiology*, 169(2):153–160, 2009.
- [192] H. Küchenhoff, S.M. Mwalili, and E. Lesaffre. A general method for dealing with

## References

---

- misclassification in regression: The misclassification SIMEX. *Biometrics*, 62(1):85–96, 2006.
- [193] R.B. Israel, J.S. Rosenthal, and J.Z. Wei. Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings. *Mathematical Finance*, 11(2):245–265, 2001.
- [194] L. Natarajan. Regression calibration for dichotomized mismeasured predictors. *The International Journal of Biostatistics*, 5(1):12, 2009.
- [195] B. Efron. Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374):312–319, 1981.
- [196] J.K. Haukka. Correction for covariate measurement error in generalized linear models - A bootstrap approach. *Biometrics*, 51(3):1127–1132, 1995.
- [197] D. Burr. A comparison of certain bootstrap confidence intervals in the Cox model. *Journal of the American Statistical Association*, 89(428):1290–1302, 1994.
- [198] Fibrinogen Studies Collaboration. Correcting for multivariate measurement error by regression calibration in meta-analyses of epidemiological studies. *Statistics in Medicine*, 28(7):1067–1092, 2009.
- [199] B. A. Barron. The effects of misclassification on the estimation of relative risk. *Biometrics*, 33:414–418, 1977.
- [200] A. Kosinski and W. Flanders. Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification: A regression approach. *Statistics in Medicine*, 18:2795–2808, 1999.
- [201] R. Chu, P. Gustafson, and N. Le. Bayesian adjustment for exposure classification in case-control studies. *Statistics in Medicine*, 29:9941003, 2008.
- [202] L.S. Freedman, D. Midthune, R.J. Carroll, and V. Kipnis. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine*, 27(25):5195–5216, 2008.
- [203] P.T. Von Hippel. How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39(1):265–291, 2009.
- [204] E. Bonora, F. Calcaterra, S. Lombardi, N. Bonfante, G. Formentini, R.C. Bonadonna, and M. Muggeo. Plasma glucose levels throughout the day and HbA1c interrelationships in Type 2 diabetes. *Diabetes Care*, 24(12):2023–2029, 2001.
- [205] World Health Organisation. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: Report of a WHO/IDF consultation. Technical report, 2006.

- [206] E. Ritz and S.R. Orth. Nephropathy in patients with Type 2 diabetes mellitus. *New England Journal of Medicine*, 341(15):1127–1133, 1999.
- [207] F.S. Fein and E.H. Sonnenblick. Diabetic cardiomyopathy. *Cardiovascular Drugs and Therapy*, 8(1):65–73, 1994.
- [208] M. Wei, L.W. Gibbons, T.L. Mitchell, J.B. Kampert, M.P. Stern, and S.N. Blair. Low fasting plasma glucose level as a predictor of cardiovascular disease and all-cause mortality. *Circulation*, 101(17):2047–2052, 2000.
- [209] The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care*, 20(7):1183–1197, 1997.
- [210] E.S. Ford, W.H. Giles, and W.H. Dietz. Prevalence of the metabolic syndrome among US adults: Findings from the Third National Health and Nutrition Examination Survey. *JAMA: The Journal of the American Medical Association*, 287(3):356–359, 2002.
- [211] J.V. Bjornholt, G. Erikssen, E. Aaser, L. Sandvik, S. Nitter-Hauge, J. Jervell, J. Erikssen, and E. Thaulow. Fasting blood glucose: an underestimated risk factor for cardiovascular death. Results from a 22-year follow-up of healthy nondiabetic men. *Diabetes Care*, 22(1):45–49, 1999.
- [212] E.B. Levitan, Y. Song, E.S. Ford, and S. Liu. Is nondiabetic hyperglycemia a risk factor for cardiovascular disease?: A meta-analysis of prospective studies. *Archives of Internal Medicine*, 164(19):2147–2155, 2004.
- [213] E.L.M. Barr, P.Z. Zimmet, T.A. Welborn, D. Jolley, D.J. Magliano, D.W. Dunstan, A.J. Cameron, T. Dwyer, H.R. Taylor, A.M. Tonkin, T.Y. Wong, J. McNeil, and J.E. Shaw. Risk of cardiovascular and all-cause mortality in individuals with diabetes mellitus, impaired fasting glucose, and impaired glucose tolerance: The Australian Diabetes, Obesity, and Lifestyle Study (AusDiab). *Circulation*, 116(2):151–157, 2007.
- [214] J. Sung, Y. Song, S. Ebrahim, and D. A. Lawlor. Fasting blood glucose and the risk of stroke and myocardial infarction. *Circulation*, 119(6):812–819, 2009.
- [215] The DECODE study group. Is the current definition for diabetes relevant to mortality risk from all causes and cardiovascular and noncardiovascular diseases? *Diabetes Care*, 26(3):688–696, 2003.
- [216] J.E. Roeters van Lennep, H.T. Westerveld, D.W. Erkelens, and E.E. van der Wall. Risk factors for coronary heart disease: implications of gender. *Cardiovascular Research*, 53(3):538–549, 2002.
- [217] E.L. Barrett-Connor, B.A. Cohn, D.L. Wingard, and S.L. Edelman. Why is diabetes

## References

---

- mellitus a stronger risk factor for fatal ischemic heart disease in women than in men? *JAMA: The Journal of the American Medical Association*, 265(5):627–631, 1991.
- [218] M.G. Goldschmid, E. Barrett-Connor, S.L. Edelstein, D.L. Wingard, B.A. Cohn, and W.H. Herman. Dyslipidemia and ischemic heart disease mortality among men and women with diabetes. *Circulation*, 89(3):991–997, 1994.
- [219] A.J. Wells, P.B. English, S.F. Posner, L.E. Wagenknecht, and E.J. Perez-Stable. Misclassification rates for current smokers misclassified as nonsmokers. *American Journal of Public Health*, 88(10):1503–1509, 1998.
- [220] D.L. Patrick, A. Cheadle, D.C. Thompson, P. Diehr, T. Koepsell, and S. Kinne. The validity of self-reported smoking: a review and meta-analysis. *American Journal of Public Health*, 84(7):1086–1093, 1994.
- [221] M.E. Martinez, M. Reid, R. Jiang, J. Einspahr, and D.S. Alberts. Accuracy of self-reported smoking status among participants in a chemoprevention trial. *Preventive Medicine*, 38(4):492 – 497, 2004.
- [222] N.E. Day, M.Y. Wong, S. Bingham, K.T. Khaw, R. Luben, K.B. Michels, A. Welch, and N.J. Wareham. Correlated measurement error—implications for nutritional epidemiology. *International Journal of Epidemiology*, 33(6):1373–1381, 2004.
- [223] M.W. Knuiman, M.L. Divitini, J.S. Buzas, and P.E.B. Fitzgerald. Adjustment for regression dilution in epidemiological regression analyses. *Annals of Epidemiology*, 8(1):56–63, 1998.
- [224] L. Jack, L. Boseman, and F. Vinicor. Aging Americans and diabetes: A public health and clinical response. *Geriatrics*, 59(4):14–17, 2004.
- [225] J.M. Bland and D.G. Altman. Statistics Notes: Measurement error proportional to the mean. *BMJ*, 313(7049):106, 1996.
- [226] A. Chesher. Non-normal variation and regression to the mean. *Statistical methods in Medical Research*, 6(2):147–166, 1997.
- [227] D.G. Altman and J.M. Bland. Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society: Series D*, 32(3):307–317, 1983.
- [228] T. Tarpey, D. Yun, and E. Petkova. Model misspecification: finite mixture or homogeneous? *Statistical Modelling*, 8(2):199–218, 2008.
- [229] R.J. Carroll, K. Roeder, and L. Wasserman. Flexible parametric measurement error models. *Biometrics*, 55(1):44–54, 1999.
- [230] S. Richardson, L. Leblond, I. Jaussent, and P.J. Green. Mixture models in measure-

- ment error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society: Series A*, 165(3):549–566, 2002.
- [231] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, 1977.
- [232] T. Lim, R. Bakri, Z. Morad, and M.A. Hamid. Bimodality in blood glucose distribution. *Diabetes Care*, 25(12):2212–2217, 2002.
- [233] M. Borenstein, L.V. Hedges, J.P.T. Higgins, and H.R. Rothstein. *Introduction to meta-analysis*. Wiley, 2009.
- [234] D. Sharpe. Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review*, 17(8):881–901, 1997.
- [235] R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986.
- [236] J.P.T. Higgins, S.G. Thompson, and D.J. Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A*, 172(1):137–159, 2009.
- [237] J.P.T. Higgins, S.G. Thompson, J.J. Deeks, and D.G. Altman. Measuring inconsistency in meta-analyses. *BMJ*, 327(7414):557–560, 2003.
- [238] W. Viechtbauer. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30(3):261–293, 2005.
- [239] J.E. Hunter and F.L. Schmidt. *Methods of meta-analysis: correcting error and bias in research findings*. Sage, 2004.
- [240] L.V. Hedges. A random effects model for effect sizes. *Psychological Bulletin*, 93(2):388–395, 1983.
- [241] C.N. Morris. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983.
- [242] R. Harris, M. Bradburn, J. Deeks, R. Harbord, D. Altman, T. Steichen, and J. Sterne. Metan: Stata module for fixed and random effects meta-analysis. Statistical Software Components, Boston College Department of Economics, December 2006.
- [243] G. Schwarzer. *meta: Meta-Analysis*, 2008. R package version 0.9-15.
- [244] T. Lumley. *rmeta: Meta-analysis*, 2009. R package version 2.16.
- [245] W. Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010.

## References

---

- [246] J. Anzures-Cabrera and J.P.T. Higgins. Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods*, 1(1):66–80, 2010.
- [247] R.F. Galbraith. Some applications of radial plots. *Journal of the American Statistical Association*, 89(428):1232–1242, 1994.
- [248] J. Hartung, G. Knapp, and B.K. Sinha. *Statistical meta-analysis with applications*. Wiley, 2008.
- [249] R. Rosenthal and D.B. Rubin. Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, 99(3):400–406, 1986.
- [250] S.W. Raudenbush, B.J. Becker, and H. Kalaian. Modeling multivariate effect sizes. *Psychological Bulletin*, 103(1):111, 1988.
- [251] J.H. Zhao et al. with inputs from K. Hornik and B. Ripley. *gap: Genetic analysis package*, 2010. R package version 1.0-23.
- [252] D. Jackson, I.R. White, and S.G. Thompson. Extending DerSimonian and Laird’s methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine*, 29(12):1282–1297, 2010.
- [253] I.R. White. Multivariate random-effects meta-analysis. *Stata Journal*, 9(1):40–56, 2009.
- [254] A. Gasparrini. *mvmeta: multivariate meta-analysis and meta-regression.*, 2011. R package version 0.2.3.
- [255] D. Jackson, R. Riley, and I.R. White. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30(20):2481–2498, 2011.
- [256] S. Bingham and E. Riboli. Diet and cancer — The European prospective investigation into cancer and nutrition. *Nature Reviews Cancer*, 4(3):206–215, 2004.
- [257] A. Thompson. Thinking big: large-scale collaborative research in observational epidemiology. *European Journal of Epidemiology*, 24(12):727–731, 2009.
- [258] J.P.T. Higgins, A. Whitehead, R.M. Turner, R.Z. Omar, and S.G. Thompson. Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine*, 20(15):2219–2241, 2001.
- [259] R. Ma, D. Krewski, and R.T. Burnett. Random effects Cox models: A Poisson modelling approach. *Biometrika*, 90(1):157–169, 2003.
- [260] C.T. Smith, P.R. Williamson, and A.G. Marson. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics In Medicine*, 24(9):1307–1319, 2005.
- [261] T. Therneau. *coxme: Mixed Effects Cox Models.*, 2009. R package version 2.0.

- [262] T. Mathew and K. Nordström. On the equivalence of meta-analysis using literature and using individual patient data. *Biometrics*, 55(4):1221–1223, 1999.
- [263] R.D. Riley, M.C. Simmonds, and M.P. Look. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology*, 60(5):431–439, 2007.
- [264] A.J. Sutton and J.P.T. Higgins. Recent developments in meta-analysis. *Statistics in Medicine*, 27(5):625–650, 2008.
- [265] J. Schwartz and A. Zanobetti. Using meta-smoothing to estimate dose-response trends across multiple studies, with application to air pollution and daily death. *Epidemiology*, 11(6):666–672, 2000.
- [266] A. Conde-Agudelo, A. Rosas-Bermudez, and A.C. Kafury-Goeta. Birth spacing and risk of adverse perinatal outcomes: A meta-analysis. *JAMA: The Journal of the American Medical Association*, 295(15):1809–1823, 2006.
- [267] M. Rota, R. Bellocco, L. Scotti, I. Tramacere, M. Jenab, G. Corrao, C. La Vecchia, P. Boffetta, and V. Bagnardi. Random-effects meta-regression models for studying non-linear dose-response relationship, with an application to alcohol and esophageal squamous cell carcinoma. *Statistics in Medicine*, 29(26):2679–2687, 2010.
- [268] A.M. Scanu and G.M. Fless. Lipoprotein (a). Heterogeneity and biological relevance. *Journal of Clinical Investigation*, 85(6):1709–1715, 1990.
- [269] S.M. Marcovina, M.L. Koschinsky, J.J. Albers, and S. Skarlatos. Report of the national heart, lung, and blood institute workshop on lipoprotein (a) and cardiovascular disease: recent advances and future directions. *Clinical Chemistry*, 49(11):1785–1796, 2003.
- [270] A. Noma, A. Abe, S. Maeda, M. Seishima, K. Makino, Y. Yano, and K. Shimokawa. Lp(a): an acute-phase reactant? *Chemistry and Physics of Lipids*, 67:411–417, 1994.
- [271] B.G. Nordestgaard, M.J. Chapman, K. Ray, J. Borén, F. Andreotti, G.F. Watts, H. Ginsberg, P. Amarenco, A. Catapano, and O.S. Descamps for the European Atherosclerosis Society Consensus Panel. Lipoprotein(a) as a cardiovascular risk factor: current status. *European Heart Journal*, 31(23):2844, 2010.
- [272] P.C. Sharpe, I.S. Young, and A.E. Evans. Effect of moderate alcohol consumption on Lp(a) lipoprotein concentrations: Reduction is supported by other studies. *BMJ*, 316(7145):1675, 1998.
- [273] E. Bruckert, J. Labreuche, and P. Amarenco. Meta-analysis of the effect of nicotinic acid alone or in combination on cardiovascular events and atherosclerosis. *Atherosclerosis*, 210(2):353–361, 2010.

## References

---

- [274] A. Bennet, E. Di Angelantonio, S. Erqou, G. Eiriksdottir, G. Sigurdsson, M. Woodward, A. Rumley, G.D.O. Lowe, J. Danesh, and V. Gudnason. Lipoprotein(a) levels and risk of future coronary heart disease: Large-scale prospective data. *Archives of Internal Medicine*, 168(6):598–608, 2008.
- [275] L.S. Jonsdottir, N. Sigfusson, V. Gunason, H. Sigvaldason, and G. Thorgeirsson. Do lipids, blood pressure, diabetes, and smoking confer equal risk of myocardial infarction in women as in men? The Reykjavik Study. *Journal of Cardiovascular Risk*, 9(2):67–76, 2002.
- [276] G. Utermann, H.J. Menzel, H.G. Kraft, H.C. Duba, H.G. Kemmler, and C. Seitz. Lp(a) glycoprotein phenotypes. inheritance and relation to lp(a)-lipoprotein concentrations in plasma. *Journal of Clinical Investigation*, 80(2):458–65, 1987.
- [277] S.N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC., 2006.
- [278] D.G. Altman and P. Royston. The cost of dichotomising continuous variables. *BMJ*, 332(7549):1080, 2006.
- [279] C.J. Howe, S.R. Cole, D.J. Westreich, S. Greenland, S. Napravnik, and J.J. Eron Jr. Splines for trend analysis and continuous confounder control. *Epidemiology*, 22(6):874–875, 2011.
- [280] L. Desquilbet and F. Mariotti. Dose-response analyses using restricted cubic spline functions in public health research. *Statistics in Medicine*, 29(9):1037–1057, 2010.
- [281] F.E. Harre, K.L. Lee, and B.G. Pollock. Regression models in clinical studies: determining relationships between predictors and response. *Journal of the National Cancer Institute*, 80(15):1198–1202, 1988.
- [282] A.M. Jurek, G. Maldonado, S. Greenland, and T.R. Church. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *European Journal of Epidemiology*, 21(12):871–876, 2006.
- [283] J.M. Bland and D.G. Altman. Statistics Notes: Measurement error. *BMJ*, 313(7059):744, 1996.
- [284] B.G. Armstrong. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occupational and Environmental Medicine*, 55(10):651–656, 1998.
- [285] J.A. Hutcheon, A. Chiolero, and J.A. Hanley. Random measurement error and regression dilution bias. *BMJ*, 340(7761):1402–1406, 2010.
- [286] J. Sexton and P. Laake. Boosted regression trees with errors in variables. *Biometrics*, 63(2):586–592, 2007.

- 
- [287] X.-F. Wang and B. Wang. Deconvolution estimation in measurement error models: The R Package *decon*. *Journal of Statistical Software*, 39(10):1–24, 2011.
- [288] A. Delaigle and A. Meister. Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association*, 102(480):1416–1426, 2007.
- [289] D. Bennett, J. Little, L. Masson, and C. Minelli. The empirical investigation of methods to correct for measurement error in biobanks with dietary assessment. *BMC Medical Research Methodology*, 11(1):135, 2011.
- [290] P. Royston, W. Sauerbrei, and H. Becher. Modelling continuous exposures with a ‘spike’ at zero: A new procedure based on fractional polynomials. *Statistics in Medicine*, 29(11):1219–1227, 2010.
- [291] B.L. Heitmann and L. Lissner. Dietary underreporting by obese individuals — is it specific or non-specific? *BMJ*, 311(7011):986–989, 1995.