The design of composite indicators of healthcare quality:

a multi-method analysis

Matthew Edward Barclay

Jesus College

January 2021

This thesis is submitted for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Abstract. The design of composite indicators of healthcare quality: a multi-method analysis

Matthew Edward Barclay

High profile composite indicators seeking to summarise aspects of the quality of healthcare organisations often have important impacts: they may be reputationally consequential for healthcare organisations, may affect prioritisation for inspection or action by regulators, and may be responsible for effecting changes in the way services are delivered. Use of composite indicators, which include examples such as the Hospital Compare Star Ratings in the US and the Quality and Outcomes Framework in the UK, is proliferating in healthcare. But to date there is limited understand of how such indicators should be designed and reported. This thesis seeks to address this gap, characterising relevant challenges in the design of composite indicators and exploring their reporting might be improved.

My examination of composite indicators is formed from two principal components. The first comprises two quantitative studies, which are conceptually linked. These studies investigate the sensitivity of the ratings and rankings of hospitals to the exact technical methods by which composite indicators are produced. I show that composite indicators can be highly sensitive to technical decisions, and that different composite indicators are affected differently by these, rarely examined, technical issues. The second component is a qualitative interview study, for which I interviewed experts in performance measurement about the process of developing a composite indicator. This study highlighted two crucial aspects of the development of composite indicators. First, the purpose of a composite indicator needs to drive every aspect of technical decisions relating to their specification. Second, the development process should be iterative: developers must be willing and able to revise the purpose of the indicator and restart the development process if it subsequently becomes apparent that the composite indicator does not match the purpose for which it is being developed.

The results of the quantitative and qualitative studies in this thesis revealed a number of challenges for the design of composite indicators, but also helped to identify possible ways for improving their design. The qualitative study catalogues important decision points which could be used as a framework for design choices for composite indicators. In combination

with the results of the quantitative studies, which inter alia show the usefulness of applications of Monte Carlo sensitivity analysis as part of the development and evaluation of composites, this catalogue of decisions holds promise for identifying and addressing problems with composite indicator in the design stage and for offering critiques of existing indicators. Together these considerations point toward designing-in an iterative approach to processes used in the development of composite indicators, in which sensitivity analysis is used to assess the implications of each design decision. Finally, my results lay the groundwork for future development of a reporting guideline for the development and reporting of composite indicators of healthcare quality.

Acknowledgements

This PhD has only been possible because of the support I have received from many people.

My supervisors, Mary Dixon-Woods and Georgios Lyratzopoulos, have been brilliant. I worked with both prior to my PhD, covering issues with individual performance measures ranging from case-mix to missing data. This research laid the ground for the in-depth assessment of composite indicators presented this thesis. They have both been very supportive throughout, from the day I first went to them to discuss the possibility of researching the design of composite indicators in healthcare right up until today as I put the finishing touches on my thesis. They have changed my perspective on research, and for the better.

My friends in Cambridge pushed me to do a PhD in the first place, and kept me going when I was struggling. There are too many names to mention here. But some people really made a difference, and I would give particular thanks to Audrey, Brett, Cong, David, Elisa, Gary, Guillaume, Jag, Karolina, Katie, Lawrance, Sarah, and Tish.

Many people gave me helpful advice about my research. I am very grateful to the participants in my interview study and those who helped in recruitment. Without them, much of my research would not have been possible or would have been far more limited.

I also want to acknowledge the role that music has had for me, as stress relief and as a way of focusing on something entirely different. My main distractions were the Grad Orchestra and the Ceilidh Band, and Thursday night sessions in the Anchor. But I am also grateful to various string quartets, college orchestras, and the occasional CUMS ensemble for reminding me of life outside public health. The last few years would have been far more boring without it.

This research project was funded by a combination of Mary Dixon-Woods' Wellcome Trust Investigator award (grant number WT097899) , the Health Foundation's award to The Healthcare Improvement Studies (THIS) Institute at the University of Cambridge and Georgios Lyratzopoulos's Departmental discretionary fund. The Health Foundation is an independent charity committed to bringing about better health and health care for people in the UK.

Contents

Declaration ... iii

Abstract. The design of composite indicators of healthcare quality: a multi-method analysis ... iv

Acknowledgements ... vi

Contents ... vii

1 Background: Why study technical choices in developing composite indicators of healthcare quality? ... 9

1.1 Thesis aims and objectives ... 10
1.2 Thesis structure ... 10
1.3 Composite indicators and their uses in healthcare quality and safety ... 10
1.4 Developing composite indicators of healthcare quality ... 14
1.5 Terminology: (individual) measures versus (composite) indicators ... 16
1.6 Problems with composite indicators ... 18
1.7 Discussion and conclusions ... 24

2 Methods ... 28

2.1 The rationale behind the quantitative studies ... 28
2.2 The impact of alternative technical approaches in the design of the Sentinel Stroke National Audit Programme Score and Level ... 54
2.3 Additional technical notes for the quantitative studies ... 70
2.4 Understanding decisions in the development and reporting of composite indicators: interview study ... 76
2.5 Summary ... 81

3 Results 1: An analysis using secondary data of the impact of alternative technical approaches in the design of the CMS Hospital Compare Star Ratings ... 82

3.1 Summary ... 82
3.2 Introduction ... 83
3.3 Summary of methods ... 83
3.4 Findings ... 87
3.5 Discussion ... 105

4 Results 2: An analysis using secondary data of the impact of alternative technical approaches in the design of the SSNAP score and level ... 109

4.1 Summary 109
4.2 Introduction 111
4.3 Summary of methods 113
4.4 Findings 119
4.5 Discussion 136

5 Results 3: Qualitative interview study of experts' views on good practice in developing composite indicators 143

5.1 Summary 143
5.2 Introduction 144
5.3 Study aims and objectives 145
5.4 Summary of methods 145
5.5 Results 146
5.6 Discussion 181

6 Discussion and reflections: Improving composite indicators of healthcare quality 192

6.1 Understanding design and reporting of composite indicators 193
6.2 Understanding the consequences of the technical approaches used to create a composite indicator 199
6.3 Researching complex methodological questions and the interplay between quantitative and qualitative approaches 204
6.4 Common problems with existing composite indicators – and potential strategies for mitigation 207
6.5 Limitations of this thesis 215
6.6 Conclusions: routes toward better designed and reported composite indicators 218

References 221

Appendices 238

Appendix 1. Current approaches to individual measure standardisation for each of the measures used in the SSNAP score and level 238
Appendix 2. Prompt guide for the interview study 249
Appendix 3. Approaches to combining measures into composite indicators 251
Appendix 4. Academic papers published during PhD 253
Appendix 5. Presentations given during PhD 255
Appendix 6. Conferences attended 256
Appendix 7. Formal professional development undertaken during PhD 256

1 Background: Why study technical choices in developing composite indicators of healthcare quality?

Composite indicators, where individual measures are compiled into a single indicator of healthcare quality [1], are a prominent feature of health policy in many countries. They are now reported across the world by a variety of organisations including governments [2–4], newspapers [5], commercial companies [6], charities [7], and clinical audits [8]. Used by patients to help choose hospitals, by policy-makers to make decisions, and by researchers to explore features of health systems, composite indicators (for example in the form of “star ratings”) have powerful reputational and other effects for organisations [9]. Consideration of the design of composite indicators is important given rising demand for simple, actionable summaries of individual quality measures [10]. Yet a glance at the academic literature (e.g. [11,12]) and specialist newspapers (e.g. [13,14]) shows that composite indicators frequently suffer avoidable flaws and limitations. Though decisions about the technical specifications of indicators can influence hospital rankings for individual measures such as hospital mortality rates [15], these choices have rarely been explored systematically in relation to composite indicators specifically.

Accordingly this thesis examines the design of composite indicators of quality in healthcare. I am specifically concerned to investigate the impact of different technical specifications upon organisational rankings and performance, and to explore how to improve the robust and transparent reporting and documentation of composite indicators. Combining a number of different research approaches in a multi-method (quantitative and qualitative) study, I show that composite indicators depend on a range of methodological choices, ranging from the choice of constructs to be measured (e.g. safety of care, patient experience, waiting times) to how the results are presented (e.g. summary scores, star ratings, report cards), and I offer scrutiny of these choices.

1.1 Thesis aims and objectives

My thesis has two broad and inter-related aims:

1. To characterise relevant challenges in the design of composite indicators.
2. To explore how reporting of composite indicators might be improved.

1.2 Thesis structure

This thesis comprises six chapters.

In this first chapter, I introduce the concept of composite indicators and set out definitions and notes on terminology.

The second chapter describes the methods used in the empirical studies in my PhD.

The third and fourth chapters examine the sensitivity to plausible alternative technical specifications of actual composite indicators in current use, and primarily intended to address my first aim.

The fifth chapter presents the results of an interview study with series of experts about the decisions involved in producing composite indicators, and primarily intended to address my second aim.

In the final chapter, I discuss and reflect on the thesis as a whole and set out my overall conclusions.

1.3 Composite indicators and their uses in healthcare quality and safety

In the sections that follow, I explore the challenges in current practice in design and reporting of composite measures of quality in healthcare. These sections are closely based on a paper I published at the initial stages of my PhD [16]. I identify a range of problems with composite indicators as currently deployed, thus setting up my later focus on identifying decisions in developing composite indicators and on describing their fundamental importance.

Composite indicators of healthcare quality and safety have only been widely produced and reported for around 30 years, though they are a continuation of a far older tradition of measuring organisational healthcare performance. Their potential use as a summary of quality in organisations is appealing to policy-makers and perhaps the public, but they are open to critique and criticism.

1.3.1 A brief history of composite indicators of quality and safety in healthcare

Examples of performance measurement in healthcare go back to the dawn of civilisation [17,18]. The *Codex Hammurabi* is an early example, dating from around 4000 years ago. This set out financial incentives for successful surgical operations, and harsh penalties for unsuccessful ones. In contrast with modern systems of performance measurement, this did not involve public reporting or peer comparisons. Instead, those who performed unsuccessful surgeries would receive punishments ranging from a monetary fine to the death penalty [17].

Public reporting of quality measures about hospitals is a more recent development. Florence Nightingale was a pioneer, both in her measurement of mortality rates in army hospitals during the Crimean War and in her work examining hospitals in London [19,20]. Ernest Codman was another early pioneer [21], publishing outcomes for his own hospital and helping found a predecessor body to the Joint Commission on Accreditation of Hospitals [22]. The challenge of collecting data meant public reporting for the next 100 years followed a similar pattern, for example the Nuffield Trust's 1962 publication *Further Studies in Hospital and Community* [23], examining the reason for admission, length of stay and longer-term outcome for patients discharged from hospitals in Aberdeen, Dundee and Glasgow. By 1965, accreditation by the Joint Commission in the US was required for hospital to treat Medicare patients [22]. This included standards on record-keeping, necessary to allow comparison of outcomes between hospitals. And in 1966 Avedis Donabedian, a health services researcher at the University of Michigan, was setting out his categorisation of healthcare performance measures into structure, process and outcome [24].

It can be argued that the modern era of performance measurement began in the late 1980s, with the advent of digital records capturing aspects of the processes of healthcare (appointments, tests, procedures, hospital stays), often created to automate billing for care. Suddenly, data on healthcare was being collected on a routine basis, rather than in one-off audits or research studies, making it far more efficient to measure performance. In the US, the Joint Commission were developing an indicator-based performance measurement system [22], while in the UK there was growing interest in the development and reporting of performance measures based on vast amounts of routine data collected by the NHS [25,26]. More routine performance measurement allowed for regular comparisons of hospitals and individual physicians [27–29]. While performance measurement was often conceptualised in terms of professionally-led audit and quality improvement, reporting of performance data especially in the US (where competition of different healthcare providers is much more established) was also promoted as a way of helping patients choose the hospital or surgeon with best outcomes for their condition [28,30,31].

The first composite indicators of healthcare quality were introduced as routine reporting of individual performance measures became common. One of the first was the US News & World Report "Best Hospitals" ranking [32,33], which from 1993 onwards was based on a combination of hospital reputation among physicians, structural measures such as the ratio of registered nurses to beds, and outcome measures – initially just mortality rates. As is common for composite indicators in healthcare, it repurposed data intended for other purposes (such as reimbursement).

My scoping of the literature suggests that there followed a slow proliferation of composite indicators of healthcare quality and safety.

- In the US:
    - IBM Watson Health began publishing its annual 100 Top Hospitals study in 1993 [34].
    - The Agency for Healthcare Research and Quality developed the PSI-90 composite safety indicator in 2008 [35].
    - The Leapfrog Group began publishing the Hospital Safety Grade in 2012 [36,37].
    - Consumer Reports began publishing the Safety Score in 2012 [37,38]

    - US News & World Report began reporting composite quality ratings for specific medical procedures (in addition to the overall quality of specific clinical areas and entire hospitals) in 2015 [39].
    - Centers for Medicare and Medicaid Services began publishing the Hospital Compare Overall Star Ratings in 2016 [2], with a precursor measuring only patient experience released in 2015 [40,41].
- In the UK:
    - The UK government began publishing the NHS Star Ratings, a composite indicator the quality of NHS hospitals, in 2001 [42].
    - The UK NHS introduced the Quality and Outcomes Framework in 2004 [43].
    - The Care Quality Commission developed its Intelligent Monitoring composite indicator in 2013 [3].
    - The Sentinel Stroke National Audit Programme (SSNAP) published the SSNAP score and level measuring the quality of stroke services in 2013 [8].
    - NHS England began publishing Overall Patient Experience Scores in 2015 [4].
    - NHS England introduced the CCG Improvement and Assessment Framework in 2016 [44], and the STP Progress Dashboard in 2017 [45].

- In the Netherlands, the newspaper AD has published annual rankings of hospitals since 2003 [5].

1.3.2 Current uses of composite indicators

Composite indicators currently appear generally to be used for the same inter-related purposes as other approaches to performance measurement. As alluded to above, these are quality assurance and performance management, quality improvement, and supporting patients in choosing where best to get treated. More broadly, composite indicators may be deployed where navigating the large selection of individual performance measures presents challenges.

One common use is in directly supporting patient choice. Most composite indicators of healthcare quality produced by charities and for-profit organisations are developed for this purpose, including the various US News & World Report composites [6,39], the AD Ziekenhuis Top 100 [5], and the Consumer Reports Safety Score [38]. The CMS Hospital Compare Star

Ratings are also principally intended to support patient choice [46]. These types of composite indicators often prioritise outcome measures over other measure types, and tend to be presented on websites designed so that patients can easily find their local hospitals.

Use in quality assurance and performance management is also reasonably common, especially in the UK. The original NHS Star Ratings were for this purpose [42], as was CQC Intelligent Monitoring [3] and the various NHS England indicators for commissioning groups [44,45]. CQC Intelligent Monitoring was a fascinating example. It was not in itself aiming to be a rating of quality; CQC produce those based on in-person inspections. But what it was meant to do was flag hospitals where there were potential concerns [47], to prioritise in-person inspections.

Composite indicators are also used in quality improvement activity. The NHS Quality and Outcomes Framework is one such example [43], as are the Sentinel Stroke National Audit Programme (SSNAP) Score and Level [8]. Such indicators may well be presented to the public – for example, the SSNAP Level used to be on the public-facing MyNHS website (now closed due to lack of use [48]) – but are distinguished from composite indicators aimed at supporting patient choice by their heavy focus on process measures of quality, and the inclusion of few outcome measures. The use of process measures makes pathways to improving performance more obvious. It is far easier to, say, work on getting acute stroke patients admitted to the stroke unit more quickly than it is to reduce mortality from stroke.

1.4 Developing composite indicators of healthcare quality

Developing a composite indicator is a multi-step process [1,18,49–51] involving aggregation of multiple organisational-level measures to create a summary score (Box 1 gives a simple mathematical description). While there are different ways to produce a composite, a typical approach in healthcare might look like this:

1. Identification of individual measures that would be appropriate to include in the composite.
2. Standardisation of the individual measures, so that all measures are on comparable scales.

3. Assignment of appropriate weights to each individual measure.
4. Combination of the identified weighted and standardised individual measures to produce the composite indicator.

Articulating these processes immediately raises a series of questions:

1. How should the individual measures be chosen? What are the implications of this choice?
2. How should individual measures be standardised?
3. Who chooses the weights given to individual measures?
4. Are weights chosen for one purpose, by one group of people, appropriate for alternative uses of the composite indicator? Is a single set of weights possible?
5. Is it ever reasonable for good performance in one domain to 'make up' for bad performance in another?

And in practice, as is clear when examining existing composite indicators, there are frequently additional steps and other complications in this process. But in principle every composite indicator examined in this thesis could be written in the mathematical form set out in Box 1.

Box 1. A mathematical definition of a simple composite indicator, adapted from Freudenberg 2003 [22]

Let p_i be an individual performance measure, and $p_{i,h}$ be the performance on this measure for hospital h. Let w_i be the associated weight, and f_i be some function from the natural scale of p_i to some 'standard' scale.

A composite indicator of performance for hospital h, I_h, is given by

$$I_h = \sum_i w_i \times f_i(p_{i,h})$$

1.5 Terminology: (individual) measures versus (composite) indicators

Hereafter in this thesis, a heuristic distinction is drawn between the terms 'measure' and 'indicator', which are often used interchangeably in the existing literature. This convention will make it easier to follow the thesis by clarifying whether sections of text discuss individual measures or composites.

- *Measure(s)* is used only to refer to individual performance measures that might be combined to produce a composite indicator (for example, postoperative mortality rate).
- *Indicator(s)* is used only to refer to a composite indicator (for example, the CMS Hospital Compare Overall Star Ratings composite indicator) and also domain indicators of overall composite indicators (for example, the mortality domain indicator of the CMS Hospital Compare Overall Star Ratings composite indicator).

1.5.1 Other types of composite indicator outside the scope of the thesis

The above definition describes the composite indicators of healthcare quality that I will discuss in this thesis. But there are two other approaches to producing performance measures that are often described as composite indicators, and it is useful to distinguish them to make clear that they are outside the scope of the thesis. These are the 'composite endpoint' and the 'enhanced outcome measure'.

1.5.1.1 The patient-level composite endpoint

Composite endpoints are collections of performance measures aggregated at the patient-level (Box 2). For example, a composite endpoint measure of surgical quality might ask if a patient had any negative outcomes following surgery – death, readmission, complications and so forth. Indeed, in the Netherlands surgical quality is judged on whether patients have a so-called 'textbook outcome', where they have no major post-surgical problems [52,53].

Composite endpoints are frequently used in clinical trials – for example cancer treatment may be considered to fail if a patient either relapses or dies. While some of the topics raised in this thesis can be applicable to patient-level composite endpoints, those considering using such measures should consult the detailed discussions of the benefits and difficulties of using composite endpoints in the clinical trials literature [54,55].

Box 2. A composite endpoint, a simplified version of the Textbook Outcome endpoint proposed by Kolfschoten and colleagues for colon cancer resection [51].

Consider patients undergoing major surgery. Three key measures of good surgical outcome may be

1. Survived 30 days post-procedure ($Surv_{30d}$)
2. No major surgical complications (No_Major_Comp)
3. Total length-of-stay < 7 days ($LOS_{<7d}$)

The composite endpoint $Good\_Surg\_Outome$ may then be defined as:

$$Good\_Surg\_Outcome = \begin{cases} 1 & \text{if } Surv_{30d} \text{ and } No\_Major\_Comp \text{ and } LOS_{<7d} \\ 0 & \text{otherwise} \end{cases}$$

The hospital-level score is then the proportion of patients with $Good\_Surg\_Outome = 1$.

Refinements to composite endpoints when used in quality measurement typically modify the patient-level scoring system such that it reflects, for example, that death is a less desirable outcome than an extended stay in hospital.

1.5.1.2 The ‘enhanced outcome’ measure

By ‘enhanced outcome’ measure, I mean the use of additional variables to give more statistical precision to an individual performance measure. Low reliability is a common problem with outcome measures of healthcare quality [56–59], and one solution to this is to use a shrinkage or reliability-adjusted measure [60]. These reliability-adjusted measures shrink extreme performance toward the overall mean producing more accurate measures under the assumption that hospital performances will usually be relatively similar. Where there is additional information about hospital performance, for example relevant process or structural measures, it may be possible to do even better. For surgical mortality measures, for example, one may expect a volume-outcome relationship and so one may intuitively wish to shrink performance toward the mean of hospitals treating a similar number of cases each year, rather than the overall mean. This idea motivated the so-called composite indicator of hospital-level postoperative mortality proposed by Dimick and colleagues [61]. This approach is not considered further in this thesis.

1.6 Problems with composite indicators

Little in the literature offers an overview of the problems with composite indicators of quality in healthcare. Though there are critical discussions, they typically focus on single major problems with individual composite indicators, and empirical investigations remain rare. For example, Rajaram, Barnard and Bilimoria challenge the design of a composite indicator of patient safety based on the validity of the individual measures the indicator is based upon [62], but do not offer a demonstration that these problems of validity lead to misclassification of hospital performance.

I drew on my background reading for the thesis to write an article in the “Problems with…” series in *BMJ Quality and Safety* [16] that was intended as a critical analysis and overview of the field rather than a formal literature review. Rather than being an aggregative synthesis of studies identified via a protocol-driven search strategy, it is an narrative overview of themes in the literature identified through multiple strategies [63,64], including database searches, reference chaining and examining websites presenting composite indicators. Here, I draw on and extend that published piece, first summarising the common problems and then highlighting areas where further research is needed, and in particular setting out the research questions I aimed to answer in this thesis. This discussion is focused primarily on composite indicators used in relation to hospitals, but is likely to be generalisable to other settings.

1.6.1 Lack of transparency

The first, and perhaps most important, current problem with composite indicators is a lack of transparency. By transparency, I mean explaining what the indicator means, how it has been produced, and why the indicator has been produced in the way it has. Transparency is required for three main reasons: to help users understand the (intended) meaning of a composite; to promote trust in the validity of the composite; and to allow for independent validation and replication of the processes used to create the indicator [65].

Lack of transparency harms understanding of the meaning of indicators when, for example, the reporting fails to make clear how differences in the composite score should be relevant to the user. There are many examples. The Leapfrog Hospital Safety Grade assigns US hospitals a rating from A through E, but even in the document “Explanation of Hospital Safety Grades” the meaning of these ratings is not explained beyond an A grade being best [66]. It is difficult to understand whether a hospital receiving an F grade is not safe, or indeed whether a hospital receiving an A grade is very safe. Similarly, it is unclear how much weight one should give to the difference between a ‘5-star’ and a ‘4-star’ hospital according to the CMS Hospital Compare Overall Star Ratings.

Trust in the validity of composites is undermined when, for examples, the processes by which decisions are made about what gets measured and how it is measured are not clear or accountable. Clarity is always needed about the role of different stakeholders in selecting measures for inclusion in composite measures, including the respective contributions of members of the public, clinicians, and payers and policy-makers. This is all the more important when composite indicators are deployed as drivers of performance improvement or linked to pay-for-performance criteria [67].

One major problem with lack of transparency is that independent reproduction of composite indicators is often rendered impossible, because the technical documentation of composite indicators frequently fails to include sufficient technical detail. Sometimes composite indicators are published without any technical documentation at all [68]. More frequently, some technical information may be published but is insufficient, and may be reported in a different place from organisational scores on the actual composite indicator [69,70].

1.6.2 What goes into baskets of measures matters

The second common problem with composite indicators is that their individual parts often do not, when taken together, give a fair summary of the whole [67]. This problem can arise via several routes.

Composite indicators purporting to provide a broad overview of organisational quality may be dominated by a few clinical areas or by surveillance measures that are unsuitable for measuring quality. This dominating effect may occur because of pragmatic decisions to rely on data that is readily to hand (a form of ‘availability bias’), as I outline in Table 1. For example, more than one in five (15/37) of the individual underlying measures for CMS Star Ratings relate to care for cardiovascular disease, including half (8/16) of the highly-weighted mortality and readmission measures [71]. When indicators are dominated in this way by measures of specific clinical fields they may incentivise hospitals to focus on measured disease areas at the expense of those not directly measured [67,72,73].

Composite indicators can also be affected by structurally absent information, such as inclusion in the indicator of cardiac surgery performance measures for hospitals not providing cardiac surgery. This is not a missing data issue, rather one of irrelevance: certain performance measures are simply not applicable to particular organisations. In the CMS Star Ratings, the same methods and measures are used to produce ratings for all hospitals publicly reporting quality information on Hospital Compare [2], including speciality hospitals. Yet such hospitals report fewer measures than general hospitals, and are substantially more likely to be classed as high-performing than the average hospital, with 87% of them receiving 4 or 5 stars in 2015 compared with 28% of all hospitals [74]. It is plausible that the relevant subset of general quality measures do not appropriately reflect the quality of care provided by specialist hospitals.

Occasionally, measures are used without clear conceptual justification: one US scheme uses operating profit margin as a measure of quality, for example, yet why this should reasonably be seen as an indicator of (clinical) quality is not clear [75].

Table 1. Specific issues with selected composite indicators of care quality.

Problem	Specific issue	CMS Overall Hospital Star Rating	AHRQ PSI90	Leapfrog Composite Patient Safety Score	MyNHS Overall Stroke Care Rating	NHS England Overall Patient Experience Score
Lack of transparency	Is there a single public document with all important methodological details?	No	No	Yes	No	Yes
What goes into the baskets of measures matters	Are specialist and general hospitals compared on the same measures?	Yes	Yes	No	Yes *	Yes
Threats arising from issues with underlying measures and data	Is missing measure information handled in a way that can introduce bias?	Yes	No	Yes	Yes	No
	Are all component measures adequately adjusted for case-mix?	No	Yes	Yes	Unclear	Yes
Banding to get measures onto consistent scales	Are measures standardised using banding?	No	No	No	Yes	No
Choosing appropriate weights to combine measures	Is the choice of weights justified?	Limited	Limited	Limited	No	No
	Has there been sensitivity analysis of the choice of weights?	No	Yes	No	No	No
Failure to present uncertainty	Is the uncertainty in the final composite rating presented?	Not in the star rating	Yes	No	No	Yes

*As this composite is for a clinical service rather than a hospital, it is reasonable to compare general and specialist trusts on the same measure

1.6.3 Threats arising from issues with underlying measures and data

Composite indicators, by their nature, obscure details about the underlying (individual) measures, yet problems with the former can render the composite meaningless. At minimum, the underlying measures must represent valid measures of quality. To achieve this, they need to be adequately and appropriately adjusted for case-mix in order to avoid bias in the overall composite. But not all composite indicators in current and widespread use meet this basic standard. Thus, for example, lack of adjustment for sociodemographic factors in readmission measures included the CMS Star Ratings means that hospitals serving more disadvantaged communities may receive lower ratings for reasons that are outside the hospital's control [11].

Problems also occur when composite indicators rely on quality measures that are not available for all hospitals or that are missing for many patients despite in principle being possible to collect. Fair comparisons rely on understanding why data to calculate these quality measures have not been collected for certain hospitals (or certain patients) in order to decide whether to use a measure and, if so, how to make appropriate adjustments to reduce bias. Surveillance bias, whereby organizations vary in efforts expended on collecting indicator data, may result in hospitals with the same underlying performance appearing different [76,77]. Notably, the proportion of patients with missing data for a given measure may vary substantially between organisations, potentially having a major impact on the comparative performance of different hospitals [62]. Sometimes disclosure rules play a part in variations in the proportion of missing data between different hospitals, frequently impacting composite indicators that use data from national public reporting schemes. In other circumstances, data are simply not collected or available. The Leapfrog Hospital Safety Grade, a composite indicator of patient safety, for example, uses information from a voluntary survey of hospitals, but underlying measures are not available for hospitals that do not complete it [7].

In practice, schemes often use *ad hoc* methods to handle problems with underlying data, with several simply calculating ratings as the weighted average of non-missing measures [2,35]. The CMS Star Ratings take this approach when producing overall summary scores, apparently favouring hospitals that do not provide or do not collect relevant data: hospitals that report a greater number of measured domains have systematically worse performance [74]. In this context, it is unclear whether hospital differences in CMS Star Ratings reflect genuine differences or bias due to improper handling of missing variables, or improper comparisons of hospitals providing different services (see also 1.6.2 regarding the choice of measures included in the 'basket').

1.6.4 Banding to get measures onto consistent scales

Many composite indicator schemes apply threshold-based classification rules to standardise disparate individual measures to a consistent scale. Measures that are naturally continuous are mapped to categorical bands before being combined into the overall composite [44,45,69]. For example, in the (recently retired) MyNHS Overall Stroke Care Rating, the individual measures were all mapped to 0 to 100 scales. Here, the continuous measure "median time between clock start and thrombolysis" was mapped to a score of 100 if <30 minutes, a score of 90 if between 30 and 40 minutes and so on [69]. This approach violates the general statistical principle that such categorisation reduces statistical power and potentially hides important differences [78]. Banding distorts apparent organisational performance: hospitals with median time to thrombolysis of 29:59 would be treated as having meaningfully different performance to those with median time 30:01. These differences are unlikely to reflect reality. The thresholds used to band performance are typically arbitrary, but the particular choice of threshold can have a serious impact on estimates of organisational performance [10,49].

1.6.5 Choosing appropriate weights to combine measures

The weighting assigned to individual measures contributing to composites is another problem area. As few hospitals perform equally well in all areas, performance can be artificially improved by giving higher weight to individual measures where a hospital performs better than average, and vice versa. The choice of weights given to individual measures is thus a key determinant of performance on the overall composite, and different weights might allow almost any rank to be achieved [79,80]. Therefore, transparency is needed about the importance attached to each measure in terms of the aim of the indicator, with supporting evidence. However, many schemes do not provide explicit justification for the weights used to create the composite (Table 1).

Not assigning any weights is also fraught with problems. The NHS England Overall Patient Experience Scores scheme, for example, does not allocate different weights to survey questions because "there was no robust, objective evidence base on which to generate a weighting" [4]. But that criticism is also applicable to the decision to adopt equal weights [50].

Similarly, the composite patient safety indicator AHRQ PSI90, since revised [81,82], originally gave greater weight to more common safety incidents [35], ignoring differences in the degree of potential harm to patients. The original specification gave a 21-fold greater weight to the incidence of pressure ulcers compared with postoperative hip fracture [35,62].

1.6.6 Failure to present uncertainty

Composite indicators are not immune to chance variation: tiny differences in individual measures can translate into differences in the final rating, but will often be due to chance [14]. Simulations show that around 30% of US hospitals might be expected to change CMS Star Rating from year-to-year due to chance alone [2]. Yet many composite indicators are presented without appropriate measures of uncertainty (Table 1, page 21), in defiance of expert recommendation and established practice for individual performance measures [27,83–85]. Of course, confidence intervals spanning multiple performance categories might lead users to view an indicator as meaningless: when comparing performance between two hospitals, it is easier to say one is three-star and the other four-star, rather than say that one is ‘between two and four stars’ and the other is ‘between three and five stars’. However, when there is a lot of uncertainty about hospital performance, hospitals might be penalised or rewarded for performance that may simply reflect the play of chance – making it especially important that reporting conventions are well-founded.

1.7 Discussion and conclusions

As is clear from my narrative review of the existing literature, much is already understood about the some of the challenges of developing composite indicators. I seek to add to this understanding, with the ultimate aim of supporting better reporting of composite indicators. The primary motivation for the empirical studies in this thesis is to identify information that should be reported as part of the composite indicator development process to satisfy the requirements for transparency. Doing so requires both understanding what the process of developing a composite indicator actually are, and understanding the practical consequences of the technical specification.

1.7.1 Understanding the consequences of the technical specification

Characterising relevant challenges in the design of composite indicators requires an understanding of the consequences of these challenges, similar to work from prior investigators examining the impact of technical decisions on individual performance measures. Examples include assessments of the way sample size limits what can be inferred from apparent variation in surgeon and hospital outcomes [56,59], exploration of how apparent regional differences in recording of comorbidities affect case-mix adjustment [86,87], and an examination of four different specifications of hospital-wide mortality rates [15].

The advent of publicly reported performance measures has led to a proliferation of research using such publicly reported organisation-level data. This genre of research includes studies examining associations between different performance measures [88–90], or associations between organisational characteristics and performance [91], and those assessing the impact of pay-for-performance initiatives [92,93].

Many of the technical decisions involved in producing composite indicators relate to the processing and combining of pre-existing organisation-level performance measures. This also means that examining the impact of how composite indicators are constructed using organisation-level scores on individual performance measures is feasible and indeed has 'face validity'. Examples include studies exploring the construction of a potential new composite [49,79,94], and others that critique single aspects of the design of existing composite indicators [95,96].

Studies that both identify potential limitations and assess the consequences of these limitations are vital, but they are rare. Even rarer are studies that empirically examine the multiple technical aspects of composite indicators. They are needed because potential problems with a specific composite indicator may not necessarily undermine the meaning of the composite score – that is, some issues for some indicators may not have much practical impact. For example, the exact weights used to combine individual measures are far less influential if these measures are highly correlated [79]. Understanding the nature and importance of such technical issues is critical to improving practices relating to composite indicators. To address my aim of characterising challenges in the design of composite indicators, I carried out quantitative studies examining the sensitivity of apparent hospital performance to the technical specifications.

1.7.2 Understanding the development of composite indicators in healthcare

Beyond characterise relevant challenges in composite indicator design, my other broad aim was to explore how reporting of composite indicators could be improved. Most of the research examined in this chapter explores technical aspects of composite indicator design. This reflects aspects of the performance measurement literature that focus on technical over conceptual issues. There is little consensus about the best way to approach developing (individual or composite) performance measures, leading Shekelle to note in 2013 that:

> "[For clinical practice guidelines and systematic reviews] there are reports from the Institute of Medicine on "Finding What Works" and "Standard for Developing Trustworthy Guidelines", and there are assessment tools such as AMSTAR and AGREE II that stakeholders can use to assess the methods of the development. There is nothing comparable for quality indicators and performance measures, and it is desperately needed." [97]

Developing tools for aiding the reporting of a composite indicator requires a wider view than simply examining technical decisions. At base, it must be possible to understand why certain approaches have been favoured over other approaches. These are not technical questions, but are centred in how people conceptualise the composite indicator. Quantitative studies cannot fully address how people do this, however: there is a need for qualitative research into how people develop composite indicators, and the issues they consider when making development decisions. To address this aim, I carried out a qualitative interview study with experts in performance measurement.

1.7.3 Adding to our understanding of composite indicator design

While the literature contains valuable information about the design and reporting of composite indicators, many gaps remain. I have prioritised focusing on methodological transparency. As set out above, this includes both understanding the practical importance of decisions and understanding what those decisions are. The next chapter sets out the methods for two quantitative studies that examine the impact of methodological approaches, and one qualitative interview study that explores the decisions involved in developing a composite indicator.

2 Methods

The development of composite indicators of healthcare quality involves multiple technical steps. As set out in the previous chapter, the design choices governing these technical decisions are rarely clearly justified, and other potentially reasonable choices are often available but remain unevaluated. To advance the understanding of design choices and their consequences, I undertook two quantitative studies and one interview study. The two quantitative studies each take an existing composite indicator of healthcare quality and explore the impact on assessments of hospital performance if alternative technical choices had been made. The qualitative study then deepens understanding, interviewing experts in quality measurement in healthcare about the decisions involved in developing a composite indicator.

2.1 The rationale behind the quantitative studies

The two studies I report here are deliberately similar. Building on the previous critique of current problems with composite indicators (see Table 1), they both examine the impact of certain aspects technical decisions on composite indicators that combine more than 40 individual performance measures into a single overall quality rating. They primarily examine the same three technical decisions (the choice of measure weights, the grouping of measures, and the way measures are standardised to consistent scales, Table 2). The main difference is that the first study examines the CMS Star Ratings, while the second study examines the Sentinel Stroke National Audit Programme score and level.

Table 2 provides an overview of both these linked quantitative studies. It very briefly summarises the current approach that CMS and SSNAP use to produce their respective composite indicators of healthcare quality, and gives details of the alternative approaches that I examine when assessing how robust the performance ratings assigned under these schemes are to the precise methods used. There are a lot of similarities, in particular that for two of the technical decisions the current approach in the CMS Star Rating was applied as an alternative approach for the SSNAP score and level (and vice versa).

One reason for carrying out two quantitative studies using different composite indicators is that the importance of the different technical decisions is likely to depend on context. Choice of weights may matter more for some indicators than others, measure groupings chosen to suit stakeholders may in some settings closely approximate those from data-driven approach, and methods of standardisation may have far more impact in one setting than another. It is useful to remember that the importance – or lack of importance – of a technical decision on one indicator may not reflect its importance more generally, and repeating a similar analysis for two different composite indicators should help make this clear.

The second reason is that the technical approaches used in these two composite indicators partially mirror each other. The alternative approaches to weighting domains and to standardising measures I examine for the CMS Hospital Compare Overall Star Ratings are inspired by the current approach used for the Sentinel Stroke National Audit Programme Score and Level, and vice versa (Table 2). The use of factor analysis to identify apparent empirical domains of quality within a set of performance measures was inspired by a paper by Samuel and colleagues about possible composite indicators of cancer care quality [94], that does not relate to a current composite indicator. Hence, given the value of examining two different composite indicators, I chose to look at this pair of existing composites.

Table 2. Technical decisions examined in the quantitative studies, with brief details of the current and alternative specifications considered. For two of the technical decisions, the current specification of one of these indicators was very similar to the alternative specification considered for the other, supporting the reasonable nature of the alternative decisions.

Technical decision	**CMS Hospital Compare Overall Star Ratings**		**Sentinel Stroke National Audit Programme Score and Level**	
	Current approach	**Alternative approaches considered**	**Current approach**	**Alternative approaches considered**
Preliminary rounding of domain scores	Not applied	Not applicable	Applied	Not applied
Weighting of domains	Prioritising outcome domains	(a) Equal (b) Monte Carlo simulation based on wide log-normal distributions centred on current weights	Equal	(a) Prioritising acute domains of care (b) Monte Carlo simulation based on uniform distributions
Grouping of measures into domains	Domains chosen to align with existing reporting of measures	Domains based on factor analysis	Key indicator domains from the audit	Domains based on factor analysis
Standardisation of measures	Z-scores (-3 to +3 as Winsorized at +/-3)	0-100 scale produced using logistic transformations	0-100 scale using bands based on guidelines and expert opinion	Z-scores (approx. -3 to +3 scale)

2.1.1 Choice of composite indicators to examine

Both the CMS Star Ratings and the SSNAP indicators of stroke care quality meet three selection criteria I have chosen to employ.

1. Data are available to re-create the composite indicator under different technical specifications
2. The indicator combines a large number of different individual performance measures
3. The indicator is in current use

These criteria exclude many other composite indicators. For example, challenges accessing the underlying data limit examination of many US healthcare quality composites produced by commercial organisations. Examining the Star Ratings in parallel with the SSNAP indicators of stroke quality was also attractive because they represent US and UK examples of composite indicators of hospital quality. Another very attractive feature of combining these two indicators was the some of the plausible alternative approaches I considered for the CMS Star Ratings were similar to the approaches used by SSNAP currently, and vice versa.

2.1.2 Choice of technical decisions to examine

There are many technical decisions involved in producing a composite indicator, for example including selection of measures; the calculation of domain scores; the handling of missing data; the case-mix-adjustment of individual measures; and the approach to assigning any performance [16]. Most previous assessments of composite indicators look at only one technical choice, and I felt it was important to examine multiple approaches. Examining three technical decisions is a pragmatic balance between feasibility and my desire to look at multiple approaches.

The choice of three specific technical decisions I examine is informed by my prior consideration of the eight most important ‘problems with’ composite indicators [16], which I summarise in Chapter 1.

As alluded above, the handling of missing data is another important technical challenge for composite indicators. I do not examine the likely influence of person-level missing data because addressing this would require patient-level information, not available for the CMS

Star Ratings scheme or SSNAP composite indicators. However, the impact of missing data at the level of domains included in the calculation of CMS Star Ratings is described.

2.1.2.1 Factor analysis for identifying measure domains

The use of factor analysis to identify measure domains is motivated by the idea that measures grouped into the same domain should be those that capture similar aspects of quality. Factor analysis allows identification of measures that empirically appear to capture related aspects. Factor analysis should not replace clinically-informed theory of how different specific individual measures should be grouped together into domains. The choice of domains into which individual measures are grouped should remain a decision for the indicator developers. However, factor analysis can help to avoid assigning individual measures to domains to which they may not necessarily relate to. In that respect it can be considered as a data-driven complement to the decision-making regarding assignment of individual measures to given domains.

2.1.2.2 Approaches to standardisation

Standardisation of individual measures may be approached in several ways. Jacobs and colleagues describe nine different approaches to standardisation of individual measures in a composite indicator [49], with seven of these approaches being potentially applicable for the CMS Star Ratings and the SSNAP indicators. I use two of these 7 approaches, Z-scoring and the use of external standards to transform measures to 0-100 scales. When interpreting the results of my quantitative studies it will be important to remember that the approaches to standardisation I consider are only two of several others that could be applied. My findings should thus be treated as illustrating the principle that technical choices are likely to matter for the performance of hospitals measured using composite measures, and that further assessment of specific technical choices in composite indicator design (including of different approaches to standardisation of individual measures) will be valuable.

The impact of alternative technical approaches in the design of the CMS Hospital Compare Star Ratings scheme

The CMS Hospital Compare Star Ratings scheme is probably the highest-profile rating scheme for hospital quality in the US. Around 80% of US hospitals are assigned a Star Rating (3606 of 4586 hospitals according to the medicare.gov website [98]), with a rating of one star being the worst possible and five stars being the highest possible rating [2]. This is higher coverage than any other overall quality rating scheme in the US, with the CMS Star Ratings including nearly 1000 more hospitals than the 2680 rated in the IBM Watson Health 100 Top Hospitals Study 2020 [99], for example. The Star Ratings are intended to support patients in choosing the best hospital for their treatment, allowing the public to compare hospitals based on their overall star rating rather than needing to examine the wide range of individual measures that are also reported for each hospital on the medicare.gov website [46]. Hospitals that receive five stars often use this in their advertising (for example [100–102]). The Star Ratings and the continuous scores used to assign a rating are also used by researchers as a summary of hospital quality.

The Star Ratings are well-known and are in common use as a measure of the overall quality of a hospital. This includes the academic literature, where there are studies examining correlations between overall hospital quality taken from the CMS Star Ratings and specific process or outcome measures [40,103,104], that identify characteristics of high-performing hospitals [74,105,106], and which assess agreement between CMS Star Ratings and other hospital quality rating systems [107–110]. Hence, it might be assumed that the Star Ratings are a valid and robust measure of quality. Yet aspects of the Star Ratings appear questionable, as I discussed in Chapter 1, and as documented in critical articles in the academic literature [11,41,65,111], and associated coverage in specialist newspapers [9,13,112–114].

It is noteworthy that the documentation of the specific technical approach to producing the CMS Hospital Compare Star Ratings is excellent, with the technical approach described in full together with the provision of SAS code and an anonymised dataset to allow approximate reproduction of the Star Ratings [115]. Yet the reasons for making specific technical decisions are not fully explained. These specific technical decisions matter. The literature on composite indicators includes various examples showing that, for example, rankings of hospitals may be sensitive to:

- The weights used to combine measures and sub-domain scores [10,49,79,80,95]
- Whether and how measures are grouped into sub-domains [50,94]
- The functions used to map measures onto consistent scales [49]

Given the specific approaches used for each of these technical decisions are not clearly justified, one might expect that these technical choices are not important for the CMS Star Ratings. If these apparently unjustified technical choices can have a great impact on hospital rankings or assigned star ratings, then this is relevant when considering whether the ratings provide a valid measure of quality. But the impact on hospital performance of taking different approaches has not been evaluated. This quantitative study aims to examine the stability of hospital rankings and assigned star ratings under alternative technical specifications of the CMS Star Ratings.

2.1.3 The CMS Hospital Compare Star Ratings dataset

The 2020 CMS Hospital Compare Star Ratings are calculated from a set of 53 individual measures [2], grouped into seven domains of quality (Table 3). The exact number of measures used to create the Star Ratings varies from year-to-year due to changes in the measures included in CMS Hospital Compare [71]. Hospital Compare reports performance measures for 4,586 hospitals. But hospitals are only rated in the Star Ratings if they have domain scores for three or more of the seven domains of quality, and only receive domain scores if they report three or more performance measures within that domain.

For this analysis, I used the January 2020 SAS Input File for the CMS Hospital Compare Overall Star Ratings, which may be freely downloaded from the qualitynet.org website run by CMS [115]. This file is intended for use in understanding the technical methods used to create the Star Ratings. It contains data for 4,586 hospitals, including a pseudonymised identifier, a score (or missing value) for each of the 53 measures, and a denominator for each of the 53 measures where relevant. CMS noted that data may differ slightly from those reported on the Hospital Compare website, as some results in this dataset may have been suppressed due to possible data inaccuracies before reporting on Hospital Compare. This explained why in my analyses of this dataset I could assign Star Ratings to 3,726 hospitals, around 100 more than the 3,606 hospitals with Star Ratings reported on Hospital Compare.

2.1.3.1 Missing data in the CMS Hospital Compare Star Ratings dataset

One common problem for performance measures used in healthcare is missing data, and the dataset that the CMS Star Ratings are based on is no exception. For composite indicators, missing data can occur both at patient-level and at hospital-level. By patient-level missing data I mean cases where patients who should contribute to a performance measure for a hospital do not have information recorded. The January 2020 SAS Input File for the Star Ratings provides no information about the extent to which this happens for the different performance measures, and both my analysis and the current approach applied by CMS do not attempt any adjustment for or exploration of patient-level missing data.

It was possible to see the amount of hospital-level missing data on the different performance measures (Table 3). None of the 53 measures in the January 2020 SAS Input File had complete information for all hospitals, or even for all hospitals for which a Star Rating could be calculated. All hospitals had missing information for at least one performance measure, with hospitals on average having missing information on 14.7 measures. CMS gave two reasons for this missing data. First, that a hospital had not reported its performance on the measure, which was common in part because reporting requirements vary by hospital type in the US. Major teaching hospitals typically must report on all measures, with smaller hospitals and specialty hospitals having different reporting requirements [74]. Second, that CMS had suppressed the measure information from the public release dataset due to small numbers or some other reason, such as data inaccuracies.

Even hospitals that receive Star Ratings demonstrate very variable rates of missing data. Of the 3,726 hospitals that I could assign Star Ratings based on the current CMS approach, the vast majority (80%) of hospitals had no information about surgical site infections following hysterectomies , and a similar majority (78%) had no information about the percentage of patients receiving radiation therapy for bone metastases. Conversely, only 5% of hospitals had no data on the median time from arrival in the emergency department until discharge, and all hospitals reported a hospital-wide readmission rate.

Table 3. Domains of quality and constituent measures used in the CMS Hospital Compare Star Ratings, with the proportion of hospitals with missing data on each individual measure.

Domain	Measure name	Measure type	% missing (all 4,586 hospitals)	% missing (3,726 hospitals I can assign a star rating)
OUTCOME DOMAINS				
Mortality	30-day mortality from acute MI	Proportion	51%	39%
	30-day mortality from CABG	Proportion	78%	73%
	30-day mortality from COPD	Proportion	23%	8%
	30-day mortality from HF	Proportion	24%	9%
	30-day mortality from PN	Proportion	13%	3%
	30-day mortality from stroke	Proportion	45%	33%
	30-day mortality from surg. comp.	Proportion	61%	52%
Safety	Complications from hip and knee surgery	Proportion	40%	29%
	CLABSI	Rate ratio	57%	48%
	CAUTI	Rate ratio	51%	40%
	SSI Colon	Rate ratio	60%	50%
	SSI Hysterectomy	Rate ratio	84%	80%
	MRSA	Rate ratio	63%	54%
	C. diff.	Rate ratio	33%	19%
	PSI-90	Rate	30%	20%
Readmission	Excess days in acute care for acute MI	Rate	55%	44%
	Excess days in acute care for HF	Rate	22%	7%
	Excess days in acute care for PN	Rate	13%	3%
	Hospital visits after OP colonoscopy	Rate	35%	22%
	Readmission following CABG	Proportion	78%	73%
	Readmission following COPD	Proportion	22%	8%
	Readmission following hip and knee surg.	Proportion	40%	29%
	Hospital-wide readmission	Proportion	5%	0%
Patient experience	Communication with nurses	Proportion	26%	12%
	Communication with doctors	Proportion	26%	12%
	Responsiveness of hospital staff	Proportion	26%	12%
	Communication about medicines	Proportion	26%	12%
	Discharge information	Proportion	26%	12%
	Care transition	Proportion	26%	12%
	Cleanliness of hosp. environment	Proportion	26%	12%
	Quietness of hosp. environment	Proportion	26%	12%
	Global hospital rating	Proportion	26%	12%
	Willingness to recommend hospital	Proportion	26%	12%
PROCESS DOMAINS				
Efficient use of medical imaging	MRI lumbar spine for lower back pain	Proportion	70%	64%
	Abdomen CT	Proportion	14%	4%
	Thorax CT	Proportion	21%	8%
	Cardiac imaging for preop. risk assessment	Proportion	55%	44%
	Brain and sinus CT	Proportion	34%	25%
Timeliness of care	ED - time arrival to departure	Time-to-event	13%	4%
	ED - time admit decision to departure	Time-to-event	13%	4%
	OP - time to specialist care with cardiac symptoms	Time-to-event	90%	88%

Domain	Measure name	Measure type	% missing (all 4,586 hospitals)	% missing (3,726 hospitals I can assign a star rating)
	OP - time to ECG with cardiac symptoms	Time-to-event	39%	33%
	ED - time arrival to discharge	Time-to-event	11%	5%
Effectiveness of care	Flu vaccinations	Proportion	9%	2%
	Staff flu vaccinations	Proportion	9%	4%
	% patients leaving ED unseen	Proportion	18%	9%
	Stroke - brain scan within 45 minutes	Proportion	65%	57%
	% receiving appropriate follow-up after colonoscopy	Proportion	38%	25%
	% Hx polyps receiving follow-up colonoscopy	Proportion	39%	26%
	% radiation therapy for bone metastases	Proportion	82%	78%
	% mothers delivering early unnecessarily	Proportion	46%	35%
	% receiving appropriate care for sepsis	Proportion	33%	18%
	% blood clot while no prevention	Proportion	73%	67%

2.1.4 Current technical specification of the CMS Hospital Compare Star Ratings

The technical specification of the CMS composite indicator is described in detail in their published technical methodology [2]. This section summarises current technical specification of the CMS Star Ratings.

CMS select 53 quality measures, which they group into seven domains of quality (Table 3). These seven domains range in size from five quality measures, in the Timeliness of care domain, to 11, the Effectiveness of care domain. Each individual measure is mapped from its native scale (e.g. Percentages, rates or time-to-event) onto a common scale by Z-scoring using a normal approximation (i.e. Standardised measures give the number of standard deviations a hospital is from the overall mean [83], see Section 2.3.1 for more details). Standardised scores (i.e. z-score values) greater than 3 are rounded down to 3, and any less than -3 are rounded up to -3, a process known as Winsorization [116]. Applying Winsorization limits the influence any extreme scores may have on the overall performance of a hospitals. A score of +3 is the best possible and -3 the worst possible on every performance measure.

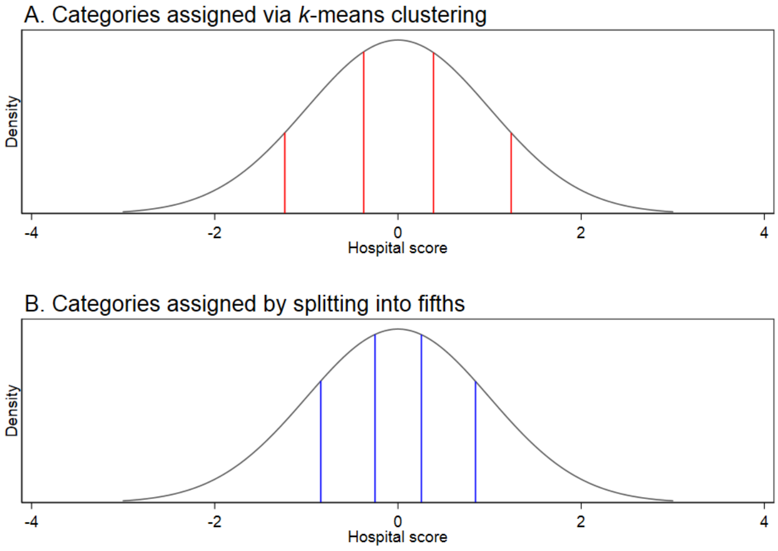
To estimate scores for each of the seven domains of quality CMS use latent variable models based on their constituent measures (see Section 2.3.2 for a brief introduction to latent variable models). Latent variable models are a ‘reflective’ approach to developing a composite indicator: the calculation of the composite domain score from the individual measures in the domain reflects structures of the data rather than a subjective valuation of different performance measures [117].

The overall summary score is calculated by taking a weighted average of the various domain scores. Outcome domains are given a higher weight than process domains (Table 3). Hospitals receive a summary score if they report at least three of the seven domains of quality, with at least one being an ‘outcome’ domain.

Finally, CMS assign the overall star rating by applying k-means clustering to the overall summary scores (see Section 2.3.3), with $k = 5$. This splits hospitals into five groups such that each hospital is closer to the mean of its group than to the mean of any other group. The group with the highest score is classed as the ‘five star’ group, the next highest score is the ‘four star’ group, and so on.

CMS prefer k-means clustering to other approaches because it provides a naturally defensible categorisation. With other approaches, such as splitting hospitals into five equal-sized groups, then there may be little distinction between performance categories. For example, if hospital performance followed a perfect normal distribution with mean 0 and standard deviation 1, then splitting into equal fifths leads to the middle performance groups covering only a narrow band of hospital scores (Figure 1), while the top and bottom categories cover a very wide range of performances. Using k-means mitigates this problem while still guaranteeing five distinct performance categories.

Figure 1. Illustrative comparison of categories assigned via k-means clustering and by splitting into fifths for a notional performance measure that follows a normal distribution with mean 0 and standard deviation 1.



2.1.5 Plausible alternative technical specifications for the CMS Star Ratings

My analysis used the current CMS approach to producing the CMS Star Ratings as the base case. I wrote Stata programs that implemented each of the steps used to turn the individual measures into the overall Star Ratings. I matched the CMS approach as closely as possible, but found that the latent variable model did not converge for the Timeliness of care domain when using the January 2020 public release dataset; so, for the score for this domain in my reproduction of the current Star Ratings I used the mean of the included measures. The failure to converge for me may be due to differences in software. I used Stata v15, while CMS use SAS.

There does not appear to be an explicit theoretical justification for the technical approach that CMS use to produce the Star Ratings, and the public-facing technical documentation of the indicator does not set out the way in which the technical approach was chosen. I examined the sensitivity of hospital rankings to the specific choices made when a) assigning weights used to combine domain-specific scores into the overall summary score, b) grouping individual measures into domains, and c) mapping individual measures from their native scale onto a consistent scale. I initially compared hospital performance under four technical specifications: my implementation of the current CMS approach, and three alternative specifications that differed only in one of the three choices outlined above (explained in detail below). I additionally examined the impact of making multiple different choices simultaneously, and carried out a simulation study to assess the impact of using a wider plausible range of domain weights.

The brief overview of the current CMS technical specification and the plausible alternatives that I considered is presented in Table 2 on page 30, and may be a helpful guide for the remainder of this section.

2.1.5.1 Technical choice 1: Weights used to combine domain scores

The current CMS Hospital Compare approach combines seven domain scores using a weighted average. The four ‘outcome’ domains each receive a weight of 0.22 (i.e. totalling to a four-outcome domain weight of 0.88), and the three ‘process’ measures each receive a weight of 0.04 (i.e. totalling to a three-process domain weight of 0.12). This weighting convention evidently gives ‘outcome’ domains more weight than ‘process’ ones, reflecting stakeholder preferences, and is aligned with CMS quality initiatives, especially with the Hospital Value-Based Purchasing program [118].

As a plausible alternative approach, I used equal weighting across the seven domains, a commonly approach when calculating domain scores for composite indicators [16].

As explained later on, there was one fewer domain (i.e. six as opposed to seven in the current specification of the Star Ratings when exploratory factor analysis was used to empirically identify quality domains (see Section 2.1.5.2)). As the domain dropped by the empirical identification analysis was a ‘process’ domain, the weights assigned to the two empirically derived process domains when aiming to match the current Star Ratings approach were increased from 0.04 (× 3) to 0.06 (× 2) so that the ratio of weights between outcome and process domains remained the same (0.88 / 0.12).

2.1.5.2 Technical choice 2: Grouping of measures into domains

The current CMS approach assigns individual measures to seven domains: four outcome domains (mortality, safety of care, readmission, and patient experience) and three process domains (effectiveness of care, timeliness of care, and efficient use of medical imaging) – see also Table 3. These align with the domains used in the CMS Hospital Value-Based Purchasing program and other national quality initiatives.

As a plausible alternative approach, I used the hospital-level data on the individual measures to generate the quality domains, which were then combined into the overall composite score. I applied exploratory factor analysis to identify empirical latent factors that explained most of the variance among the individual measures [119].

Missing data presented a challenge to this alternative approach. Standard approaches to exploratory factor analysis require complete data on all measures for all hospitals, but as alluded earlier most hospitals in the January 2020 public release for the CMS Star Ratings had some missing performance measure information (Table 3). That is, for most hospitals there was at least one, and usually several, performance measures for which they had no performance information at all. Multiple imputation offered one possible way of handling this [120,121], but in a factor analysis context maximum likelihood methods may be a better solution [120], and a maximum likelihood approach was straightforward to implement in Stata [122]. I used the expectation-maximisation algorithm to estimate the covariance matrix between all measures [123], giving correct results if measures are missing-at-random [121].

The number of factors to retain was decided after inspecting scree plots and considering eigenvalues [124]. While no strict criteria were applied, enough factors were retained to reach the ‘elbow’ of the scree plot, with this being interpreted generously so that most factors with eigenvalues above 1 were retained.

The promax rotation was applied to estimate how strongly each measure was associated with the underlying factors [125]. This process generated empirically coherent domains into which individual measures could be assigned. On the occasions where an individual measure was found to have a similar degree of association with more than one empirically-derived domain I assigned the measure to the domain where it appeared more conceptually relevant.

2.1.5.3 Technical choice 3: Standardisation to consistent scales

When standardising measures, the current Star Ratings approach treats all individual measures that feed into each domain similarly [2], regardless of their nature, i.e. with regard to whether they constitute rates, proportions or time-to-event measures. The approach CMS currently use, known as ‘Z-scoring’, measures how different performance at a particular hospital is from the average relative to the differences observed across all hospitals in a particular year [83]. If scores are normally distributed, then around 95% of hospitals will fall between ±1.96 standard deviations and 99.8% of hospitals between ±3.

There are many approaches to standardisation and Z-scoring has some limitations. For example, Jacobs, Smith and Goddard describe nine different standardisation methods used in producing composite indicators [49]. These range from simply using the raw data (i.e. not standardising at all) to identifying explicit performance thresholds to achieve specific scores. They argue that:

> “There is no need for any transformation if it is possible to specify a weight that indicates the relative value to the composite of an extra unit of attainment in that dimension *at all levels of attainment*. Otherwise a transformation is required. The objective is to make the transformed variable such that an extra unit of attainment is of equal value at all levels of attainment.” ([49], page 36, emphasis in original)

Standardisation of individual measures for combination into composite indicators is not just useful in order to get measures onto a common scale, but to also make it so that differences of the same size on the standardised scale have the same importance (up to a constant factor that can be addressed via weighting). This is because:

> “the process of the transformation of indicators is linked to the interpretation of the weights attached to the indicators and is therefore crucial in terms of the incentives which may be generated by the implicit weights.” ([49], page 41)

In terms of the incentives provided to rated organisations, if performance is measured on a 1-100 scale it is desirable for a change in score from 2 to 1 to be of the same importance as a change in score from 62 to 61. Otherwise there is an incentive for organisations to focus improvement efforts in certain areas, and these areas may not be those that are most important.

The Star Ratings include two different types of measures that might plausibly be treated differently. Of the 53 individual measures (see Table 3), 48 are proportion and rate-based measures such as death or healthcare-acquired conditions. A further five measures represent time intervals, such as time in the emergency department from arrival to departure.

For the 48 proportion and rate-based measures, one could deem differences between higher and lower values to be equally meaningful whether the lower value is near zero or approaching 100% (or, for rate-based measures,100 events per 100 units of person-time). For example, a one-percentage-point reduction in post-operative mortality equates to one fewer death per 100 operations, whether it relates to a reduction from 2% to 1% or from 62% to 61%. But with Z-scoring a one percentage-point change will appear far more important for a performance measure with standard deviation of one percentage-point than for a performance measure with a standard deviation of three percentage-points. Hence, for these measures, Z-scoring distorts comparisons between different performance measures.

As a plausible alternative approach that avoids this distortion, I standardised all event (e.g. mortality) or rate (e.g. of healthcare acquired infection) measures according to the proportion of people having an event, so that differences between levels of do represented deaths avoided or safety events that did not happen. For example, a mortality rate of 0.18 would equate to a standardised score of 100*(1-0.18) = 82, while a mortality rate of 0.73 would equate to a score of 100*(1-0.73) = 27.

For the five of the 53 individual measures in the CMS Star Ratings representing time intervals, such as time in the emergency department from arrival to departure, a difference of one hour might be much more consequential if it is between, say, two and three hours than if it is between 15 and 16 hours. Z-scoring is not able to reflect the challenge that the impact of an additional hour of expected waiting time depends on how long the expected waiting time is. Hence, for these measures, Z-scoring risks distorting comparisons between different levels of performance on the same performance measure.

As a plausible alternative that reflected the variable importance of an additional period of waiting, I mapped time interval measures to the 0-100 scale using an appropriate logistic transformation: differences between middle-of-the-range performances were treated as more important than differences between excellent performances, or between poor performances. The choice of transformation depended on the measure, but was motivated by the idea that differences between groups of hospitals with either excellent (or very poor) levels of performance were not important, so they should be small on the standardised scale.

For example, for the time-to-event measure ‘ED – time arrival to departure’, my alternative approach to standardisation was motivated by the idea that hospitals with median intervals of two hours or less had excellent performance (independently of whether one hospital has a mean of 60’ and another of 90’ minutes), and therefore all deserved a top score. Conversely, hospitals with median scores of ten hours or more had very poor performance deserving of a minimal score (again, independently of between-hospital differences within hospitals comprising the 10+ hours group).

With these considerations in mind, I used the following function to standardise the time-to-event measure ‘ED – time arrival to departure’. Let d be the standardised score for this measure and t time from arrival in ED to departure in minutes. Then the standardised score d is calculated as:

$$d = \begin{cases} 100 \text{ if } t < 120 \text{ minutes} \\ 100 \times \left(1 + \operatorname{expit}(-5) - \operatorname{expit}\left(-5 + \dfrac{t-120}{48}\right)\right) \text{ if } t \geq \ 120 \text{ minutes} \end{cases}$$

This equation looks complex, but in practice it simply describes a smooth curve that falls slowly up to around four hours, then drops off rapidly until it levels off at around ten hours (Figure 2A). There were 4005 hospitals with a known score for this performance measure, with mean performance 272 minutes (interquartile range 209 to 315 minutes). Mean standardised performance was 79 (interquartile range 73 to 97).

Standardised scores for the other four time-to-event measures were calculated as follows.

ED – time admit decision to departure (Figure 2B). There were 3988 hospitals with a known score for this performance measure, with mean performance 101 minutes (interquartile range 55 to 132 minutes). Mean standardised performance was 78 (interquartile range 69 to 98).

$$d = \begin{cases} 100 \text{ if } t < 30 \text{ minutes} \\ 100 \times \left(1 + \operatorname{expit}(-5) - \operatorname{expit}\left(-5 + \frac{t-30}{24}\right)\right) \text{ if } t \geq \ 30 \text{ minutes} \end{cases}$$

OP – time to specialist care (Figure 2C). There were 460 hospitals with a known score for this performance measure, with mean performance 63 days (interquartile range 42 to 92 days). Mean standardised performance was 78 (interquartile range 74 to 96).

$$d = \begin{cases} 100 \text{ if } t < 14 \text{ days} \\ 100 \times \left(1 + \operatorname{expit}(-5) - \operatorname{expit}\left(-5 + \frac{t-14}{14}\right)\right) \text{ if } t \geq \ 14 \text{ days} \end{cases}$$

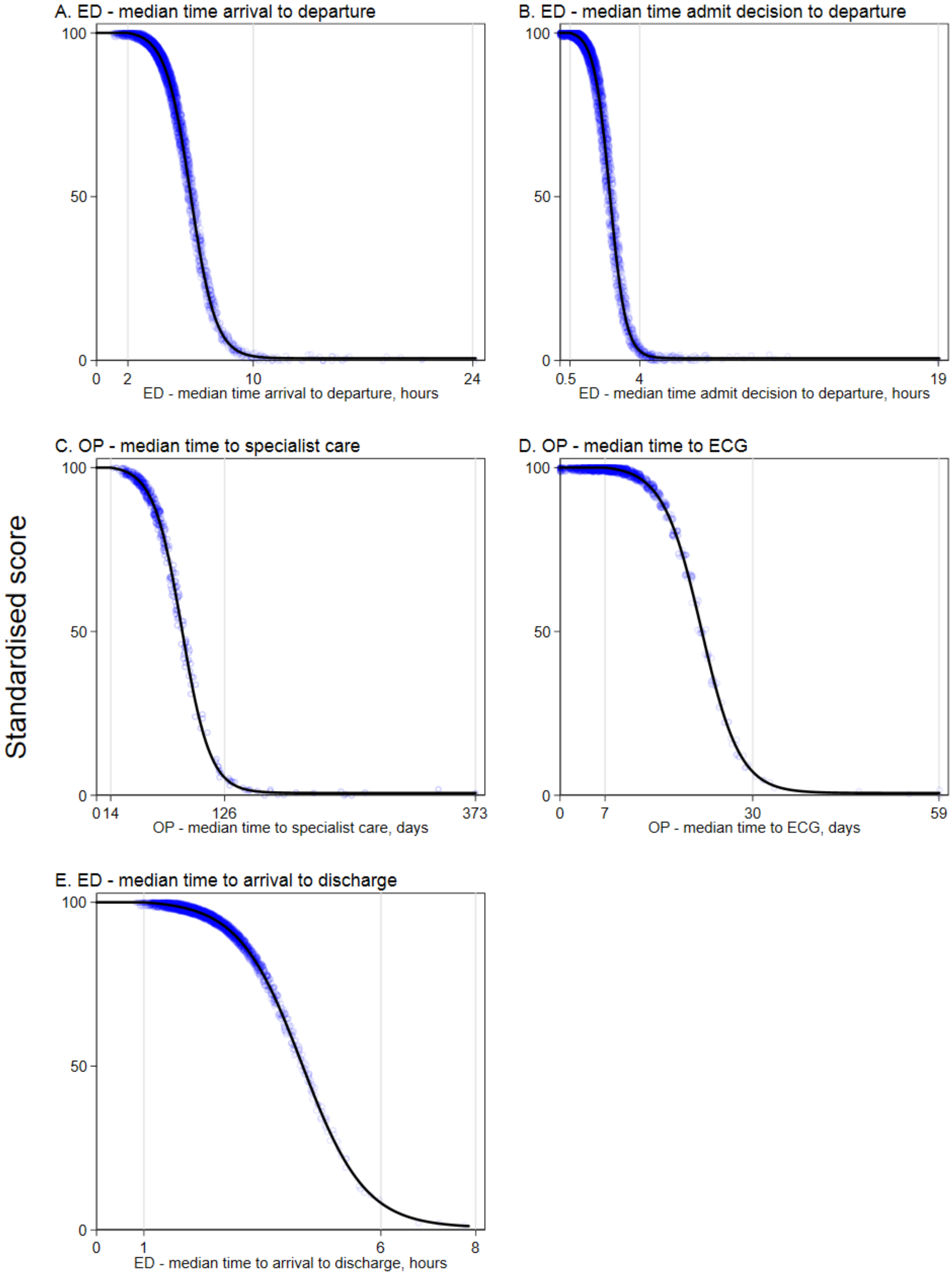
OP – time to ECG (Figure 2D). There were 2814 hospitals with a known score for this performance measure, with mean performance 8 days (interquartile range 5 to 14 days). Mean standardised performance was 97 (interquartile range 99 to 100).

$$d = \begin{cases} 100 \text{ if } t < 7 \text{ days} \\ 100 \times \left(1 + \operatorname{expit}(-5) - \operatorname{expit}\left(-5 + \frac{t-7}{3}\right)\right) \text{ if } t \geq \ 7 \text{ days} \end{cases}$$

ED – time arrival to discharge (Figure 2E). There were 4067 hospitals with a known score for this performance measure, with mean performance 140 minutes (interquartile range 110 to 164 minutes). Mean standardised performance was 93 (interquartile range 92 to 98).

$$d = \begin{cases} 100 \text{ if } t < 60 \text{ minutes} \\ 100 \times \left(1 + \operatorname{expit}(-5) - \operatorname{expit}\left(-5 + \frac{t-60}{40}\right)\right) \text{ if } t \geq \ 60 \text{ minutes} \end{cases}$$

Figure 2. Comparison of observed and standardised scores under the plausible alternative standardisation approach for the five time-to-event measures used in the CMS Hospital Compare Star Ratings. The black line shows the standardisation function, and the blue circles show observed scores for individual hospitals.



2.1.5.4 Changing multiple decisions at the same time

In addition to examining the impact of changing each decision individually, I assessed the impact of changing two or all three of the decisions simultaneously. In total, this meant eight different summaries of performance were produced for US hospitals based on the same underlying data. One was my reimplementation of the current CMS approach to producing the Star Ratings, while the other seven differed in one, two or three technical aspects.

2.1.6 Measuring impact of technical choices

I used three different approaches to understand the impact that changing a technical decision has on the CMS Hospital Compare Star Ratings. The idea was to show how taking a different appropriate technical approach, one that was at least as appropriate as the approach currently used in producing the star ratings, impacted the ranking and star rating received by each individual hospital.

2.1.6.1 Exploratory data analysis / visualisation of perturbance of hospital ranks

The first analytical step was exploratory, comprising data visualisation. I visualised the impact of the technical choices by drawing scatter plots of hospital ranks. I compared the rank under each of the plausible alternative approaches with the rank under my reimplementation of the current CMS approach to producing the Star Ratings.

2.1.6.2 Kendall’s Tau rank-based correlation coefficient

Second, I used a rank-based correlation coefficient as a summary measure of the agreement between different technical specifications of the composite indicator. In the literature, both Spearman’s rank correlation coefficient and Kendall’s Tau have been used for similar purposes [107,126]. I used Kendall’s Tau because of its clearer real-world meaning in terms of the number of concordant and discordant pairwise comparisons between two different rankings conventions.

2.1.6.3 Changes in assigned Star Ratings

The third measure was based on ‘major’ changes in the assigned Star Ratings. This was intended to show the practical importance of the changes in performance visible in the exploratory data analysis and summarised by Kendall’s Tau. It might be that substantial changes in hospital ranks tended to have little impact on the star ratings assigned to hospitals (i.e. a given hospital might have very different ranks under different technical specifications, but always receive the same star rating). To address this, I examined the number of hospital with a major change in star rating between my implementation of the current CMS approach and each of the alternative specifications.

In the context of my study, I considered a ‘major’ change in star rating to be the reclassification of a hospital being from four or five stars (out of five) to one or two stars (or vice versa). I viewed this as reclassifying a hospital being from ‘good’ to ‘bad’ (or ‘bad’ to ‘good’), and chose to look at such a large change in performance category because it was not clear that differences between neighbouring ratings (e.g. between four and five stars, or between two and three stars) were viewed as important.

2.1.7 Probabilistic sensitivity analysis via Monte Carlo simulation

The analysis described so far compared performance under my implementation of the current CMS approach with performance when specific alternative approaches were taken to between one and three technical decisions.

Yet considering only single specific alternative choices does not give a principled summary of the possible impact of alternative technical specifications on the performance of individual hospitals, because there may be many options that were not considered. There are many ways that 53 individual measures could be grouped into higher order domains and an uncountably infinite number of ways to standardise measures. Similarly, there are an uncountably infinite number of ways to weight domains.

The need to evaluate a range of possible technical choices has led some authors to propose the use of probabilistic sensitivity analysis, typically via Monte Carlo simulation, to assess the impact of design choices across different aspects of the specification of a composite indicator [1,49,50,95].

In the final part of the analysis, I used Monte Carlo simulation to assess the possible impact of a wide range of choices of domain weights both alone, and in combination with the two different options for grouping and standardisation considered described above. In total I performed four different Monte Carlo simulations:

1. Only the weights used to combine measure domains were varied, sampling weights from probability distributions centred on the policy-based weights used in the current star ratings.
2. Weights are varied as in 1., and in addition the way measures were grouped was chosen at random from the two options considered in section 2.1.5.2.
3. Weights are varied as in 1., and in addition the approach to standardisation was chosen at random from the two options considered in section 2.1.5.3.
4. Weights are varied as in 1., and in addition both the way measures were grouped and the approach to standardisation were chosen at random from the two options considered above.

Each Monte Carlo simulation used 10,000 draws, ensuring minimal Monte Carlo error.

2.1.7.1 Probabilistic sensitivity analysis of choice of weights

In the Monte Carlo simulations I focused primarily on the weights given to domains, similar to the few prior applications of probabilistic sensitivity analysis of composite indicators in healthcare [49,95]. In one such prior example, Proudlove and colleagues applied Monte Carlo simulation to assess the robustness of a composite indicator measuring quality of hospital maternity care based on 38 individual measures grouped into four domains [95]. They drew weights for each domain from uniform distributions, and summarised results using boxplots and by calculating the proportion of draws in which each hospital is in the best (worst) decile of scores. In another, Jacobs, Smith and Goddard compared several deterministic weight specifications in a composite indicator based on ten individual measures of quality [49], using Monte Carlo simulation to estimate the uncertainty in hospital scores under each specification.

Previous applications of probabilistic sensitivity analysis of weights used in the context of composite indicators weighting conventions have tended to either use uniform distributions or sample from a discrete set of weights [49,95,127]. Sampling from a discrete set of weight choices is quite a limited sensitivity analysis, while drawing from uniform distributions is only appropriate if there is no good prior reason to call one domain of quality more important than another domain of quality. The former is not the case for the Star Ratings as domain weights were chosen to be in line with existing quality initiatives, especially a pay-for-performance programme. In contrast to other composite indicator schemes, this means the currently used weights do have some justification and thus that conducting a sensitivity analysis using weights drawn from uniform distributions may not be ideal.

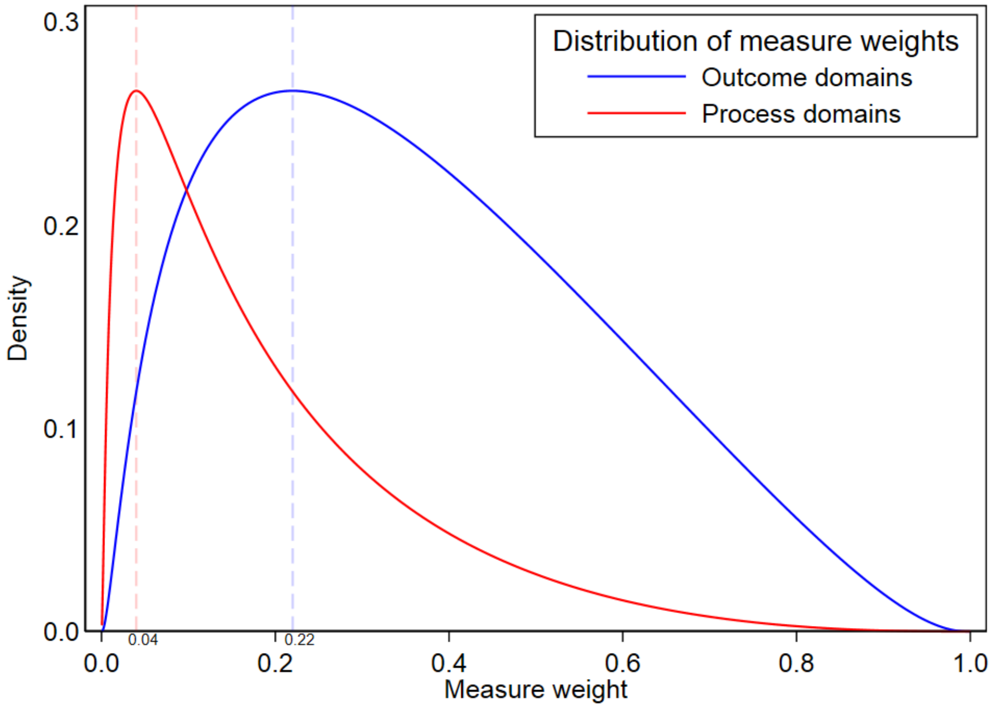
Therefore, for this sensitivity analysis, weights were drawn from appropriate log-normal distributions for which the mean value was the weight in current use. Let wt_{domain} be the weight given to domain of quality $domain$ by the CMS Hospital Compare Star Ratings. For outcome measures $wt_{domain} = 0.22$, for process measures using the current Star Ratings groups $wt_{domain} = 0.04$, and for process measures using the alternative grouping based on factor analysis $wt_{domain} = 0.06$. The weights used in the Monte Carlo simulations, $mc\_wt_{domain,sim}$, were drawn from the distribution

$$mc\_wt_{domain,sim} \sim \text{expit}\left(\text{normal}(\text{logit}(wt_{domain}), 1.5^2)\right)$$

This Monte Carlo simulation took the current weights as an appropriate starting point and then explored what happened if weights were varied around this starting point. To better appreciate this, it is worth considering the distribution on the ‘weight’ scale for each domain of quality in Figure 3. Drawing from these distributions meant that 95% of weights for current outcome domains (mortality, readmission, safety of care, and patient experience) were between 0.05 and 0.81, while 95% of weights for current process domains (timeliness of care, efficient use of medical imaging, and effectiveness of care) were between 0.01 and 0.61. Across many simulations the drawn weights on average matched those used in the existing Hospital Compare Star Ratings.

Weight choice was independent for each domain. While in the current CMS approach, all outcome domains receive a weight of 0.22, the simulation was not restricted to give all outcome (or all process) domains the same weight. In the current CMS approach, the total weight across all domains adds up to 1. In the simulation, weights were rescaled so that the total of the weights was 1. For example, if by some unlikely fluke the drawn weight for each of the seven domains was 0.1, so that the total weight across all domains was 0.7, the weights would be upscaled by dividing by 0.7 so that the weight given to each domain when combining them would be around 0.14. This kept the composite score on a consistent scale between different Monte Carlo draws.

Figure 3. Distribution of domain weights used in the Monte Carlo simulation when the measures are grouped into the domains used in the current CMS Hospital Compare Star Ratings. The mean weight for the four outcome domains is 0.22 and the mean weight for the three process domains is 0.04, chosen to match those in the current specification. But there is a very large amount of uncertainty around what the correct weight should be.



2.1.7.2 Probabilistic sensitivity analysis for grouping of individual quality measures into quality domains and for approach to their standardisation

Setting up an appropriate simulation to explore decisions other than weighting is challenging. When examining weights, there is only one uncertain element: the weight assigned to each domain. For standardisation, there are a far wider range of uncertain elements. For grouping of measures, there are points where human input may be desirable. For example, an individual measure may have similar factor loadings on two different factors, and in such cases it may be helpful to apply human judgement to ensure the measure is grouped with individual measures with which it appears conceptually linked. This is straightforward as a one-off event, but becomes prohibitively time-consuming for 10,000 different factor analyses based on 10,000 sets of measures standardised via randomly chosen approaches.

A probabilistic sensitivity analysis of standardisation of measures should consider not just the overall approach (e.g. Z-scoring or defining an explicit value function or some other approach), but also the specific versions of such approaches. For example, there are several Z-scoring techniques depending on the type of measure and the shape of the data [83], and one may wish to consider both whether different approaches meaningfully affect the results and whether the uncertainty in the population mean and standard deviation are important. If one is using a standardisation function based on explicit value judgments for different levels of performance, then any uncertainty in those value judgments should also be explored in sensitivity analysis.

Probabilistic sensitivity analysis of grouping of measures should address both the uncertainty of the overall approach (e.g. Stakeholder preferences or data-driven approaches) and its specific details. Efforts to elicit stakeholder preferences introduce uncertainty, and this should be included, and so should the possible impact of different groups as formal stakeholders or of using different approaches to eliciting preferences [128]. Factor analysis involves a range of decisions, from how many factors to retain to how to rotate the results to produce an interpretable set of factors and how to assign measures to a specific group based on the results. Often, these decisions involve judgment and may not be fully automatable to allow for a reasonable Monte Carlo simulation.

Given the above, it was not practical to design a Monte Carlo simulation that examined all the uncertainty involved in standardising measures or grouping measures into domains. But it was nonetheless possible to examine certain important aspects of this uncertainty via a more limited simulation. For example, a simulation could produce a composite indicator using Z-scoring half the time and using some explicit value functions the other half the time. This could, in principle, then be extended to capture a range of plausible options [127], and other technical decisions.

My probabilistic sensitivity analysis of grouping and standardisation aimed to demonstrate how these issues contributed additional uncertainty above that contributed by the choice of weights. As set out above, I carried out three further Monte Carlo simulations. One drew random domain weights and chose at random between two different approaches to grouping measures into domains. One drew random domain weights and chose at random between two different methods of standardisation. One drew random domain weights and chose at random between the two different approaches to grouping measures and the two different methods of standardisation.

2.1.7.3 Measuring the outcome of the sensitivity analysis

The main approach I used to quantify uncertainty in the sensitivity analyses was to estimate the typical range of performance for each hospital. By this, I mean the range of performance in which a hospital lies in half of the Monte Carlo draws, with the hospital falling below this range in one quarter of the draws and above this range in the other quarter. If this range of performance ran 1000 places for a given hospital, from 1700th place to 2700th out of around 3700 ranks, then in half the simulations it had performance inside this range and in half the simulations it lay outside this range – so a wider range meant that the hospitals' apparent performance was more sensitive to the precise technical approach used to produce the composite indicator. I contextualised this range with a plot showing the 25th to 75th percentile of ranks for each hospital, against the mean rank for that hospital.

Analyses with similar approaches have used 2.5th to 97.5th percentile ranges and minimum to maximum (i.e. 1-100) range [79,95]. I felt there was a risk that the 2.5th to 97.5th percentile range could be mistaken for a 95% confidence interval, and that using 2.5th – 97.5th ranges emphasised very rarely materialised scenarios (where performance of a given hospital was rated dramatically differently between the compared approaches). Therefore I viewed the choice of the 25th and 75th percentile as a measure of impact on performance ranks as more appropriate.

2.1.8 Availability of data and analysis code

All data and Stata code used to carry out this analysis are available online at https://bitbucket.org/mattebarclay/cms-star-ratings-methodology-final-code-and-data/src/master/ [129].

2.2 The impact of alternative technical approaches in the design of the Sentinel Stroke National Audit Programme Score and Level

The Sentinel Stroke National Audit Programme (SSNAP), run by the Royal College of Physicians as part of the suite of national clinical audits in the Healthcare Quality Improvement Partnership, aims to improve the quality of stroke care by making related timely information available to clinicians, commissioners, patients and the public [130]. Two composite indicators produced using data from SSNAP, the SSNAP score and the SSNAP level, are a key part of this. They summarise the performance of hospital stroke services on a set of 44 individual measures grouped into 10 domains to reflect different aspects of stroke care [8].

- The *Overall Score* ranges between 0 and 100, based on performance on each of the 10 domains with an adjustment for ascertainment and audit compliance. A score of 100 is the best possible score which – given the design of the composite indicator – would suggest that acute stroke care could not meaningfully be improved, and a score of 0 the worst possible score. In July-September 2019, achieved scores for routinely admitting teams ranged from 24 to 96, with a mean of 73 [131].
- The *Overall Level* assigns each hospital a letter grade ranging from 'A' for hospitals performing well across the board down to 'E' for poorly-performing hospitals, and is solely based on the Overall Score. Hospitals are expected to achieve an A or B rating, indicating care that is 'first class' (A) or 'good or excellent in many aspects' (B) [8]. A rating of 'C' indicates some areas of care require improvement, and ratings of 'D' and 'E' suggest several areas of care require significant improvement.

These composite indicators aim to provide a simple summary of national trends in quality of care over time in the SSNAP national report and allow hospitals to assess their overall performance compared with their peers [8]. The SSNAP Overall Level has historically been presented on the patient-facing website MyNHS, and updates are often accompanied by local press coverage of high-performing hospitals [132–134]. As SSNAP is a clinical audit, the

SSNAP score and level assess the processes of care that patients receive against particular standards. These include whether a patient receives clot-busting drugs and the therapy delivered for each patient in hospital and at home. In 2019, the audit captured information on the care of more than 90% of all patients who had a stroke [8].

The SSNAP score and level summarise the performance of hospital stroke services on a set of 44 individual measures grouped into 10 domains to reflect different aspects of stroke care [8]. The Overall Score ranges between 0 and 100, based on performance on each of the 10 domains with an adjustment for ascertainment and audit compliance. The Overall Level assigns each hospital a letter grade ranging from ‘A’ to ‘E’.

The two SSNAP composite indicators represent interesting examples of composite indicators for examination because they summarise a large collection of performance measures describing a great number of aspects of the quality of stroke care, and because they compare performance on each of these measures against benchmarks derived from clinical guidelines rather than applying purely statistical approaches to standardisation. This approach to standardisation inspired the alternative approach to standardisation I applied to the CMS Star Ratings, as described earlier. In contrast to the methods of standardisation used in many other composite indicators, this approach allows for average performance to change over time and it is possible for all hospitals to receive very good ratings. In the initial release of the SSNAP Overall Level, no hospital received an A grade. By 2018/19, 22% of hospitals were assigned an A [8].

I aimed to examine the sensitivity of hospital rankings on the SSNAP score and the rating hospitals are assigned by the SSNAP level to some of the decisions that were implicitly or explicitly made when defining the technical method. Similar to the analysis of the CMS Hospital Compare Star Ratings set out in Section 0, I assessed sensitivity to the functions used to map measures onto consistent scales, the way measures were grouped into sub-domains, and the weights used to combine sub-domain scores.

2.2.1 The SSNAP dataset

All data underlying the SSNAP level are made publicly available on the audit’s website [135]. These quarterly data releases contain all the information used to assign SSNAP levels for each hospital trust that routinely admits stroke patients in UK, as well as the number of patients who contributed data to the audit from each hospital.

There are two different versions of each performance measure for each hospital: the ‘patient-centred’ and the ‘team-centred’ measure [131]. Differences in SSNAP level based on the two different definitions are small. The patient-centred version includes information from all patients who received any stroke care from a hospital, while the team-centred version only includes that information if it was judged as relevant to the local team. The rationale is that this may “encourage an open dialogue between teams treating patients along a care pathway” [131]. For example, if a patient was admitted to an acute stroke unit at Addenbrooke’s Hospital and then transferred to the Norfolk and Norwich University Hospital (NNUH) for recovery care and discharge planning, then this patient’s full pathway would be included in the ‘patient-centred’ performance measures for both hospitals. So the patient-centred time-to-stroke-unit measures for NNUH would include this patient’s experience despite this part of the pathway happening at Addenbrooke’s, and the patient-centred discharge process measures for Addenbrooke’s would include this patient’s experience despite this part of the pathway happening at NNUH. For the team-centred measures, only the parts of the patient’s treatment and recovery pathway that a hospital had direct control over are included.

My analysis used the patient-centred version of the underlying performance measures from July to September 2019 [131], restricting only to the 135 hospitals identified as routinely admitting teams. For routinely-admitting teams, the differences between patient-centred and team-centred performance measures were relatively small, while hospitals that did not routinely-admit stroke patients tended not to have team-centred performance measures for acute aspects of care.

2.2.2 Current calculation of the SSNAP score and level

Technical documentation of the SSNAP score and level is spread across several reports, web-pages and spreadsheets [69,136]. There is no publicly accessible report (from the producers of SSNAP indicators) that provides the full methodological details. Although the actual technical method by which the SSNAP level is produced can be found in an appendix to the quarterly summary reports [69], the reasoning behind the technical method does not seem to be publicly available. I have therefore reviewed and synthesised the available information to produce the following understanding of the technical issues in determining the SSNAP level and score.

2.2.2.1 Selection, grouping, and standardisation of measures

The SSNAP score and level composite indicators are based on 44 measures chosen by the Royal College of Physicians Intercollegiate Stroke Working Party [136]. These 44 measures are grouped into 10 domains, with most domains containing a mix of proportion measure and time-to-event measures (Table 4). For example, the ‘Scanning’ domain includes the measures ‘Proportion of patients scanned within one hour of clock start’, ‘Proportion of patients scanned within 12 hours of clock start’, and ‘Median time between clock start and scan’.

Measures are standardised to a 0-100 scale using absolute thresholds [69]. The Intercollegiate Stroke Working Party, responsible for the design of the SSNAP score and level, produce clinical guidelines for stroke care [137]. The thresholds used in SSNAP were chosen to be in accordance with these clinical guidelines and the relevant quality standards for stroke care [138]. Although updated clinical guidelines and quality standards were released in 2016 [137,139], until now the technical design of the SSNAP composite indicators have not changed since the first release in 2013. Appendix 1, on page 238, gives details of the way each individual measure is standardised.

Table 4. SSNAP domains and a summary of the measures included within each domain.

Domain	**Summary of included measures**	**Number of proportion measures**	**Number of time-to-event measures**
Scanning	Three measures relating to time between clock start and scan	2	1
Availability of stroke unit	Two measures on time-to-admission and one on proportion of time spent in a stroke unit	2	1
Thrombolysis	Four measures on proportion of patients receiving thrombolysis and one on time until thrombolysis given	4	1
Specialist assessments	Four measures of whether certain specialist assessments happened and two on how long until specific assessments occurred	4	2
Occupational therapy	Four measures relating to receipt of occupational therapy	3	1
Physiotherapy	Four measures relating to receipt of physiotherapy	3	1
Speech and language therapy	Four measures relating to receipt of speech and language therapy	3	1
Multidisciplinary team working	Eight measures on receipt of specialist assessments and specialist therapies	5	3
Standards by discharge	Three measures on discharge screening and post-discharge planning	3	0
Discharge processes	Four measures on planning for support after discharge	4	0

2.2.2.2 Calculation of domain scores and domain levels

Measures within each domain are combined using equal weights, to produce a score for each of the ten domains. These scores are used to allocate domain levels based on fixed performance thresholds which differ for each domain (for example, a score of 95 is needed to receive an A on the Scanning domain, while a score of 80 will receive an A on the Thrombolysis domain). The scores required to achieve each SSNAP domain level are described in Table 5.[1]

Table 5. Domain score thresholds to achieve each SSNAP domain level in the current calculation of the SSNAP score and level.

Domain	**Score threshold**				
	A	**B**	**C**	**D**	**E**
Scanning	95	85	70	55	0
Stroke unit	90	80	70	60	0
Thrombolysis	80	70	60	45	0
Specialist assessments	90	80	75	65	0
Occupational therapy	80	75	65	60	0
Physiotherapy	85	75	70	60	0
Speech and language therapy	75	65	55	50	0
Multidisciplinary team working	85	80	75	65	0
Standards by discharge	95	80	70	55	0
Discharge processes	95	85	75	60	0

[1] This use of different thresholds across domains may suggest that reference points used to standardise the indicator are not correct (else the same thresholds could be used for all domains). For example, currently a score of 95 is required to receive an A in the scanning domain. But the standardisation of individual measures could be tweaked such that performance that would currently receive a 95 would instead receive a score of 90. Then, the same thresholds as used for the stroke unit domain could be applied to the scanning domain.

2.2.2.3 Calculation of the overall SSNAP score and level

The overall SSNAP score and level are based on the SSNAP domain levels, rather than the domain scores that lie behind the domain levels. Each domain on which a hospital receives an A counts as 100 points, B as 80, C as 60, D as 40 and E as 20.

The overall SSNAP score is the total number of points received divided by 10, as there are ten domains. In the audit, this score is then adjusted to reduce the points awarded to hospitals with low case ascertainment at the hospital level. These adjustments affected 41 of the 136 routinely-admitting hospitals.

I did not apply these adjustments, and so the SSNAP score presented in my analysis is an overestimate for some hospitals. It was reasonable for me to ignore these adjustments because they are applied after the calculation of the SSNAP score, and so are not affected by any other aspect of the composite indicator. They serve to incentivise participation in the audit, rather than providing an indication of the quality of stroke care.

The overall SSNAP level is based on the SSNAP score as follows:

- A: SSNAP score of 80 or higher.
- B: SSNAP score of 70 or more, but less than 80.
- C: SSNAP score of 60 or more, but less than 70.
- D: SSNAP score of 40 or more, but less than 60.
- E: SSNAP score of less than 40.

2.2.3 The technical choices being considered

The base case for my analysis was a reimplementation of the current calculation of the SSNAP score and level, which I programmed in Stata.

I examined the impact of an alternative technical approach at four different points. These decisions were the:

- Preliminary rounding of domain scores
- Weights used to combine domains
- Approach to assigning individual measures to domains
- Approach to standardisation of individual measures

The current approach to each of these positions is conceptually and/or technically problematic as set out below. The alternative technical approach examined addressed these issues. As for the CMS methods chapter, the brief overview presented in Table 2 may be helpful guide for the details that follow.

2.2.3.1 Preliminary rounding of domain scores

Scores on individual domains in the SSNAP are assigned to bands of performance (Table 5), which are then used in calculation of the overall SSNAP score and level. This is a type of preliminary rounding, which is not a recommended step in calculation of composite measures [49]. The documentation of the SSNAP does not explain why this step is applied, and it is a problematic approach for two reasons. First, due to the use of different thresholds to achieve the same performance band in different domains, it is effectively a re-standardisation of a set of scores that are already standardised. If this is necessary, it is best addressed by changing the standardisation of the individual measures rather than applying a second round of standardisation. Second, the use of bands means that performance of hospital units may be sensitive to the exact location of relatively arbitrary thresholds [10,49], and the location of these thresholds distorts incentives for quality improvement at individual hospitals [49].

While some form of re-standardisation may be required, the use of bands is not. The plausible alternative to banding I considered was to assign scores by linear interpolation between the performance thresholds [16]. This preserved the understanding, encoded in the current thresholds, of the importance of performing at a certain level but removed the sudden change in apparent performance that occurred when a threshold was crossed. This should reduce sensitivity to the exact location of performance thresholds [16], and reduce the distortion of the implied incentive for quality improvement [49].

Let S_i be the score assigned for level i, and P_i be the performance required to achieve level i. Then the interpolated score d_j for hospital j with raw domain score p_j, $P_i \leq p_j < P_{i+1}$, was

$$d_j = S_i + (S_{i+1} - S_i) \times \frac{p_j - P_i}{P_{i+1} - P_i}$$

In practice, this meant that a hospital with a raw domain score exactly halfway between two thresholds, would be assigned an interpolated score exactly halfway between the score given for meeting the lower threshold and the score given for meeting the upper threshold. For example, referring to Table 5, a raw domain score of 95 is required to receive an A (100 points) and of 85 to receive a B (80 points). A hospital with a raw domain score of 90 is halfway between these thresholds, so would receive an interpolated score of halfway between 80 and 100: 90. Similarly, a hospital with a raw domain score of 87 would receive an interpolated score 2/10 of the way between 80 and 100: 84.

The interpolation approach I applied necessarily led to hospitals receiving a higher score in the interpolated approach than when using the bands. I viewed this as the interpolated approach appropriately recognising how close hospitals were to the threshold. But this is debatable, and the approach could be tweaked to address this, for example requiring hospitals to reach the middle of the performance band to receive the score that would otherwise be achieved by crossing the threshold.

2.2.3.2 Weights assigned to domains

Under the existing methodology, the scores for the different SSNAP domains are combined using equal weights to produce the overall score and grade. The rationale for using equal weights is not obvious. Choice of domain weights is a contentious issue for composite indicators [16,79,80,95]. Other composite indicators use weights selected for ‘importance’ in some way, for example by strength of association with patient harms [81], or to specifically overweight outcome domains compared with process domains [2].

My assessment of the impact of domain weights on the SSNAP grade had two components.

The first component compared the score in my reimplementation of current SSNAP approach, calculated using equal weights, with a single plausible alternative approach that gave more weight to ‘acute’ domains and less weight to ‘recovery’ domains. This alternative weighting scheme would be reasonable if policy makers prioritised these acute domains, but of course it would be equally reasonable to consider a set of alternative weights that prioritised recovery domains or some other subset of the domains. This was simply a specific example to help understand the impact of domain weights. This single comparison allowed the impact of the change of weights on the SSNAP scores and levels received by hospitals to be understood in detail, but was unreasonable in that many other alternative weighting approaches are equally appropriate.

The second component addressed this limitation, carrying out a Monte Carlo simulation to examine the impact of a wide range of domain weights. The Monte Carlo simulation used weights drawn from uniform distributions (in contrast with the sensitivity analysis of the CMS Hospital Compare Star Ratings where certain domains were typically given more weight), because there was no particular reason to prioritise one SSNAP domain over another [69]. These random weights were used to calculate SSNAP scores and levels; individual uniform distributions produced weights between zero and one, although domain weights were rescaled such that the mean of all the domain weights was one. By repeating a large number (10,000) times, a range of plausibly achievable performances are estimated for each hospital.

Apart from the use of uniform distributions for choice of weights, the Monte Carlo specification was similar to that used in analysis of the CMS Hospital Compare Star Ratings (described in Section 2.1.7.1).

2.2.3.3 Grouping of measures into domains

Measures in the SSNAP composite indicator are assigned to ‘key indicator’ domains, aiming to give an overview of different aspects of stroke care. These domains have a potential flaw. Consider the ‘Physiotherapy’ domain. Two of the measures in this domain are ‘Percentage of patients reported as requiring physiotherapy’ and ‘Median minutes per day receiving physiotherapy’. While these two measures are both about physiotherapy in some way, it is not obvious that they will be correlated. In fact, they may be more correlated with measures in other domains, such as ‘Percentage of patients reported as requiring speech and language therapy’.

If the measures within each domain relate to a single aspect of stroke care, then the measures within each domain should be correlated. Domains that measure multiple aspects of care are not ideal, as poor performance on one aspect may be masked by better performance on a different aspect. This makes the overall domain score difficult to interpret.

If different domains relate to distinct aspects of quality, then measures that are in different domains should not be strongly correlated. If a single aspect of quality is measured by multiple domains, then in effect that aspect of quality becomes far more important than other aspects. This may be fine, as it can be addressed by choice of domain weights, but it is something that should be done deliberately.

My plausible alternative to the current key indicator domains was to derive domains from exploratory factor analysis [125], the same alternative approach I suggested for the CMS Star Ratings above. Domains produced using factor analysis will contain correlated measures, and correlations between measures in different domains will generally be small. Exploratory factor analysis uses the observed correlations between the different performance measures to find a way of producing a smaller number of factor scores that explain most of the variation in the individual performance measures, producing results unique up to rotation. There are various options for rotating factor analytic results to produce more interpretable factors. I chose to apply the promax rotation [125], which is a standard form of rotation that allows resulting factors to be correlated. Following factor analysis, measures ere assigned to the domain on which they had the highest loading.

Using domains derived from exploratory factor analysis introduced a new challenge. The raw scores on domains in the SSNAP score and level are grouped into performance bands before being combined into the overall SSNAP score, and these performance bands are different for each domain. I identified plausible bands for the domains derived from factor analysis based on the domains that measures in each empirical domain are assigned in the current specification of the SSNAP score and level (Table 6). The impact of this choice was likely to be small, as the differences in score threshold between domains are typically small.

Table 6. Domain score thresholds required to receive each SSNAP domain level used with the empirically-derived domains. The domain name in brackets is the current SSNAP key indicator domain from which the score thresholds have been sourced..

Domain (domain from current specification using same thresholds)	**Score threshold**				
	A	**B**	**C**	**D**	**E**
Good care (MDT working)	85	80	75	65	0
Stroke unit (Stroke unit)	90	80	70	60	0
Rapid thrombolysis (Thrombolysis)	80	70	60	45	0
Receipt of therapy (Occ. therapy)	80	75	65	60	0
Identification of therapy need (Occ. therapy)	80	75	65	60	0
Receipt of speech and lang. therapy (SLT)	75	65	55	50	0
Time receiving therapy (Physiotherapy)	85	75	70	60	0

2.2.3.4 Standardisation to consistent scales

The SSNAP score and level uses banding to provide an absolute approach to standardisation [69]. Measures are standardised to a 0-100 scale based on fixed reference points adapted from clinical guidelines [138]. For example:

- The ‘Stroke unit’ domain measure ‘% patients directly admitted within 4 hours’ is standardised simply by leaving it as-is.
- The ‘Thrombolysis’ domain measure ‘% all stroke patients given thrombolysis’ is standardised by multiplying the observed percentage by five if it is less than 20% and by mapping percentages greater than 20% to 100 (i.e. 20% or above receives the maximum possible score).
- The ‘Physiotherapy’ domain measure ‘Median minutes per day receiving physiotherapy’ was standardised such that a median of 40 minutes or more received a score of 100, a median between 32 and 40 minutes received a score of 90, and so on.

For some domains, where the individual measure measured compliance against a target, it is possible to score over 100. While the approach to standardisation is based on clinical guidelines [138], the justification for the specific way measures are standardised is not clear.

In principle, absolute approaches to standardisation have the key advantage of allowing tracking of performance at an individual hospital or group of hospitals over time, as well as identifying the best hospitals in any given year. But the current implementation for the SSNAP score and level has limitations. First, as mentioned in Section 2.2.3.1, the use of bands is not ideal [10,49]. Second, the fixed reference points used for SSNAP may be out of date. There have been clear improvements in performance since the SSNAP score and level were first introduced in 2013, but the fixed reference points remain the same. While this allows measurement of trends over time, there may be important differences in performance between two hospitals that are both in the top performance category for a given performance measure.

My plausible alternative approach was to standardise measures based on the observed variation between hospitals, that is to use *Z-scoring* [83], reasonable under the assumption that the current performance thresholds had become outdated. Z-scoring is common in healthcare performance measurement, and is for example the current approach used in producing the CMS Star Ratings [2]. Scores standardised using Z-scoring relate to quantiles of the normal distribution, so a hospital at a level of performance on a specific measure expected to be better than 99.8% of other hospitals (if hospital performance followed a normal distribution, which may not be the case) would receive a score of approximately +3 for that measure, while a hospital with a level of performance expected to be worse than 60% of other hospitals would receive a score of -1.

Consider the measure ‘Median minutes per day receiving physiotherapy’. Mean performance across hospitals in the Audit in July to September 2019 was 35 minutes, with standard deviation of around 6 minutes (Table 7). At Addenbrooke’s Hospital, the median performance was 33 minutes per day.

- The current SSNAP approach standardises measures to a 0 to 100 scale. A median performance of 33 minutes per day is in the second-highest category, equating to a standardised score of 90 out of 100. Hence, under the current standardisation, Addenbrooke’s receives 90 – almost top marks.
- Z-scoring standardises measures to values typically between -3 and +3. A median performance of 33 minutes per day was roughly two minutes less than the average. As the standard deviation of the measure was six minutes, Addenbrooke’s Hospital would receive a Z-score of about $-\frac{2}{6} = -\frac{1}{3}$, or to be precise -0.36.

Typical performance on this measure was so high in July to September 2019 that a level of performance that received 90 out of a possible 100 under the current SSNAP standardisation was worse than the average. Z-scoring addressed this possible ceiling affect in the current standardisation, given that 14 of the 44 measures had mean standardised scores of over 90 out of 100 in July to September 2019 (Table 7). It also addressed the problem of standardising via banding.

Table 7. Mean (SD) performance on each measure included in the SSNAP composite indicator in July to September 2019, as observed, as standardised in the current SSNAP composite indicator, and under an alternative 'Z-score' standardisation approach.

Domain	Measure	Observed performance		Current SSNAP standardisation		Z-score standardisation	
		Mean	(SD)	Mean	(SD)	Mean	(SD)
Scanning	% patients scanned within 1 hour	55.7	(12.5)	95.1	(9.9)	0.0	(1.0)
	% patients scanned within 12 hours	95.6	(3.4)	95.6	(7.0)	0.0	(1.0)
	Median time until scanned	0.9	(0.4)	87.7	(13.1)	0.0	(1.0)
Stroke unit	% patients directly admitted within 4 hours	55.8	(16.4)	55.8	(16.4)	0.0	(1.0)
	Median time until arrival on stroke unit	3.9	(2.6)	63.6	(22.7)	0.0	(1.0)
	% patients spending at least 90% of stay on a stroke unit	84.4	(9.6)	84.4	(9.6)	0.0	(1.0)
Thrombolysis	% all stroke patients given thrombolysis	13.0	(5.6)	63.1	(23.5)	0.0	(1.0)
	% eligible patients given thrombolysis	91.7	(10.4)	91.7	(10.4)	0.0	(1.0)
	% patients thrombolysed within 1 hour	57.5	(22.5)	57.5	(22.5)	0.0	(1.0)
	% applicable patients admitted within 4 hrs AND get thrombolysis	55.7	(16.4)	55.7	(16.4)	0.0	(1.0)
	Median time until thrombolysis	1.0	(0.3)	67.2	(17.5)	0.0	(1.0)
Specialist assessments	% patients assessed by a stroke specialist within 24 hours	83.6	(12.2)	83.6	(12.2)	0.0	(1.0)
	Median time until assessed by a stroke specialist	10.0	(5.4)	71.7	(18.1)	0.0	(1.0)
	% patients assessed by a stroke nurse within 24 hours	91.5	(7.4)	91.5	(7.4)	0.0	(1.0)
	Median time until assessed by a stroke nurse	1.6	(1.9)	84.2	(15.6)	0.0	(1.0)
	% applicable patients given a swallow screen within 24 hours	75.8	(13.7)	75.8	(13.7)	0.0	(1.0)
	% applicable patients given a formal swallow assessment within 72 hours	89.2	(10.0)	89.2	(10.0)	0.0	(1.0)
Occupational therapy	% patients reported as requiring occupational therapy	83.9	(12.1)	83.9	(12.1)	0.0	(1.0)
	Median minutes per day receiving occupational therapy	40.5	(5.7)	94.3	(6.7)	0.0	(1.0)
	Median % days on which occupational therapy is received	66.5	(15.1)	66.5	(15.1)	0.0	(1.0)
	% compliance against therapy target for occupational therapy	88.7	(28.0)	88.7	(28.0)	0.0	(1.0)
Physiotherapy	% patients reported as requiring physiotherapy	83.9	(11.8)	83.9	(11.8)	0.0	(1.0)
	Median minutes per day receiving physiotherapy	35.3	(5.6)	88.0	(8.5)	0.0	(1.0)
	Median % days on which physiotherapy is received	74.8	(13.1)	74.8	(13.1)	0.0	(1.0)
	% compliance against therapy target for physiotherapy	80.8	(21.5)	80.8	(21.5)	0.0	(1.0)
Speech and language therapy	% patients reported as requiring speech therapy	50.7	(11.1)	50.7	(11.1)	0.0	(1.0)
	Median minutes per day receiving speech therapy	33.0	(5.7)	83.9	(9.6)	0.0	(1.0)

Domain	Measure	Observed performance		Current SSNAP standardisation		Z-score standardisation	
		Mean	(SD)	Mean	(SD)	Mean	(SD)
	Median % days on which speech therapy is received	55.1	(15.7)	55.1	(15.7)	0.0	(1.0)
	% compliance against therapy target for speech therapy	57.9	(25.5)	57.9	(25.5)	0.0	(1.0)
MDT working	% applicable patients assessed by occupational therapist within 72 hours	92.3	(11.9)	92.3	(11.9)	0.0	(1.0)
	Median time until assessed by occupational therapist	22.6	(6.2)	67.4	(10.2)	0.0	(1.0)
	% applicable patients assessed by a physiotherapist within 72 hours	95.5	(4.0)	95.5	(4.0)	0.0	(1.0)
	Median time until assessed by physiotherapist	21.2	(3.3)	69.9	(6.1)	0.0	(1.0)
	% applicable patients assessed by a speech therapist within 72 hours	90.1	(9.2)	90.1	(9.2)	0.0	(1.0)
	Median time until assessed by speech therapist	24.1	(5.1)	65.2	(9.0)	0.0	(1.0)
	% applicable patients with rehab goals agreed within 5 days	92.5	(9.7)	92.5	(9.7)	0.0	(1.0)
	% applicable patients assessed by all relevant specialists in a timely manner	63.0	(17.2)	63.0	(17.2)	0.0	(1.0)
Standards by discharge	% applicable patients screened for nutrition and seen by dietician by discharge	84.8	(19.0)	84.8	(19.0)	0.0	(1.0)
	% applicable patients with a continence plan drawn up within 3 weeks	93.9	(9.0)	93.9	(9.0)	0.0	(1.0)
	% applicable patients who have mood and cognition screening by discharge	92.3	(12.6)	92.3	(12.6)	0.0	(1.0)
Discharge processes	% applicable patients receiving a joint health and social care plan on discharge	94.2	(11.3)	94.2	(11.3)	0.0	(1.0)
	% patients treated by a stroke-skilled Early Supported Discharge team	36.5	(23.1)	72.7	(36.7)	0.0	(1.0)
	% applicable patients in atrial fibrillation discharged on anticoagulants	98.2	(4.3)	98.2	(4.3)	0.0	(1.0)
	% patients discharged alive who are given a named person to contact	97.2	(7.5)	97.2	(7.5)	0.0	(1.0)

2.2.4 Measuring impact of technical choices

I measured the impact that changing a technical decision to a plausible alternative had on the SSNAP hospital score and level in two ways, and measured the impact on the SSNAP level in one way.

My assessment of the impact on the SSNAP score was similar to the way I assessed impact on the score underlying CMS Hospital Compare Star Ratings. I first assessed the impact that changing a technical decision to a plausible alternative had on the SSNAP score graphically, comparing the SSNAP score (and rank on the SSNAP score) under the plausible alternative specification against that under my reimplementation of the current SSNAP approach. I then quantified the concordance between the two approaches using Kendall's Tau rank-based correlation coefficient. Kendall's Tau is a useful measure of correlation in a performance measurement setting because of its interpretation in terms of concordant and discordant rankings. For example a Kendall's Tau of 1 implies all pairwise comparisons of hospitals are the same in both measures, while one of 0.5 implies one quarter of comparisons have a different hospital on top. The results of these comparisons were comparable with my similar assessments of the impact of technical choices on the score underlying the CMS Hospital Compare Star Ratings.

I assessed the impact on the SSNAP level by tabulating the SSNAP level under the plausible alternative specification against the SSNAP level under my reimplementation of the current SSNAP approach. I considered summarising this using Cohen's Kappa or some similar measure [140], but with a five-category scale I found it more straightforward to directly describe reclassification between the different approaches. In contrast with results for the underlying scores, it would not make sense to compare the amount of reclassification of the SSNAP level with the amount of major reclassification in the CMS Hospital Compare Star Ratings, because the threshold for reclassification was different.

2.2.5 Availability of data and analysis code

All data and Stata code used to carry out this analysis are available online at https://bitbucket.org/mattebarclay/ssnap-methodology-final-code-and-data/src/master/ [141].

2.3 Additional technical notes for the quantitative studies

The calculation of composite indicators involves many technical steps. In my description of the methodology for these two quantitative studies I have deliberately chosen to give a high-level overview that assumes the reader is familiar with several statistical techniques, including Z-scores, latent variable models, and k-means clustering. This section gives a brief overview of these methods, aiming to support intuition rather than give a detailed mathematical or statistical understanding. Additionally, my analyses of technical decisions ignore the statistical uncertainty in the underlying measures, and I briefly justify why this is appropriate.

2.3.1 Standardisation using Z-scores

Standardisation can describe any process applied so that scores on different measures are comparable. Standardisation using Z-scores is common in the production of composite indicators of healthcare quality [49,83]. Z-scoring takes advantage of the fact that – thanks to the central limit theorem – hospital-level scores on performance measures usually approximate a normal distribution.[2] The Z-score for each hospital on each performance measure simply describes the number of standard deviations higher or lower a hospital's score is than the mean across all hospitals.

Z-scoring is mathematically simple. Let $\hat{p}_{i,h}$ be the observed score on performance measure P_i for hospital h.[3] Let $\bar{p}_i$ be the mean of the observed scores across all hospitals and $\hat{\sigma}_i$ be the

[2] Hospital-level scores on performance measures typically approximate a normal distribution, as these measures typically have large enough sample sizes that the central limit theorem applies. Exceptions do exist. For example, some proportion-based performance measures having such high performance on average that it is not reasonable to assume a normal-distribution, and these exceptions can be handled using appropriate methods [83]. For the CMS Hospital Compare Star Ratings a normal approximation is used for all measures.

[3] A note on mathematical notation. The hats, for example on $\hat{\sigma}_i$, are used to show these are estimates from a sample rather than a known fixed value. The bar on $\bar{p}_i$ is similarly to remind

standard deviation of the observed scores. Then the estimated Z-score on performance measure P_i for hospital h, $\hat{z}_{i,h}$, is

$$\hat{z}_{i,h} = \frac{\left(\hat{p}_{i,h} - \bar{p}_i\right)}{\hat{\sigma}_i}$$

Say we intend to construct a composite indicator from just two performance measures, for example ‘the percentage of patients who die within 30 days after admission for myocardial infarction’ and ‘the median time in hours spent in the emergency department from arrival until discharge’. A score of 10 (percent mortality after myocardial infarction admission) on one measure means something very different from a score of 10 (hours waiting on average) on the other measure. By Z-scoring, we may find that a hospital is one standard deviation better than average on ‘the percentage of patients who die within 30 days after admission for myocardial infarction’ and half a standard deviation worse than average on ‘the median time in hours spent in the emergency department from arrival until discharge’. The Z-scores have a consistent meaning, and can be combined to produce a composite indicator that can be interpreted in terms of a given hospital’s tendency to be better or worse than average across both performance measures.

The CMS Hospital Compare Star Ratings apply an additional step when calculating the Z-scored measures. Scores are Winsorised at ± 3. That is, any standardised performance measures with a Z-score of 3 or more are rounded down to 3, and any with a Z-score of -3 or less are rounded up to -3 [83]. This is intended to prevent a problem that may arise if any hospital has an extreme Z-score. If a hospital has a Z-score of, say, 100 on one measure, this measure will be very influential over the score on the final composite indicator. In practice, Winsorising at ± 3 is unlikely to have much influence as – so long as hospital-level scores on

this is just the sample mean rather than a known value. In principle, Z-scores could be calculated to show distance from a target value rather than the sample mean, or the standard deviation in a specific year or of a subset of hospitals could be used instead of those of the entire current sample.

performance measures are approximately normally distributed – only two of every thousand hospitals will be affected by the Winsorisation.

2.3.2 Latent variable modelling

Latent variable modelling is a technique with roots in psychometrics going back to Spearman's work on measuring general intelligence in 1904 [142]. The motivation is to use observable information to make inferences about some underlying but unobservable construct based on a set of observable measures that are theoretically linked to the underlying construct [143,144]. An example will help us unpack this.

Consider the psychological construct "anxiety". It is unreasonable to ask people how anxious they are and expect consistent responses from different people. But there may be theoretical reasons to expect anxiety to be realised in more or less concrete ways. One psychometric anxiety score asks people to rate the extent they agree with statements such as "I can sit at ease and feel relaxed" and "I get sudden feelings of panic" [145]. There will be individual-level differences but on average it is expected that people with higher anxiety will be less able to sit at ease and feel relaxed and will be more likely to get sudden feelings of panic. In fact, there might reasonably be a linear relationship between underlying anxiety and responses to statements such as "I get sudden feelings of panic".

Turning this into mathematical notation, there is an unobservable construct X ("anxiety") and a set of n observed measures $x_i, i = 1, \dots, n$, ("I get sudden feelings of panic", "I can sit at ease and feel relaxed") measured with error $\epsilon_i \sim N(0, \sigma_i^2)$. Let α_i and β_i be real constants. Then each of these observed measures has a theoretical link[4] to the construct X such that

$$x_i = \alpha_i + \beta_i X + \epsilon_i$$

[4] The link between the latent variable and each observed measure can in principle take any form (eg. It may be non-linear or involve correlated error terms), but in the CMS Hospital Compare Star Ratings the link is assumed to be simple and linear as in this example.

Ignoring the error terms ϵ_i, these are simple simultaneous equations and uniquely solvable with enough data points. The error terms make it impossible to solve exactly, but it remains possible to find the maximum likelihood solution [143,144].

There are two main challenges with latent variable modelling approaches.

1. Construct validity. If the underlying construct that the latent variable model is trying to find does not really exist, then our results are not going to be meaningful. This may be a particular challenge when applying latent variable modelling in healthcare performance measurement: Is hospital-wide "Safety of care" a valid construct?
2. Content validity. The underlying construct may be valid, but do the various measures used in the latent variable model truly have a strong theoretical link to the construct? Is it certain that the latent variable model is measuring the right underlying construct?
    a. Systematic bias is a specific realisation of this problem that is highly relevant for use of latent variable models in healthcare performance measurement. If performance on the individual measures is more about, say, socio-economic context than the quality of care itself, then the underlying construct identified by latent variable modelling might be 'socio-economic context' rather than 'quality of care'. Similar concern applies to other forms of systematic bias such as surveillance effects [76] or inadequate case-mix adjustment [146] – even regional differences in diagnostic practices [87].

The main advantage of latent variable modelling is the handling of missing data. If the theoretical link between the construct and the individual measures is correct and any missing individual measures for hospitals are missing at random, then the model will produce unbiased estimates of performance on the underlying construct even if some of the individual measures are not reported.

2.3.3 Assigning star ratings via k-means clustering

The CMS Hospital Compare Star Ratings use k-means clustering to assign the star rating. Hospitals are split into k clusters based on the overall score such that each hospital is assigned to the cluster with the nearest mean.

A simple algorithm is used to implement k-means clustering.[147] Let $\boldsymbol{x}_h$ be the vector of scores for hospital h, $h = 1, \ldots, n$. Let $S_i^{(t)}$ be the set of hospital-vectors $\boldsymbol{x}_h$ in cluster i, $i = 1, \ldots, k,$ at step t. Let $\boldsymbol{m}_i^{(t)}$ be the mean of hospitals-vectors in $S_i^{(t)}$. Choose starting means $\boldsymbol{m}_i^{(1)}$ from within the domain of the observed scores, for example by choosing k hospitals at random and using their scores as the starting means. The algorithm then proceeds by alternating between the "assignment step" and the "update step"

Initial assignment step. First, assign each hospital to an initial cluster by putting it in the cluster with the closest mean.

$$S_i^{(1)} = \left\{\boldsymbol{x}_h : \left\|\boldsymbol{x}_h - \boldsymbol{m}_i^{(1)}\right\| \leq \left\|\boldsymbol{x}_h - \boldsymbol{m}_j^{(1)}\right\| \;\; \forall j \neq i, 1 \leq j \leq k \right\}$$

The starting means $\boldsymbol{m}_i^{(1)}$ will not be the same as the mean of the different hospital-vectors $\boldsymbol{x}_h$ in the cluster $S_i^{(1)}$, except by chance.

Note: The double vertical lines denote the length of the vector, so $\left\|\boldsymbol{x}_h - \boldsymbol{m}_i^{(1)}\right\|$ is the Euclidean distance between point $\boldsymbol{x}_h$ and point $\boldsymbol{m}_i^{(1)}$.

Update step. Update the cluster means by calculating the mean of the hospital-vectors assigned to each cluster.

$$\boldsymbol{m}_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_h \in S_i^{(t)}} \boldsymbol{x}_h$$

Note: The single vertical lines denote the size of the set, so $\left|S_i^{(t)}\right|$ is the number of hospitals assigned to set $S_i^{(t)}$.

Assignment step. Assign each hospital to an updated cluster by putting it in the cluster with the closest mean.

$$S_i^{(t)} = \left\{\boldsymbol{x}_h : \left\|\boldsymbol{x}_h - \boldsymbol{m}_i^{(t)}\right\| \leq \left\|\boldsymbol{x}_h - \boldsymbol{m}_j^{(t)}\right\| \;\; \forall j \neq i, 1 \leq j \leq k \right\}$$

As at step 1, the means $\boldsymbol{m}_i^{(t)}$ will not necessarily be the same as the means of the hospital-vectors in set $S_i^{(t)}$ – and if so then we return to the update step. If the update step does not change anything (so $\boldsymbol{m}_i^{(t+1)} = \boldsymbol{m}_i^{(t)}$) then the algorithm has converged.

This algorithm will always give a set of k clusters (assuming the dataset includes at least k distinct points). There is no reason to believe these clusters are meaningful or stable over time.

2.3.4 Statistical uncertainty is not considered

The quantitative analyses aim to show the impact of the specific technical choices, not other important issues such as the statistical uncertainty of the individual performance measures. The individual component measures of the Star Ratings or SSNAP level are taken as known and fixed, and the analyses are purely descriptive. The impact of making specific changes to the technical specification of the indicator on the observed ranks and ratings are shown, rather than additionally accounting for the statistical uncertainty in the constituent measures of the composite indicator.

Previous work looking at the impact of methodological uncertainty in composite indicators also tends to ignore issues of statistical uncertainty. Some researchers take the philosophical view that, as many performance measures are based on all procedures carried out in an organisation, statistical uncertainty is not relevant [95]. I propose that it is helpful to view measured performance at an individual institution as a sample drawn from a wider super-population of performance in different years and at different institutions, and so the concept of statistical uncertainty remains important in general. Yet the issues I examine in these specific studies are unrelated to and unaffected by the statistical uncertainty of the component pieces of the composite indicator. Accounting for this uncertainty in the analysis would not add to the understanding of the importance of the methodological choices.

2.4 Understanding decisions in the development and reporting of composite indicators: interview study

Examination of composite indicators in current use reveals that lack of transparency and poor reporting are serious and common problems [16,65], and that conceptual and technical limitations are also persistent [11,14,62]. To an extent these problems may arise because of failures to recognise the choices, some of which may be implicit or tact, being made in the development of an indicator. Existing guides to composite indicator development focus almost exclusively on technical issues [1,18,49–51], but, to support better decisions, it is important to gain insight into conceptual concerns and approaches to reporting too. The design of composite indicators involves balancing concerns of multiple different stakeholders, ranging from highly technical statistical issues down to how indicators align with existing performance measurement initiatives. Capturing the perspectives of the different stakeholders in composite indicator design is thus a valuable way to identify the decisions that need to be made in developing composite indicators and thus provide a basis for improved reporting.

Several research approaches could be considered for investigating stakeholder perspectives. The most common in research linked to guideline development is perhaps a systematic review [148]. Quantitative surveys are also commonly used to gather views on important methodological issues (for example [149–151]), as are various Delphi-type studies [152–154]. Finally, there are a few examples of qualitative interview studies on perspectives on quantitative methodological questions [155–157]. I considered each of these possible approaches before deciding to undertake an interview study.

I determined that a systematic review of decisions in the development of composite indicators was not practical, not least because appropriately capturing methodological details from papers in many different fields would be very challenging. Only a few documents aim to give a full overview of the development process for composite indicators of healthcare quality, and these tend to have been written by single-profession groups [1,18,49–51]. The full technical details of the development of existing indicators do not generally appear in the academic literature, and may not be published at all. Even where methodological details are published, the literature discussed in chapter 1 showed that decisions are frequently not documented.

A quantitative survey exploring opinions on different aspects of the development of composite indicators was considered premature. To design such a quantitative survey it would be first necessary to understand what the different aspects of the development process are, so that questions about them can be asked. But, as set out above, the few existing documents describing the full development process had important limitations. A survey aiming to explore the importance and acceptability of different issues in composite indicator development, conceptually similar to Nikolakopoulou and colleagues' survey of the acceptability of different evidence synthesis techniques to support future trial design [149], would be both fascinating and useful. But to do that also requires an understanding of what the different steps in the process are, and that is still lacking.

Qualitative interviews offer a promising approach to researching complex methodological questions for which the published guidance is limited. Their application for this purpose is relatively novel, with the publication in recent years of a small handful of interview studies on quantitative methodological questions such as the design of adaptive clinical trials [155,156], or the development of core outcome sets [157]. These studies highlight the value of qualitative approaches in capturing a diverse range of perceptions and experiences, providing a robust exploration of the relevant issues and potentially setting the stage for future quantitative surveys or Delphi studies to examine the importance of the different themes from interviews.

I undertook a qualitative study with the aim of seeking to identify the range of choices that are made when developing and reporting composite indicators by interviewing people who are expert in the design of composite quality indicators, and quality measures more broadly.

2.4.1 Study setting

I conducted a telephone interview study with international participants with relevant expertise in composite indicators.

2.4.2 Study design

Semi-structured interviews were used to gather views on the decisions involved in developing and reporting composite indicators in healthcare. The sample size was planned to be between 12 and 20 experts identified through purposive and snowball sampling [158]. The final sample size of 14 was guided by considerations of information power [159], although plans for a 15th interview were thwarted by the coronavirus pandemic. Information power is the idea that, in a qualitative study, the necessary sample size depends on several issues including the sample specificity, the quality of dialogue, the aims, and the analysis strategy [159]. Participants in this study all had relevant expertise, the interviews covered many relevant details often from different perspectives, the aims were relatively limited and the analysis strategy simple – and so a small sample was appropriate. Interview transcripts were analysed using the Framework method [160], starting with a framework based on my initial views on composite indicator development as set out in chapter 1.6 and iteratively revised to produce the final framework.

2.4.3 Data collection and sampling

2.4.3.1 Telephone interviews

Data were collected via semi-structured telephone interviews using a flexible prompt guide (Appendix 2). The guide was developed based on the issues identified in examination of existing composite indicators, literature review, and multiple conversations with colleagues and supervisors, followed by some informal piloting. While it was planned that revisions to the prompt guide would be made throughout the study to accommodate new information, the initial prompt guide proved adequate for all interviews.

Interviews were digitally recorded using an encrypted recording device, and recordings were transcribed by 1st Class Secretarial Services under a confidentiality agreement. I also took notes during the interviews, primarily to inform follow-up questions, and wrote interview summaries immediately following each interview.

2.4.3.2 Interview participants

A key reason for using interviews rather than a systematic review was to ensure the study captured a wide range of views about the decisions involved in developing and reporting composite indicators. To this end, the sample was deliberately chosen to cover a wide range of domains, recognising that some participants have expertise and experience in multiple relevant areas.

As participants contributed their ideas and expertise, they were invited to have group authorship credit on any presentations or publications that will arise from this interview study.

2.4.3.3 Eligibility criteria

- English-speaking
- Able and willing to take part in a telephone interview
- Expert in the design or development of quality measures in healthcare
    - Expertise known to members of the study team
    - Expertise identified from their contributions to the published literature
    - Recommended as having relevant expertise by a different participant (snowballing)
- Professional expertise in at least one of:
    - clinical medicine
    - statistics
    - healthcare quality improvement
    - health services research
    - clinical epidemiology
    - analysis of routine data and production of routine indicators
    - hospital administration/manager
    - other relevant area identified as the study proceeds

2.4.3.4 Sampling

A mix of purposive and snowball sampling was used to recruit participants [158,161,162].

Initially, ten potential participants with diverse expertise were identified from the literature and were sent invitational emails by my supervisor Professor Dixon-Woods, as an approach by a senior academic was viewed as more likely to lead to positive responses than an email from a PhD student. The initial sample was supplemented using a 'snowball' approach [161,162], asking participants in the study if they know of other people likely to have a useful viewpoint on the reporting of composite quality indicators in healthcare. Participants were asked to email their suggestions directly to ask if they would be willing to take part, and provided a standard invitation email to be forwarded by participants to possible new candidates for interview.

Snowball sampling was particularly useful in recruiting participants from settings that were less research-active [161]. Identifying potential participants involved in hospital administration was more challenging than identifying potential participants who were health services researchers, for example, because administration professionals were less likely to have an active web presence.

2.4.4 Analysis of interview transcripts

Interview transcripts were analysed using the Framework Method [160]. This is a specific type of thematic analysis where qualitative data are summarised into a matrix. Each row represents a participant, each column a specific ‘code’, and each cell a summary of participant responses relevant to that code. The structured output gives an appealing summary of themes in the data and allows easy comparison across participants. NVivo 12 was used to structure the analysis and manage transcripts. NVivo was valuable because it allowed easy labelling of passages relevant to certain categories, and then made it straightforward to relabel categories and merge categories when this became necessary.

2.4.5 Applying the Framework Method

Initially, interview transcripts were reviewed and coded into a working analytical framework [160]. There were three key elements to this initial coding. First, the framework was deliberately very broad. Second, new categories were added to the framework as necessary, and revisions were made to ensure the categories continued to make sense. Third, there was no attempt at abstraction; when charted, the framework matrix displayed the raw text.

The initial framework was based on the study team's (I and my two supervisors) initial perceptions of the decisions in developing composite indicators. Once five interviews had been coded into this framework, the process of iteratively refining and revising the framework began, a process of looking for mismatches between the themes arising in the interview transcripts and the concepts encoded in the initial framework [163]. The aim of the first round of revisions was to ensure that the large 'bucket' categories initially added to the table contained broadly compatible ideas, potentially merging very similar categories and splitting out important ideas where single categories seemed to contain multiple different themes. Subsequent revisions continued until the framework no longer required updating.

2.5 Summary

My approach to examining the design of composite indicators in healthcare principally relies on my quantitative background. The range of aspects I considered in the two quantitative studies of the sensitivity of hospitals' apparent performance on existing composite indicators to reasonable alternative technical specifications was wider than in previous studies of the robustness of composite indicators. This detailed examination was possible because of my experience of working with healthcare performance measures when I planned my PhD research. The results of these quantitative studies, detailed in the following two chapters, empirically illustrate the degree of stability of the specific composite performance ratings. The qualitative study, in contrast, aimed to describe the decisions involved in developing composite indicators. The results of my qualitative study, set out in chapter five, help more broadly contextualise the results of the two quantitative studies.

3 Results 1: An analysis using secondary data of the impact of alternative technical approaches in the design of the CMS Hospital Compare Star Ratings

3.1 Summary

The CMS Hospital Compare Star Ratings are the highest profile summary rating of hospital quality in the US. However, they are sensitive to often implicit decisions about technical specifications that are rarely subject to scrutiny. In this chapter I examine and quantify the impact of different technical decisions when constructing the composite CMS Star Rating indicators.

The Hospital Compare January 2020 data release on 4,586 US hospitals was used for this analysis. Of these hospitals, 3,726 reported sufficient information to allow calculation of a star rating under the current CMS methodology, and were included in subsequent analysis. To undertake the work, I applied changes to the technical specification of the indicators, considering different approaches to grouping, weighting and standardising performance measures. The impact of these changes was first assessed visually, looking at changes in hospital ranks based on the summary score underlying the Star Ratings, and then quantified using Kendall’s Tau. Changes in assigned Star Rating category were additionally examined.

Swings in categories that could be classified as extreme (top two Star Ratings to bottom two Star Ratings; bottom/top) were observed in the ratings of one in nine (12%, 455/3726) US

hospitals when changing all three of the technical decisions considered. The Star Ratings of hospitals with one or more missing domain scores were more sensitive to technical specification changes, with an extreme swing in rating observed in one in five (20%, 115/586) of hospitals missing three or four of the seven domains of quality .

The findings empirically illustrate the substantial degree to which CMS Hospital Compare Star Ratings are sensitive to methodological choices about technical specifications. While reasonable alternative specifications can affect most hospital ranks, they particularly influence the ranks of hospitals that do not report all domains. Star Ratings for hospitals that do not report all measure domains should be considered to have limited validity.

3.2 Introduction

In this chapter, I examine how different aspects of the technical specifications of composite indicators can affect reported hospital performance in the context of a high-profile composite indicator, the CMS Hospital Compare Star Ratings. The Star Ratings are an excellent technical example of a composite indicator of healthcare quality because they summarise 53 individual performance measures covering multiple different aspects of quality, ranging from mortality rates to waiting times in the emergency department. Their intended use, their prominence, and their wider use in health services research means they are one of the most important composite indicators of overall hospital quality. The Star Ratings are intended to support decision-making about patients’ choices of hospitals for treatment, and for such use it is important that they present a robust and valid measure of quality. They are also often used in research studies exploring associations between hospital quality, as measured by the Star Ratings, and hospital characteristics or outcomes for specific conditions [40,74,103–106]. To the extent that these studies affect policy decisions on the optimal structure and administration of hospitals, it is critical that the strengths and limitations of the Star Ratings are well understood.

3.3 Summary of methods

The methods for this study are presented in detail in Section 0. Here I briefly reiterate the details of the Star Ratings and the approach I take in this analysis. Table 2 on page 21 provides an even briefer summary of the technical specifications compared in this chapter, while Table 8 summarises the potential drawbacks of the current specification, and the way my proposed alternative specifications addressed these perceived limitations.

3.3.1 Methods overview

The Hospital Compare Star Ratings are based on 53 individual measures across seven domains [2]. CMS apply a series of technical steps to turn these 53 individual measures into a single Star Rating for each hospital (see Section 2.1.4 for an overview of the technical processes involved).

In this chapter I describe my assessment of the impact of making an alternative technical choice for three of these technical steps. I reimplemented the current CMS approach to use as a base score. I then calculated new composite indicators that followed the same steps as the current CMS approach, except for one (or more) of these three technical steps where these new indicators took an alternative, but plausible, approach (Table 8). I compared hospital performance on each of these new composite indicators with performance on my reimplementation of the current CMS approach.

The three technical steps were:

- How domain scores were weighted before combining them into an overall summary score.
- How different individual measures were grouped into distinct domains, such as ‘mortality’ or ‘timeliness of care’.
- How measures were transformed onto a common scale, so that heterogeneous measures such as ‘median time from admit decision to ED departure’ and ‘MRSA infections’ may be combined into the composite.

Table 8 summarises the current approach CMS use at each of these technical steps, and highlights some potential problems with this current approach. It briefly describes an alternative approach that would address these potential problems – this analysis primarily examined the impact of adopting these alternative approaches. I examined the impact of adopting the alternative approach at just one technical step, and also at two or three technical steps concurrently. I additionally carried out a probabilistic sensitivity analysis using Monte

Carlo simulations. All analyses were carried out in Stata v15.1 [164]. The final analytical code and datasets are available online [129].

Table 8. Potential disadvantages of three technical steps in the calculation of the CMS Hospital Compare Star Ratings, and an alternative step that addresses the key drawbacks.

Current feature of the CMS Hospital Compare Star Ratings	**Potential problems with current feature**	**Possible alternative solution and justification**
Standardisation of different measures is based on Z-scoring	The context of the individual measures is ignored. “Better than average” performance on one measure may be objectively poor, while on others may be truly excellent. Changes in measure performance over time can not be tracked. Star Ratings calculated in different years are not truly comparable, as the reference points have differed.	Identify fixed reference points for good performance on each measure based on the context and meaning of the measure. Use these reference points to standardise measures to a common scale, for example with 0 equating to very poor care and 100 being excellent care.
The domains within which the individual measures are grouped are a priori defined normatively	Individual measures in a given domain may not all reflect the same aspects of quality. If a domain has five measures, of which four relate to aspect of quality Q_1 and one relates to aspect of quality Q_2, then good performance on the four measures of Q_1 may mask poor performance on the single measure of Q_2.	Use exploratory factor analysis to identify domains including only measures that empirically relate to the same aspect of quality.
Each of the 4 outcome indicator domains are given a weight that is 7 times greater than that given to the 3 process indicator domains	Hospital performance may be highly sensitive to selection of weights, but the choice of weights is not well justified.	Verify whether choice of weights matters, for example by comparing with alternative weight specifications or by carrying out a probabilistic sensitivity analysis.

3.3.2 Data

Publicly available aggregated CMS Hospital Compare data, from the January 2020 update to the CMS Hospital Compare Star Ratings [115], were used in this analysis. The CMS Hospital Compare Star Ratings are based on 53 individual measures that are normatively grouped into seven domains. Four of these are ‘outcome’ (mortality, re-admission, safety of care, and patient experience), and three are ‘process’ domains (timeliness of care, efficient use of medical imaging, effectiveness of care). Each outcome domain receives a weight of 0.22 in the final composite score, while each process domain receives a weight of 0.04 therefore the overall rating is heavily weighted towards the outcome domains which make up 0.88 of the total score (0.22 X 4) if all domains are reported. As an additional rule, for the composite score to be produced, a hospital must have enough data for three of the domains of quality, with at least one of those domains being either mortality, re-admission or safety of care.

Individual measures included in the Star Ratings relate to different scales. Of all measures:

- 37 were proportions or percentages, such as 30-day mortality from acute MI.
- Six were rates, such as excess days in acute care for acute MI per 100 admissions.
- Five were rate ratios, such as the ratio of the observed MRSA rate to the expected rate.
- Five were time-to-event statistics/metrics, such as median time in ED from arrival to departure.

3.3.3 Measuring impact of technical choices

I assessed the impact of adopting plausible alternative technical approaches on the score underlying the CMS Hospital Compare Star Ratings graphically, and quantified this assessment using Kendall’s Tau. I assessed the impact on the Star Ratings by describing the frequency of ‘extreme’ changes, operationalised as a hospital being reclassified from 4-5 stars to 1-2 stars, or from 1-2 stars to 4-5 stars. I summarised the probabilistic sensitivity analyses by describing the 25th to 75th centile range of ranks for each hospital across the Monte Carlo simulation, and by considering changes in assigned star rating.

3.4 Findings

The January 2020 CMS Hospital Compare Star Ratings dataset included data on 4,586 US hospitals. Of these, 3,726 reported sufficient measures in enough domains for me to assign them a star rating when implementing the current CMS methodology using the publicly available data (Table 9, and the discussion of missing data in this dataset in Section 2.1.3.1).

The segmentation of the 3,726 included hospitals by number of domains contributing to the overall star rating is shown in Table 9. Among hospitals I could assign a star rating:

- two in three (63%, 2,338/3,726) were assigned a score for all seven domains of quality
- one in eight (13%, 476/3726) were assigned a score for six of the seven domains
- one in eleven (9%, 326/3726) were assigned a score for five of the seven domains
- one in six (16%, 586/3726) were assigned a score for three or four of the seven domains

Among the 1,338 hospitals with three to six domains reported (i.e. 3,726 included hospital overall minus 2,338 with scores in all seven domains, see above), certain domains were more likely to be missing than others. Among these hospitals, almost all (97%) had a score for the readmission domain (Table 9), while just 21% had a score for the safety of care domain.

Hospitals with domain scores for all seven quality domains appeared on average to be worse performers. The assigned star ratings are based on a summary score (the weighted average of the domain scores). Ranking hospitals on this summary score, the mean rank for hospitals that reported all seven domains was 2,086 (the 56th percentile, given 3,726 is the worst possible rank); by comparison, for hospitals that only reported three or four of the seven domains the mean rank was 1,378 (the 37th percentile, Table 10).

Table 9. Percentage of hospitals for which domain scores could be calculated for the CMS Hospital Compare domains of quality, by number of domains for which a domain score could be calculated.

	All hospitals	All seven domains reported	Three to six domains reported	Six domains reported	Five domains reported	Three or four domains reported
Number of hospitals	3726	2338	1388	476	326	586
Percent with score for each CMS Hospital Compare domain						
Mortality	89%	100%	71%	96%	82%	45%
Safety of care	70%	100%	21%	38%	10%	12%
Readmission	99%	100%	97%	100%	100%	92%
Patient experience	88%	100%	67%	99%	72%	39%
Efficient use of medical imaging	80%	100%	47%	67%	56%	26%
Timeliness of care	94%	100%	84%	100%	90%	68%
Effectiveness of care	94%	100%	85%	100%	90%	69%

Table 10. Mean, minimum and maximum hospital ranks by the number of CMS Hospital Compare domains reported, for the current specification and under each of the three main alternative specifications considered.

	Mean rank (out of 3,726)	Mean percentile rank	Best rank	Worst rank
Current specification				
All seven domains reported	2086.0	56th	33	3724
Six domains reported	1731.5	46th	25	3723
Five domains reported	1333.8	36th	3	3719
Three or four domains reported	1377.8	37th	1	3726
Only changing weights given to measure domains				
All seven domains reported	2125.3	57th	16	3717
Six domains reported	1694.1	45th	42	3722
Five domains reported	1298.8	35th	2	3724
Three or four domains reported	1270.8	34th	1	3726
Only changing approach to grouping measures				
All seven domains reported	2222.3	60th	28	3792
Six domains reported	1735.9	47th	32	3791
Five domains reported	1247.7	33rd	9	3776
Three or four domains reported	1182.0	32nd	1	3795
Only changing approach to standardising measures				
All seven domains reported	1750.7	47th	85	3720
Six domains reported	2374.5	64th	32	3718
Five domains reported	2197.7	59th	36	3725
Three or four domains reported	1712.4	46th	1	3726

3.4.1 Technical choice 1: Weights used to combine domain scores

The current specification of the CMS Hospital Compare Star Ratings uses unequal weights to combine domain scores, giving a weight of 0.22 to each of the four ‘outcome’ domains (mortality, readmission, safety of care, and patient experience) and a weight of 0.04 to each of the three ‘process’ domains (timeliness of care, efficient use of medical imaging, and effectiveness of care). Missing domain scores are handled by rescaling the weight given to domains with known scores, in proportion to the weight they would receive if all domains were reported. I compared the current CMS approach with an alternative approach giving equal weight to each of the domains of quality used in the Star Ratings.

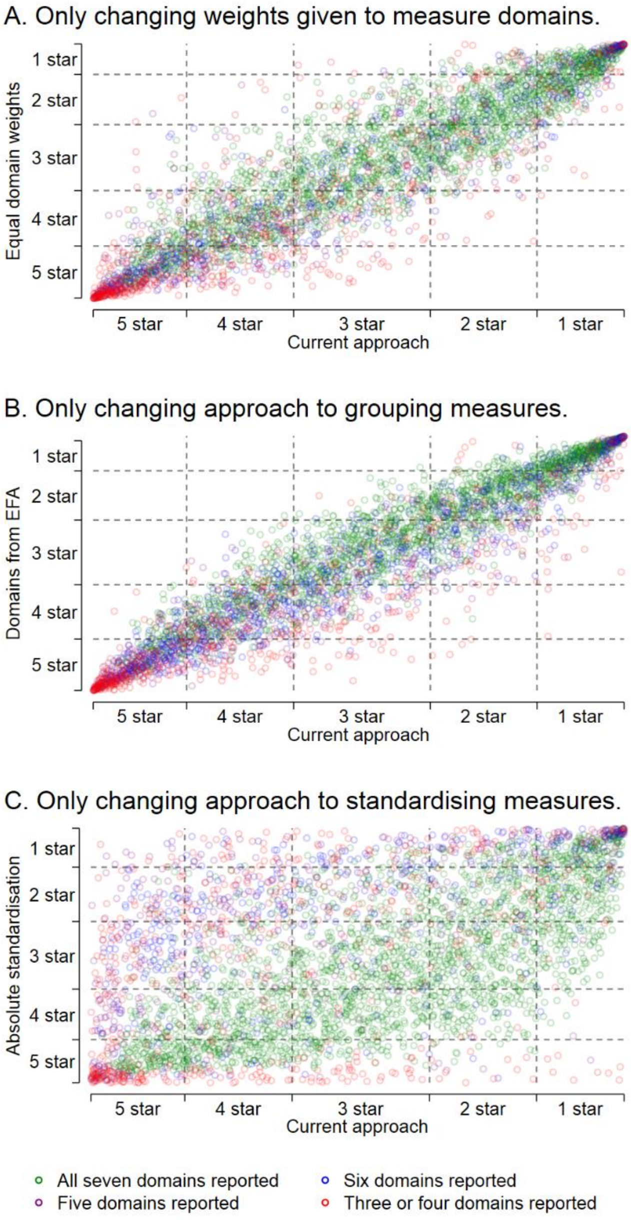
My analysis found that summary scores using the current preference-based weights were strongly correlated with summary scores produced using equal weights, with a Kendall’s Tau of 0.75 across all hospitals (Table 11 – see first row within second super-row) and Tau values of similar order for all missing-domain hospital categories. Examining the plot of ranks under the alternative specification against rank under the current specification indicates that there was substantial re-ordering and a degree of reclassification between star rating categories (Figure 4A). Yet for most hospitals, the rank under the current CMS specification was a reasonable guide to the rank under the alternative specification of domain weights, with the strength of correlation being fairly consistent across hospitals with different numbers of reported domains.

Changing the weight specification did not lead to substantial degree of extreme changes in the assigned Star Ratings category (Table 11 – see first row within first super-row). Only 18 of the 3726 hospitals (or 0.5%) were reclassified from 4/5 stars to 1/2 stars (or vice versa), with an apparent over-representation in this small group of hospitals reporting only three or four quality domains (13/18).

Table 11. Number of hospitals moving from 4 or 5 stars to 1 or 2 stars (or vice versa) when changing from current specification, and Kendall's Tau correlation coefficient between current specification and each other potential specification. Numbers calculated for all hospitals, and hospitals in missing domain groups as currently reported by CMS.

	All hospitals		**Hospitals with all seven domains reported**		**Hospitals with three to six domains reported**		**Hospitals with six domains reported**		**Hospitals with five domains reported**		**Hospitals with three or four domains reported**	
Total hospitals with CMS Star Rating	***N***	***(%)***	***N***	***(%)***	***N***	***(%)***	***N***	***(%)***	***N***	***(%)***	***N***	***(%)***
	3726		*2338*		*1388*		*476*		*326*		*586*	
Hospital performance changing from 4/5 to 1/2 stars, or vice versa												
Only changing weights given to measure domains	18	(0.5)	2	(0.1)	16	(1.2)	2	(0.4)	1	(0.3)	13	(2.2)
Only changing approach to grouping measures	10	(0.3)	1	(0.0)	9	(0.6)	0	(0.0)	1	(0.3)	8	(1.4)
Only changing approach to standardising measures	296	(7.9)	95	(4.1)	201	(14.5)	63	(13.2)	59	(18.1)	79	(13.5)
Changing weights and grouping	41	(1.1)	26	(1.1)	15	(1.1)	1	(0.2)	1	(0.3)	13	(2.2)
Changing weights and standardisation	509	(13.7)	362	(15.5)	147	(10.6)	49	(10.3)	28	(8.6)	70	(11.9)
Changing grouping and standardisation	448	(12.0)	190	(8.1)	258	(18.6)	78	(16.4)	59	(18.1)	121	(20.6)
Changing weights, grouping and standardisation	455	(12.2)	211	(9.0)	244	(17.6)	75	(15.8)	54	(16.6)	115	(19.6)
Kendall's Tau correlation coefficient	**Tau**		**Tau**		**Tau**		**Tau**		**Tau**		**Tau**	
Only changing weights given to measure domains	0.75		0.75		0.72		0.79		0.72		0.67	
Only changing approach to grouping measures	0.79		0.80		0.77		0.83		0.81		0.71	
Only changing approach to standardising measures	0.39		0.52		0.39		0.43		0.46		0.32	
Changing weights and grouping	0.70		0.67		0.71		0.77		0.75		0.65	
Changing weights and standardisation	0.35		0.30		0.34		0.44		0.35		0.26	
Changing grouping and standardisation	0.24		0.27		0.16		0.13		0.18		0.18	
Changing weights, grouping and standardisation	0.25		0.27		0.17		0.15		0.19		0.19	

Figure 4. Impact on CMS Hospital Compare Star Ratings of (A) only changing the weights given to measure domains (B) only changing the approach to grouping measures (C) only changing the approach to standardisation. Hospitals are coloured according to the number of measure domains that could not have a score calculated, with hospitals that reported all seven domains in green, those that reported six domains in blue, those that reported five in purple, and those that reported three or four domains in red. The grey dashed lines show the boundaries of different star rating categories, with the bottom left square including hospitals that received five stars under both specifications and the bottom right square those which received five stars under the alternative specification (Y-axis) but one star under the current specification (X-axis). For all of the three "single changes" in specification of the Star Ratings methods considered, the impact on hospitals reporting all sevel quality domains was smaller than the impact on hospitals that had three or more domains missing. In panel C, it is clear that there is very little correlation between summary scores in the alternative approach using a different method of standardisation and the current approach except if hospitals report all domains. In particular, a substantial number of hospitals that only report three or four domains move from among the worst under the current specification to among the best if a different approach to standardisation is used. Changing the approach to grouping measures into domains (panel B) or the weights given to measure domains (panel A) has less apparent impact.



3.4.2 Technical choice 2: Grouping of measures into domains

The seven domains used by the CMS Hospital Compare Star Ratings are: ‘Mortality’, ‘Safety of care’, ‘Readmission’, ‘Patient experience’, ‘Efficient use of medical imaging’, ‘Timeliness of care’ and ‘Effectiveness of care’ (Figure 5, Table 12). When identifying domains empirically through correlations between measures using exploratory factor analysis, two key differences can be seen. First, there only appeared to be six empirical domains (one of which contained just two measures) compared with the seven domains under the current approach. Second, only two of the current domains matched the domains identified empirically. For two domains (Mortality’ and ‘Patient experience’) all the individual measures that are currently included in them normatively did also load onto them empirically. However, measures in all other domains appeared to correlate with several different underlying latent factors. Table 12 presents the individual measures and the domains with which they were associated by each approach.

My analysis showed that using the scores of the six empirical domains when producing the star ratings, instead of the seven normative domains currently used by CMS, had an impact on hospital performance (Figure 4B). Overall, the composite indicator produced using this alternative approach was strongly correlated with that produced using the current CMS approach, with a Kendall’s Tau correlation of 0.79 (Table 11 – see second row within second super-row). But the rank discordance between the two approaches to grouping measures was greater for hospitals reporting fewer domains (Figure 4B).

While only 10 hospitals (0.3% of all hospitals with summary scores) changed from among the top two-star ratings to among the bottom two, or vice versa, this small group chiefly comprised hospitals that only reported three or four domains (8/10, Table 11 – see second row within second super-row).

Figure 5. Sankey diagram showing, on the left, the 7 Hospital Compare domains that are used currently in the calculation of the composite, together with number of individual measures which contribute in each domain and, on the right, the 6 domains identified using exploratory factor analysis, with the accompanying number of individual measures. The Hospital Compare domains 'Mortality' and 'Patient experience' were empirically confirmed by the factor analysis as comprising the exact same measures, but measures included in the other five normative domains were split into four different empirical domains. Table 12 gives details of the measures and domains they were assigned to.

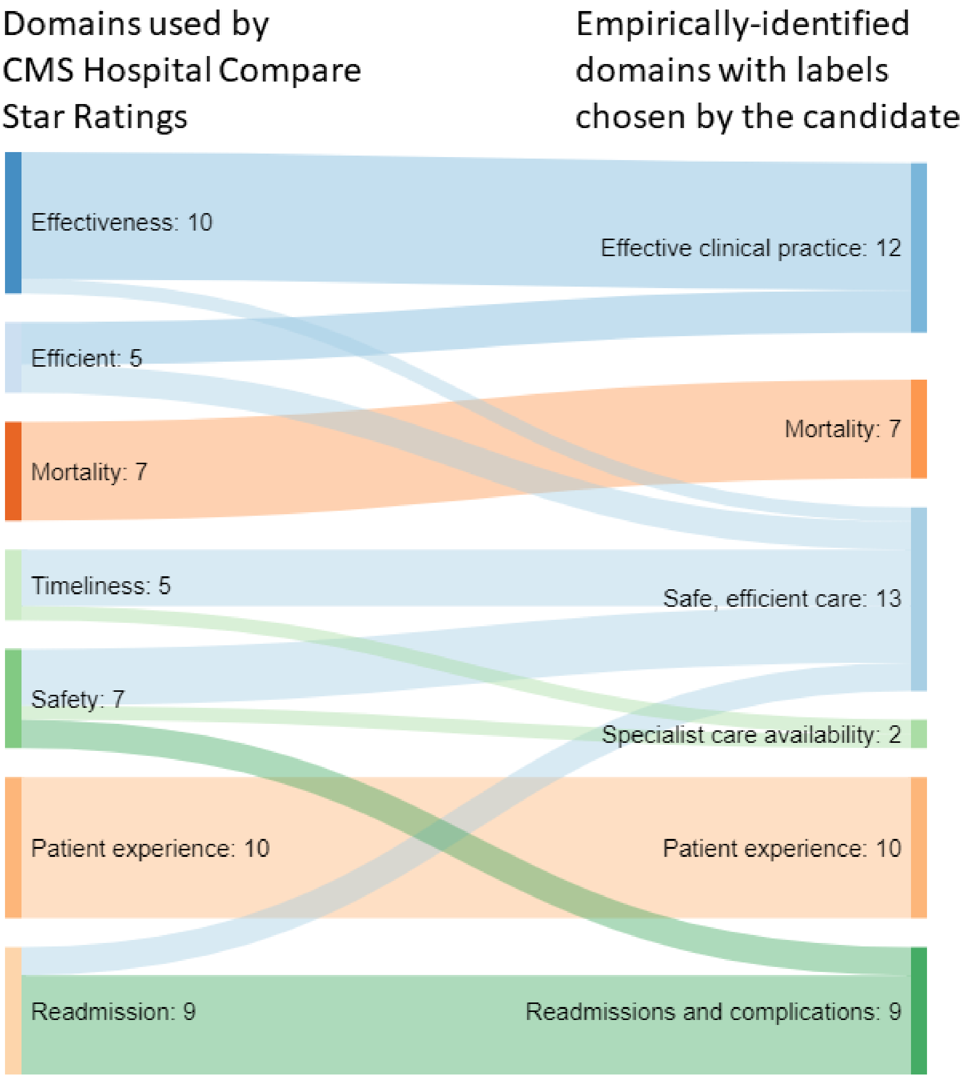

Domains used by CMS Hospital Compare Star Ratings
Empirically-identified domains with labels chosen by the candidate
Effectiveness: 10
Efficient: 5
Mortality: 7
Timeliness: 5
Safety: 7
Patient experience: 10
Readmission: 9
Effective clinical practice: 12
Mortality: 7
Safe, efficient care: 13
Specialist care availability: 2
Patient experience: 10
Readmissions and complications: 9

Table 12. Measures in the CMS Hospital Compare Star Ratings and corresponding domains, both according to the current Star Ratings and based on exploratory factor analysis.

Performance measure	CMS Hospital Compare Star Ratings domain	Empirically-identified domain (based on Z-score standardised measures)
Patients receiving flu vaccinations	Effectiveness of care	Effective clinical practice
Patients receiving appropriate care for sepsis	Effectiveness of care	Effective clinical practice
Patients having a blood clot while not receiving prevention	Effectiveness of care	Effective clinical practice
Patients with a history of polyps receiving follow-up colonoscopy	Effectiveness of care	Effective clinical practice
Mothers delivering early unnecessarily	Effectiveness of care	Effective clinical practice
Patients with bone metastases receiving radiation therapy	Effectiveness of care	Effective clinical practice
Patients receiving appropriate follow-up after colonoscopy	Effectiveness of care	Effective clinical practice
Staff receiving flu vaccinations	Effectiveness of care	Effective clinical practice
Stroke patients receiving a brain scan within 45 minutes	Effectiveness of care	Effective clinical practice
Patients leaving the ED unseen	Effectiveness of care	Safe, efficient care
OP abdomen CT scans that were "double" scans	Efficient use of medical imaging	Effective clinical practice
Cardiac imaging stress tests before low-risk surgery	Efficient use of medical imaging	Effective clinical practice
OP thorax CT scans that were "double" scans	Efficient use of medical imaging	Effective clinical practice
OP brain CT scans with a sinus CT at the same time	Efficient use of medical imaging	Safe, efficient care
MRI lumbar spine for low back pain as first option	Efficient use of medical imaging	Safe, efficient care
30-day mortality from AMI	Mortality	Mortality
30-day mortality from CABG	Mortality	Mortality
30-day mortality from COPD	Mortality	Mortality
30-day mortality from heart failure	Mortality	Mortality
30-day mortality from pneumonia	Mortality	Mortality
30-day mortality from stroke	Mortality	Mortality
30-day mortality from surgical complications	Mortality	Mortality
Care transition	Patient experience	Patient experience
Cleanliness of hospital environment	Patient experience	Patient experience
Communication about medicines	Patient experience	Patient experience
Communication with doctors	Patient experience	Patient experience
Communication with nurses	Patient experience	Patient experience
Discharge information	Patient experience	Patient experience
Global hospital rating	Patient experience	Patient experience
Quietness of hospital environment	Patient experience	Patient experience
Responsiveness of hospital staff	Patient experience	Patient experience
Willingness to recommend hospital	Patient experience	Patient experience
Excess days in acute care following acute MI	Readmission	Readmission and complications
Excess days in acute care for heart failure	Readmission	Readmission and complications

Performance measure	CMS Hospital Compare Star Ratings domain	Empirically-identified domain (based on Z-score standardised measures)
Hospital-wide readmission	Readmission	Readmission and complications
Readmission following CABG	Readmission	Readmission and complications
Readmission for COPD	Readmission	Readmission and complications
Readmission following hip and knee surgery	Readmission	Readmission and complications
Excess days in acute care for pneumonia	Readmission	Safe, efficient care
Hospital visits after outpatient colonoscopy	Readmission	Safe, efficient care
Hip and knee surgery complications	Safety of care	Readmission and complications
MRSA infections	Safety of care	Readmission and complications
Catheter-associated urinary tract infection	Safety of care	Safe, efficient care
Central line-associated blood stream infection	Safety of care	Safe, efficient care
PSI-90	Safety of care	Safe, efficient care
Surgical site infection after colon surgery	Safety of care	Safe, efficient care
Surgical site infection after hysterectomy	Safety of care	Safe, efficient care
C. difficile infections	Safety of care	Specialist care availability
ED - time admit decision to departure	Timeliness of care	Safe, efficient care
ED - time arrival to departure	Timeliness of care	Safe, efficient care
ED - time arrival to discharge	Timeliness of care	Safe, efficient care
OP - time to specialist care for suspected acute MI	Timeliness of care	Safe, efficient care
OP - time to specialist care for suspected acute MI	Timeliness of care	Specialist care availability

3.4.3 Technical choice 3: Standardisation to consistent scales

When I took an alternative approach to standardisation from that used currently by CMS, a substantial degree of reordering of hospital ranks was evident (Figure 4). The rank of a hospital on the summary score underlying the star rating on the current approach correlated poorly with the rank the same hospital would receive under the alternative approach to standardisation, with a Kendall's Tau correlation coefficient of 0.39 (Table 11 – see row three within second super-row). Concordance appeared higher for hospitals reporting all seven domains of quality (Kendall's Tau = 0.52) than for those which were not reporting one or more domains (Tau = 0.39).

Changing the approach to standardisation led to substantial reclassification of hospitals' star ratings. Overall, one in twelve hospitals (8%, 296 of 3726) were reclassified from one or two stars to four or five stars or vice versa. The amount of misclassification greater for hospitals which did not report all seven domains (Figure 4C, Table 11 – see row three within first super-row). Among hospitals which reported all seven domains of quality around 4% (95 of 2338) were reclassified from 4-5 to 1-2 stars or from 1-2 to 4-5 stars, but this proportion was 15% (201 of 1388) among for hospitals that reported six or fewer of the domains.

Hospitals that did not report as many of the domains of quality were more sensitive to changes in the standardisation of individual measures. As noted previously, on average, under the current specification hospitals with greater number of missing domains had better performance compared those without/lower number of missing domains (Table 10 – see columns 2-3 in particular within all rows except those in the bottom super-row cluster). When using the alternative approach to standardisation there was no long a clear association between number of domains reported and the hospital ranks, with the mean rank in hospitals reporting all domains (1751) comparable to that among hospitals only reporting three or four domains (1712, Table 10 – see bottom super-row cluster, columns 2-3 in particular). As described earlier (see Section 3.4, Table 9), across the 3726 hospitals included in this analysis, 99% reported enough information for a score to be produced for the readmission domain, compared with just 70% for the safety of care domain. Using the current approach to standardisation, this may not appear to matter as by definition the performance of the average hospital on each of the individual measures and hence on the domain scores was 0. Under the alternative approach to standardisation, it was intended to be possible for hospitals on average to do better on some domains than on others. For example, among hospitals for which a domain score could be calculated, average performance on the readmission domain was 93.2 out of 100, compared with 99.3 out of 100 on the safety domain and 86.2 out of 100 on the patient experience domain (Table 13). Because domains had different probabilities of being missing, these differences in average performance did not average out.

Table 13. Percentage of hospitals for which domain scores could be calculated for the CMS Hospital Compare domains of quality and mean hospital domain score under the alternative absolute standardisation approach, by number of domains for which a domain score could not be calculated. This table partly reproduces Table 9.

	All hospitals	Hospitals with all seven domains reported	Hospitals with three to six domains reported	Hospitals with six domains reported	Hospitals with five domains reported	Hospitals with three or four domains reported
Number of hospitals	***3726***	***2338***	***1388***	***476***	***326***	***586***
Percent with score for each CMS Hospital Compare domain						
Mortality	89%	100%	71%	96%	82%	45%
Safety of care	70%	100%	21%	38%	10%	12%
Readmission	99%	100%	97%	100%	100%	92%
Patient experience	88%	100%	67%	99%	72%	39%
Efficient use of medical imaging	80%	100%	47%	67%	56%	26%
Timeliness of care	94%	100%	84%	100%	90%	68%
Effectiveness of care	94%	100%	85%	100%	90%	69%
Mean domain score under the alternative absolute standardisation approach						
Mortality	87.7	87.6	87.7	87.7	87.8	87.8
Safety of care	99.3	99.3	99.4	99.4	99.4	99.5
Readmission	93.2	93.5	92.6	92.9	92.6	92.4
Patient experience	86.2	85.6	87.5	86.5	88.3	88.7
Efficient use of medical imaging	93.0	92.6	94.4	93.9	94.7	95.0
Timeliness of care	84.1	80.0	92.3	88.7	93.9	95.6
Effectiveness of care	87.1	87.6	86.0	86.9	85.0	85.8

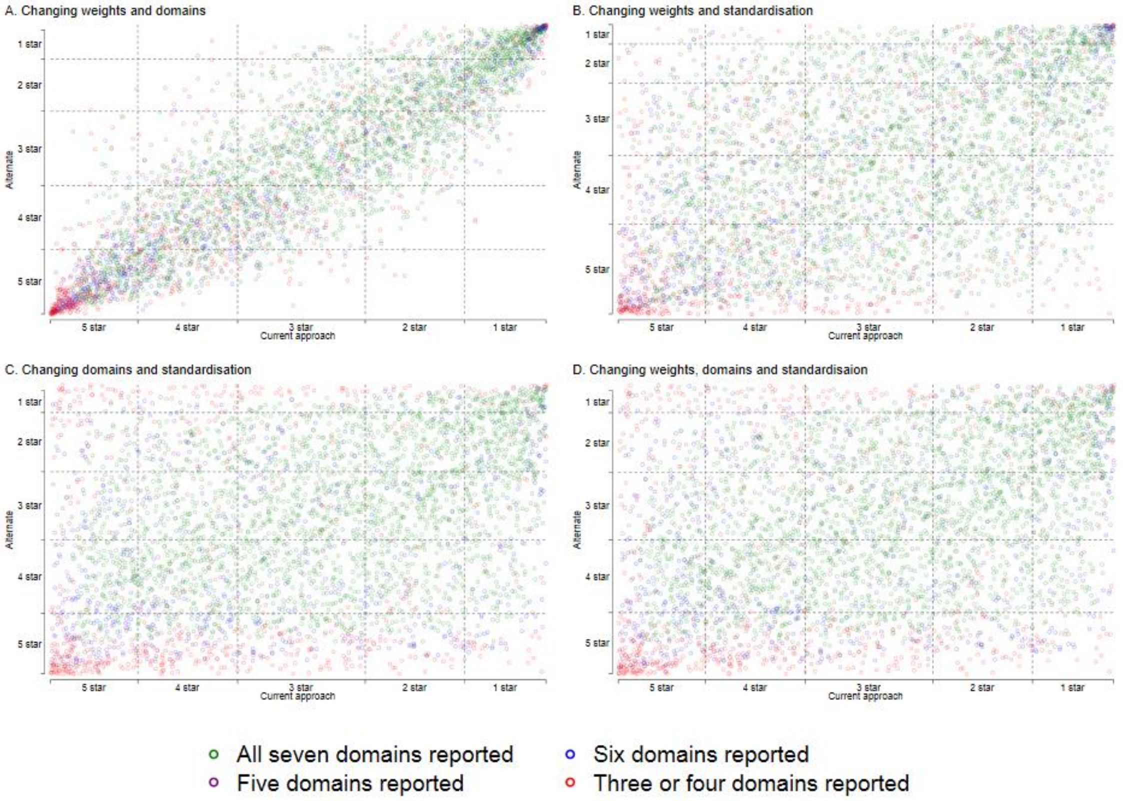
3.4.4 Changing multiple decisions at the same time

As perhaps might be expected, combinations of alternative approaches led to greater discordance in hospital performance than changing any one approach on its own (Table 11 – see bottom four rows within the second super-row, and Figure 6) .For example, when comparing performance on composite indicators produced with different approaches to standardisation, grouping, and weighting, the Kendall's Tau correlation was 0.25, compared with a minimum Kendall's Tau of 0.39 for any of these changes individually. Once multiple changes to the specification were considered, a hospital's rank under the current CMS specification was not a reasonable guide to its likely rank under an alternative specification that differed in two or more technical choices (Figure 6, left-most column).

Changing from the current specification to an alternative specification with a different approach to standardisation and a different approach to weighting domains led to the largest amount of extreme reclassification, with 14% of all hospitals moving from 4-5 to 1-2 stars or from 1-2 stars to 4-5 stars (Table 11 – see bottom four rows within the first super-row).

The above observations were not surprising; the differences in the standardisation approach were the most extreme changes, and may be expected to have the most impact on the time-based measures that were all in the 'process' domains which were the domains that received more weight in the move to equal weighting. Yet while extreme reclassifications of Star Rating were more likely under this change, other changes involving standardisation led to worse agreement in hospital ranks overall. For example, changing the way measures were grouped together with the way measures were standardised gave a Kendall's Tau of 0.24 (Table 11 – see second super-row, values in second column), compared with a Tau value of 0.35 if changes were made to the domain weights together with the way measures were standardised.

Figure 6. Scatter plots of hospital ranks on alternative specifications that differ from the current CMS approach in two or three technical decisions against the hospital rank under the current CMS approach. Each dot represents a single hospital, with vertical position representing its rank under the alternative specification considered and the horizontal position its rank under the current approach. This figure is intended to aid interpretation of the summary statistics presented in Table 11. For example, the Kendall's Tau of 0.7 when weights of domains and the way measures are grouped into domains is changed represents good agreement on average but with many individual hospitals having substantial shifts in rank, as can be seen in panel A. The Kendall's Tau of 0.35 when the weights of domains and the approach to standardisation are changed reflects that there is very little correlation between rank under this alternative specification and rank under the current CMS approach, as can be seen in panel B.



3.4.5 Probabilistic sensitivity analysis via Monte Carlo simulations

The probabilistic sensitivity analysis I conducted confirmed that hospital performance on the Star Ratings was sensitive to technical choices, and that accounting for the different plausible technical approaches simultaneously tended to increase the uncertainty in a given hospitals rank compared with the perturbation of either weighting, grouping or standardisation specifications performed individually (Figure 7). For example, in the Monte Carlo simulation that only considered different approaches to domain weighting, for most hospitals the average difference between the 25th and 75th centile of hospital-specific simulated ranks was relatively narrow. But when two different approaches to grouping domains were considered in addition to approaches to domain weighting, for some hospitals this 50% range of the ranks spanned both very high and very low rankings for the same hospital.

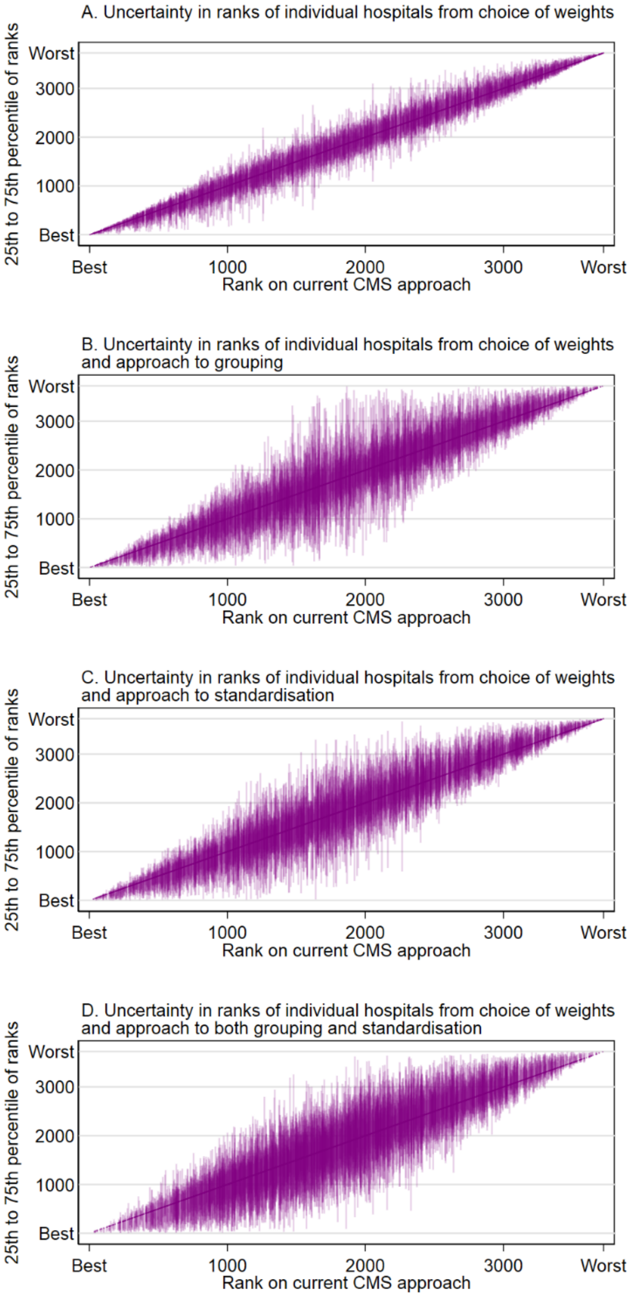
The first Monte Carlo simulation only considered plausible alternative weighting schemes, using weights drawn from distributions centred on the weights currently used in the Star Ratings. This showed that the average difference between the 25th and 75th centile of hospital-specific simulated ranks was 405 places (Figure 8). As the worst rank was 3,726, this represented a substantial but small change of around 11 percentage points, and for two in five hospitals this represented a difference of one (1,478 of 3,726) or two (9 of 3,726) star rating categories.

The second simulation considered plausible alternative weighting schemes and two different approaches to grouping measures into domains (the current CMS approach and factor analysis), and the third simulation considered weighting schemes and two different approaches to standardising measures (the current CMS approach and an approach based on fixed reference points for good performance). The average differences between the 25th and 75th centile of hospital-specific simulated ranks were similar for both these simulations, being 755 and 712 respectively. These corresponded to around 20 percentage point shifts in average hospital rank performance (i.e. 755/3,726 or 712/3,726).

The final simulation included uncertainty about all three technical decisions (weights, grouping, and standardisation). This showed the largest average difference between 25th and 75th centile of hospital-specific simulated ranks, 947 ranks. This corresponded to around a 25 percentage point shift along the overall distribution of ranks.

Examining the width of the 25th to 75th percentile range of ranks (Figure 7), it appeared that in each of the four probabilistic sensitivity analyses the ranks of hospitals near the middle of the distribution were the most perturbed, with the ranks of hospitals with very high or very low performance typically being less affected. This is partially artefactual, as the only way to receive a high (or low) rank on average was to consistently receive a high (or low) rank across simulations, while to receive a middling rank on average hospitals could either be consistently rated toward the centre or have scores that varied between very high and very low. There was essentially an endogenous relationship between the degree to which a hospital has a narrow 25th-75th simulated rank range and the probability it appeared towards either the bottom or the top end of the diagonal line. The histogram of the 25th-75th centile range of ranks for individual hospitals showed that a small number of hospitals received relatively stable rankings across the majority of technical specifications considered (Figure 8), and a small number of hospitals had apparent performances that were highly sensitive to the exact specification of the indicator, with between 4 (perturbing weights only) and 245 hospitals (perturbing all three technical decisions) having a change in ranks of 1863 (out of a total of 3726 ranks, i.e. ≥50%) or greater in each of the four simulations.

Figure 7. There are two elements to this complex data visualisation. First, in each panel figure, each hospital is represented by a (almost invisible, due to number of points that can be visualised) dot on the diagonal. Hospitals are arranged left to right on this diagonal, in order of best to worse simulated average rank (arising from the 10,000 Monte Carlo simulations that the composite indicator of each hospital was subjected to). Second, for each hospital, the 25th to 75th centile of their 10,000 simulated ranks is visualised with vertical lines. The length of these vertical lines indicates the range (number) of ranks (as shown on the y axis) between the 25th and 75th centiles of its simulated ranks. The top panel shows the uncertainty (denoted by the 'width' of the 25th-75th range of possible performance ranks under plausible alternative indicator specifications) just considering changes in choice of domain weights; the middle two panels (with visibly wider uncertainty) show the uncertainty under choice of domain weights combined with either plausible changes to the approach to grouping of indicators into quality domains (2nd from top panel) or the approach to standardisation (3rd from top panel). Lastly the bottom panel (with yet wider uncertainty) shows the uncertainty if all three of weights, approach to grouping, and approach to standardisation were considered.



A. Uncertainty in ranks of individual hospitals from choice of weights

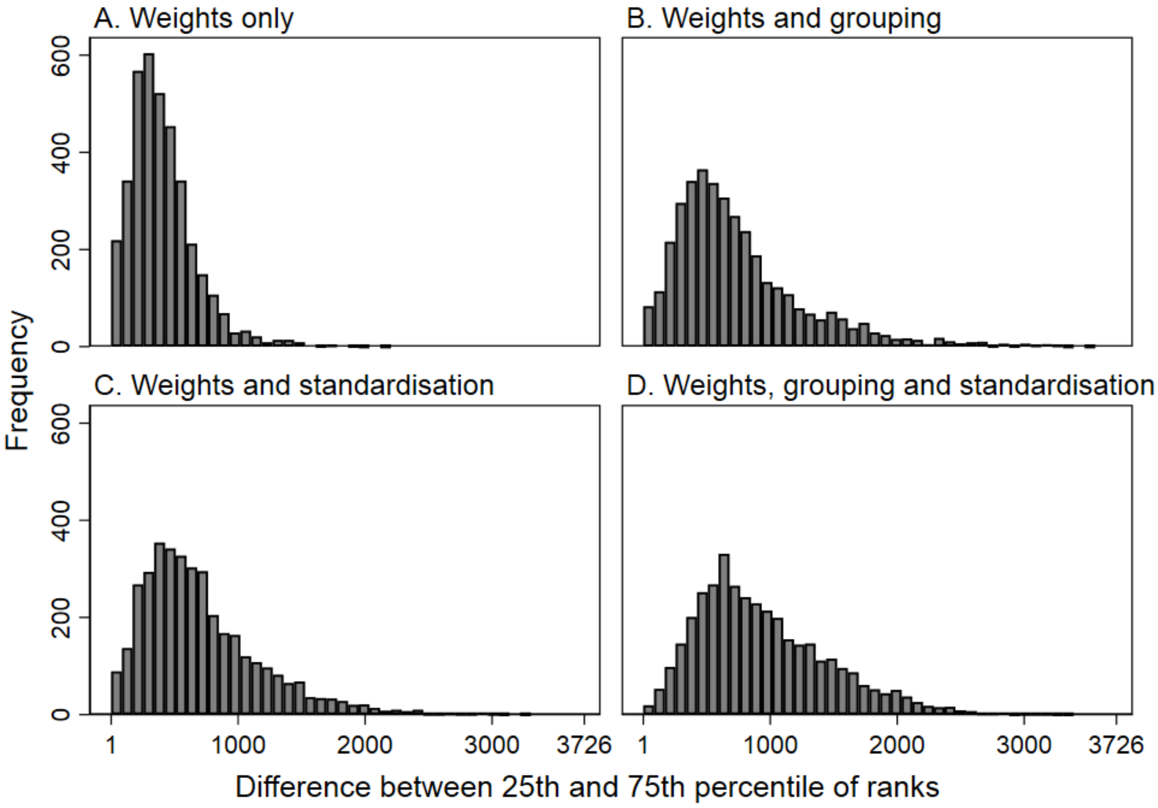
B. Uncertainty in ranks of individual hospitals from choice of weights and approach to grouping

C. Uncertainty in ranks of individual hospitals from choice of weights and approach to standardisation

D. Uncertainty in ranks of individual hospitals from choice of weights and approach to both grouping and standardisation

25th to 75th percentile of ranks
Worst
3000
2000
1000
Best
Best
1000
2000
3000
Worst
Rank on current CMS approach

Figure 8. This is an alternative visualisation of some of the information from Figure 7. The figure shows the distribution of differences between the 25th and 75th centile of simulated ranks for each hospital, which in Figure 7 are represented by the length of the vertical purple lines. As can be seen, almost all these differences are under 1000 ranks when only the choice of weights is perturbed. Yet when two or more choices are changed, the typical range of ranks increases and a substantial number of hospitals have differences between the 25th and 75th centile of simulated ranks of more than 1000 ranks.



3.5 Discussion

This analysis of CMS Hospital Ratings – one of the highest profile and consequential composite indicator schemes globally – shows that the summary scores and star ratings used are highly sensitive to reasonable alternative methodological decisions about technical specifications. Yet, as noted earlier, little documentation exists about these decisions or the reasoning behind them. This work demonstrates the importance both of close attention to technical decisions in the design of composite indicators, and the need for improved transparency about impact of those decisions.

Among the technical choices I examined, plausible alternative standardisation of individual measures had the biggest single impact, while the impact of plausible changes in measuring grouping or domain weighting was lower but non-trivial. Monte Carlo simulation covering a wide range of domain weights gave similar results, showing that changes in weights could easily lead to around 40% of hospitals changing Star Rating category. The impact of changing technical methodological approaches appeared largely driven by relatively large number of hospitals with missing domain information, while those reporting all domains were far less affected. Notably, between 10 (0.3% of hospitals, when only changing approach to grouping measures into domains) and 509 (14% of hospitals, when changing both approach to setting domain weights and approach to measure standardisation) hospitals changed classification from one or two stars to four or five stars, or vice versa, under any of the alternative specifications considered.

For the CMS Hospital Compare Star Ratings, hospitals with missing domain information appear to be disproportionately affected by technical specification changes. This raises questions about whether the Star Ratings provide a misleading summary of quality for these hospitals. This is of particular importance as hospitals that do not report all domains typically receive higher scores under the current specification of the CMS Star Ratings, with the mean rank of hospitals with reporting four or fewer of the seven domains being 1,378, which is 708 ranks (equivalent to 19 centiles of the distribution of ranks) higher than the equivalent mean rank for hospitals that reported all seven quality domains.

3.5.1 Strengths and limitations

This study uses the published data underlying the CMS Hospital Compare Star Ratings to examine the sensitivity of the Star Ratings' summary scores to technical design decisions, using the current specification of the Star Ratings as the base case. A particular strength of the study is that it both examines single pairs of options in detail and uses Monte Carlo simulations to verify that a wider range of options gives similar results.

While the study examines only three of the multitude of technical choices involved in producing a large composite indicator like the Star Ratings, keeping the number of comparisons relatively small made it possible to describe more clearly the impact of the changes considered. Most previous analyses of the sensitivity of composite indicators to technical approaches have only assessed single dimensions of technical specifications [10,81,165–167].

Complex figures such as Figure 4 and Figure 7 aim to summarise the key results of this study, but as a result can be challenging to interpret – especially as the large number of US hospitals means the impact on individual hospitals is difficult to see. There are similar examples in existing research that may make the figure design more intuitive. Figure 4 is similar to the more intuitive figure 1 of Abel, Saunders and Lyratzopoulos's 2014 examination of the impact of case-mix adjustment on cancer patient experience scores [168]. Similarly, Proudlove and colleagues' figure 2 uses boxplots to summarise performance from Monte Carlo simulations of the impact of weighting on a composite indicator of maternity department performance [95], which is similar to my Figure 7 but again with fewer hospitals. Finally, the next chapter – which examines a UK composite – includes figures which are easier to interpret because they include fewer hospitals, with Figure 12 and Figure 13 helpful for interpreting Figure 7.

3.5.2 Context of the literature

Detailed examination of multiple technical aspects of composite indicators in the literature is rare, apart from a few examples. Jacobs, Smith and Goddard (2004) discuss multiple technical aspects of the construction of composite indicators [49], and apply Monte Carlo simulation to demonstrate how these issues can affect a composite indicator. But they used a composite indicator based on just 10 measures that they constructed specifically as an example, rather than an indicator in actual use, and did not examine multiple technical choices simultaneously. Their application of Monte Carlo simulation was also different: it was used to evaluate the statistical uncertainty in the composite indicator under specific technical approaches rather than the sensitivity of the composite indicator to those technical approaches. In 2005, Saisana, Saltelli and Tarantola described a detailed sensitivity analysis of the UN Technology Achievement Index [127], based on eight measures and produced for 72 countries. Further, there are some published sensitivity analyses covering multiple aspects of composite indicators outside healthcare, for example those used in environmental research [169–171].

The analysis presented here of the CMS Star Ratings represents one of the few analyses to examine several different technical specifications of a composite indicator that is currently used for rating healthcare organisations. It is notably different from sensitivity analyses of economic composite indicators or those used in environmental research, first because it covers so many comparative reporting units (3,726 hospitals compared with 72 countries, for example [127]), and second because the Star Ratings are based on 53 individual measures compared with between 5 and 10 for many composite indicators outside healthcare.

Despite the availability of data required to examine and recreate the technical specification of the CMS Star Ratings, there is to my knowledge only one existing analysis that compares performance under the current specification of the Star Ratings with performance under an alternative specification. Recently, Adelman compared an efficient frontier approach to deriving domain scores with the current latent variable modelling approach applied by CMS [172,173]. In essence, an efficient frontier approach derives hospital-specific weights for each measure such that the hospitals' domain score is maximised. This is a very different way of thinking about weighting from the approaches considered in this chapter, but had a broadly similar impact to the choice of weights and approaches to grouping measures examined in this chapter, with for example 1064 of the 3692 hospitals in Adelman's analysis being in a different quintile when the efficient frontier approach was used than when the current CMS latent variable modelling approach was applied [172].

Most assessment of the impact of technical decisions on hospital performance on composite scores has focused on the issue of weighting. Rumball-Smith and colleagues suggest providing "personalized hospital ratings" [174], where users can set their own domain weights to prioritise specific areas. This is an appealing application, but given the impact of approaches to standardisation and to the grouping of measures, my results imply that for personalised ratings, the presentation of a composite indicator should also make it possible to specify approaches to standardising and grouping measures as well.

Others propose calculating "ranking intervals" or other interval summaries of performance [79,95], or applying dominance criteria [80], that is, describing one hospital as better than another only if it has better performance across a wide range of technical specifications of the composite indicator. As demonstrated above (e.g. Figure 7), Monte Carlo simulation makes it relatively easy to produce such summaries of performance for certain plausible technical decisions, especially around measure weights.

While published sensitivity analyses of healthcare composite indicators are uncommon, some guides to the development of composite indicators do suggest that such analyses are carried out [1,50]. An example of this in practice is the US Baby-MONITOR composite indicator of NICU quality [175]. The developers of this indicator compared five different approaches to weighting and aggregating measures, discovering that they all gave similar results, and used this to justify the appropriateness of their base case approach [175]. However, such practice does not seem to be standard, nor does it typically lead to public-facing documentation or publications.

3.5.3 Conclusions

The apparent performance of hospitals on the CMS Hospital Compare Star Ratings is highly sensitive to technical decisions about their design, especially in how individual measures are standardised to a consistent scale. The ratings generated for hospitals that lack domain scores for every domain used in the Star Ratings are particularly problematic, as they appear to be especially sensitive to the precise technical specification. The available technical documentation does not clearly justify why these decisions, rather than other apparently equally plausible alternative technical decisions, have been made. These findings open up to question the robustness of CMS Star Ratings as a guide to hospital performance.

While this analysis highlighted specific problems with the sensitivity of the CMS Star Ratings to alternative, but reasonable, technical specifications, it remains unclear whether other composite indicators of quality would be similarly sensitive. Addressing this requires examinations of other composite indicators. The next chapter explores the sensitivity of a UK composite indicator of healthcare quality to a similar set of technical decisions.

4 Results 2: An analysis using secondary data of the impact of alternative technical approaches in the design of the SSNAP score and level

4.1 Summary

The Sentinel Stroke National Audit Programme (SSNAP) is an important UK national clinical audit. The composite indicators it uses to summarise hospital performance – the SSNAP score and the SSNAP level – are interesting examples of their kind. The SSNAP score provides a numeric summary of hospital performance, ranging from 0-100. The SSNAP level assigns a grade, with A being the best and E the worst. However, the design of these composite indicators is poorly documented, and their sensitivity to the decisions used to create them has not previously been examined. In this chapter I examine and quantify the impact of different technical decisions when constructing the SSNAP score and level, composite indicators that aim to measure the quality of stroke care.

The SSNAP clinical audit data for Jul-Sep 2019 was used for this analysis. These data allowed SSNAP scores and level to be assigned to 134 of the 136 hospitals that routinely admit stroke patients; only one hospital had any missing information for the measures used in constructing the SSNAP composite indicators. The impact of four different alternative technical specifications was assessed: avoiding preliminary rounding of the domains scores; changing weights given to domain scores; changing the way measures are grouped into domains; and changing the way that measures are standardised to consistent scales. As has been made apparent earlier in the thesis, the latter three of the four alternative specifications examined correspond to similar analysis for the CMS Star Ratings. Also similar to the analysis of the CMS Star Ratings, the impact on the SSNAP score was first assessed visually and then quantified using Kendall’s Tau, and the impact on the SSNAP level was examined subsequently.

The findings indicated that the SSNAP *scores* were relatively robust to the various alternative technical approaches examined in this study. In general, expected ranks under the alternative specifications were relatively concordant with ranks on the current specification, in notable contrast to the CMS Star Ratings. However, the SSNAP *levels* were far more sensitive to these technical decisions. For example, around one in four hospitals would have their performance level reclassified if different weighting schemes were applied.

Compared with the CMS Star Ratings, the SSNAP score provides a summary of hospital quality that is robust to alternative plausible technical specifications. The SSNAP level, however, is far less robust, and differences between performance ratings on this indicator are hard to interpret. For example, relatively small changes to the technical specification of how SNNAP levels are assigned can lead to substantial reclassification of hospitals between B and C grades. Sensitivity analysis as applied in this study potentially provides a helpful tool to allow development of more robust composite ratings.

4.2 Introduction

The Sentinel Stroke National Audit Programme (SSNAP) collects and reports information on the quality of stroke care in order to promote quality improvement [130]. This reporting includes two linked composite indicators of the quality of stroke care. These are the SSNAP score, a number between 0 (representing the worst possible performance) and 100 (the best possible performance), and the SSNAP level, a rating between A and E, with A corresponding to ‘world-class’ care and E to ‘care that requires substantial improvement in several aspects’.

The SSNAP composite indicators are methodologically interesting for two main reasons. First, given their role in quality improvement as part of the suite of national clinical audits, it is important to assess their robustness. Second, they are complex composites, summarising many individual performance measures and with interesting approach to standardisation. This complexity makes them difficult to understand at a glance, and makes them more intriguing, from a scholarly perspective, than possible other options for this kind of analysis – including the CCG Improvement and Assessment Framework or the AHRQ PSI-90 Composite Safety Indicator that take a simple approach to combining around ten individual measures [35,44]. Finally, the SSNAP composites are of interest as they provide a naturally contrasting comparison to the CMS Star Ratings, where for two key technical decisions (weighting and standardisation) the developers of SSNAP use the ‘plausible alternative’ approaches I examine for the CMS Star Ratings in Chapter 3.

The SSNAP score and level are based on a collection of measures on which it has been decided that good performance indicates good quality care. In identifying the constituent measures and combining them to produce the overall score and rating, the developers of the indicators are deploying – potentially implicit – assumptions both of good stroke care, and of the relative importance of different aspects of stroke care. These underlying assumptions about standards of care quality are encoded in the technical design of the indicator, impacting each issue from the choice of constituent measures to the score required for a hospital to receive a SSNAP level of B rather than C.

Comparing the outcomes of the current steps versus these plausible alternatives could allow insights into the impact of the current technical specification of the composite indicator on apparent performance. If these plausible alternative specifications have a major impact on apparent hospital performance, then the summaries of quality provided by the SSNAP score and level may need further scrutiny. For example, if SSNAP data are used to prioritise targets for quality improvement, users might end up making suboptimal decisions. Similarly, Fisher and colleagues used the SSNAP score to adjust for unit quality when examining if adopting core components of an early supported discharge service led to a more response and intensive service [176], but if the SSNAP score does not provide a robust measure of quality it becomes more difficult to interpret their results.

In this chapter I examine the impact of alternative approaches for four decisions on hospital ranks on the SSNAP score and on hospital classification on the SSNAP level. These decisions were the:

- Preliminary rounding of domain scores
- Weights used to combine domains
- Approach to assigning individual measures to domains
- Approach to standardisation of individual measures

As alluded to above, the weights used to combine domains and the approaches to standardisation and assigning measures to domains each encode beliefs about the meaning of quality in stroke care and about the links between different measures. Different approaches to standardisation change the value assigned to performing at certain levels on individual measures; different ways of assigning measures to domains represent decisions about which individual measures reflect which aspect of quality; and different domain weight naturally represent different prioritisations of the different aspects of quality. Preliminary rounding of domain scores is included because it has been recommended against and so it is interesting to see how it affects hospital performance in this setting [49].

4.3 Summary of methods

The methods for this study are set out in detail in Section 2.2. In summary, SSNAP data for July-September 2019 were used to calculate two composite indicators, the SSNAP score and level, under their current and multiple alternative technical specifications. The remainder of this section provides a reminder of the current specification of the SSNAP score and level, and then identifies potential issues with the current specification and briefly details alternative specifications that would address these potential limitations. All data and analysis code used to carry out the analysis are available online [141].

Table 2 on page 30 offers a brief summary of the technical specifications that I compared in this analysis and may be a helpful guide to this section. Table 15 below provides a similar summary, with additional details on the perceived limitations of the current SSNAP specification and the way that my proposed alternative specifications addressed these issues.

4.3.1 Current specification of the SSNAP score and level

SSNAP calculate score and level via an intricate process, summarised in the flow chart I present in Figure 9. They base these composite indicators on 44 individual measures chosen by the Royal College of Physicians Intercollegiate Stroke Working Party [136]. These individual measures are grouped into 10 domains, with most domains containing a mix of proportion and time-to-event measures (see Table 4 on page 58 for details). For example, the ‘Scanning’ domain includes the measures ‘Proportion of patients scanned within one hour of clock start’ and ‘Median time between clock start and scan’. The exact individual measures and domains are described in full in Table 19.

SSNAP currently standardise individual measures to a 0-100 scale using absolute quality thresholds applied to individual measures (see Appendix 1) [69], with a few example thresholds shown in Figure 9. The Intercollegiate Stroke Working Party, responsible for the design of the SSNAP score and level, produce clinical guidelines for stroke care [137], and SSNAP use thresholds chosen to be in accordance with these clinical guidelines and the relevant quality standards for stroke care [138]. While updated clinical guidelines and quality standards were released in 2016 [137,139], the technical design of the SSNAP composite indicators have not changed since their first release in 2013.

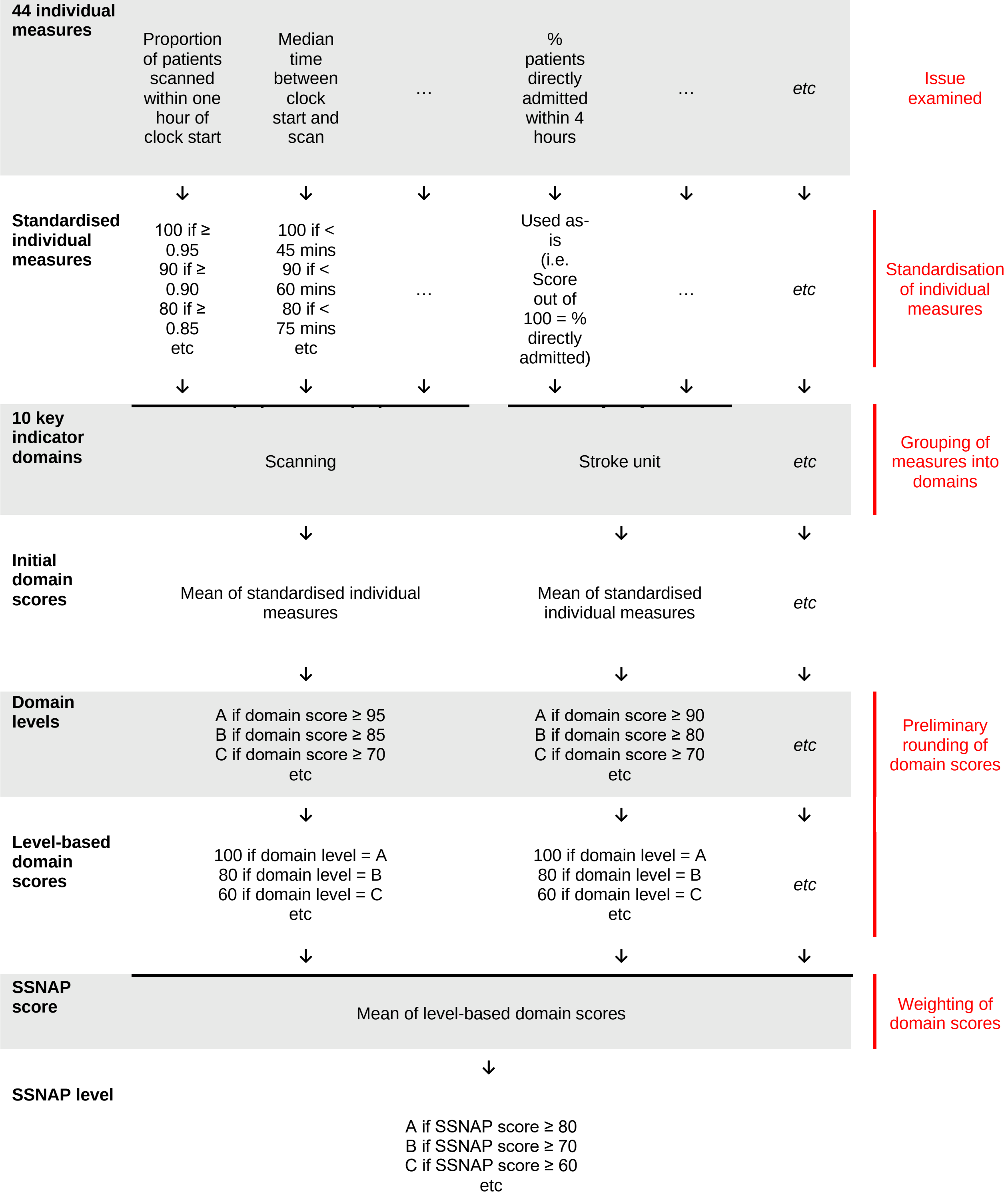
SSNAP combine measure within domains using equal weights, producing an initial score for each of the ten domains (Figure 9). They use these initial scores to assign domain levels based on fixed performance thresholds which differ for each domain (for example, a score of 95 is needed to receive an A on the Scanning domain, while a score of 80 will receive an A on the Thrombolysis domain). These domain levels are then converted back into a level-based domain score that they use to derive the summary composite score (A = 100, B = 70, C = 60, D = 40, E = 20). I found working through an example helpful to understand this process. At Nottingham University Hospitals NHS Trust the average standardised performance across the three measures in the scanning domain in Jul-Sep 2019 was 79.6 out of 100 (Table 14). This corresponded to a scanning domain level of C. As the scanning domain level was C, the Scanning domain score used in deriving the overall SSNAP score and level for this hospital was 60.

SSNAP combine these rounded domain scores using equal weights into an overall summary score. They then adjust this summary score for audit compliance and ascertainment, and the adjusted score is converted into a summary grade (80-100 = A, 70-80 = B, 60-70 = C, 40-60 = D, 0-40 = E).

Table 14. Worked example of the calculation of the Scanning domain raw score, domain level, and rounded score for use in deriving overall SSNAP level. Data are for Nottingham University Hospitals NHS Trust in Jul-Sep 2019.

Individual measures included in the 'Scanning' domain	**Raw performance of individual measures**	**Standardised performance of individual measures (see Figure 9 and Appendix 1, page 238)**	**Initial scanning domain score (mean of the 3 standardised performance measures in column 3)**	**Domain level assigned based on raw/initial domain score** $\begin{pmatrix} \mathrm{A} \geq 95 \\ \mathrm{B} \geq 85 \\ \mathrm{C} \geq 70 \end{pmatrix}$	**Conversion of domain level into score used in deriving overall SSNAP level**
Proportion of patients scanned within one hour of clock start	39.4%	78.8	79.6	C	60
Proportion of patients scanned within 12 hours of clock start	96.1%	100			***Alternative Interpolated value***
Median time between clock start and scan	1:41 (hours:minutes)	60			*72.8*

Figure 9. Flow chart showing the current approach to calculating the SSNAP score and level. The right hand column, with text in red, shows the technical decisions that are examined in this chapter.



44 individual measures	Proportion of patients scanned within one hour of clock start	Median time between clock start and scan	…	% patients directly admitted within 4 hours	…	*etc*	Issue examined
Standardised individual measures	100 if ≥ 0.95 90 if ≥ 0.90 80 if ≥ 0.85 etc	100 if < 45 mins 90 if < 60 mins 80 if < 75 mins etc	…	Used as-is (i.e. Score out of 100 = % directly admitted)	…	*etc*	Standardisation of individual measures
10 key indicator domains	Scanning			Stroke unit		*etc*	Grouping of measures into domains
Initial domain scores	Mean of standardised individual measures			Mean of standardised individual measures		*etc*	
Domain levels	A if domain score ≥ 95 B if domain score ≥ 85 C if domain score ≥ 70 etc			A if domain score ≥ 90 B if domain score ≥ 80 C if domain score ≥ 70 etc		*etc*	Preliminary rounding of domain scores
Level-based domain scores	100 if domain level = A 80 if domain level = B 60 if domain level = C etc			100 if domain level = A 80 if domain level = B 60 if domain level = C etc		*etc*	
SSNAP score	Mean of level-based domain scores						Weighting of domain scores
SSNAP level	A if SSNAP score ≥ 80 B if SSNAP score ≥ 70 C if SSNAP score ≥ 60 etc						

4.3.2 Potential issues with the current calculation of the SSNAP score and level

I used publicly available SSNAP data relating to patients who had strokes in July-September 2019 to examine the impact of alternative technical approaches at four different decision points in the construction of the composite indicators (Figure 9, Table 15) [131]. These decisions were the:

- Preliminary rounding of domain scores
- Weights used to combine domains
- Approach to assigning individual measures to domains
- Approach to standardisation of individual measures

The current approach that SSNAP use for each of these has either potential technical or potential conceptual problems, or both (Table 15). The alternative technical approach examined addresses these issues. Table 15 briefly summarises the current SSNAP approach, outlines some potential problems, and describes the alternative approaches that I considered that addressed the potential limitations of the current approach.

Table 15. Current features of the SSNAP score and level, potential issues introduced by this feature, and a possible alternative solution that addresses the potential problems.

Current feature of the SSNAP score and level	**Potential problems with current feature**	**Possible alternative solution and justification**
Preliminary rounding of domain scores	Threshold boundaries reduce stability and may distort quality improvement priorities [49]. In particular, there is an incentive to game the indicator by focusing on domains that are close to performance thresholds [177].	Use linear interpolation between boundaries so that scores are not rounded and there are no major changes in the contribution to the SSNAP score caused by small changes in the domain score [16].
Each domain receives the same weight in the overall SSNAP score	The different domains of quality may not be as important as each other. For example, acute domains such as 'Scanning' and 'Thrombolysis' may be more important than recovery domains in improving outcomes.	Either: (a) Try to make domain weights reflect some measure of importance, for example giving acute domains more weight than recovery domains. (b) Verify that domain weights are not important by performing probabilistic sensitivity analysis of the impact of domain weights on the SSNAP score and level received by hospitals.
The domains within which the individual measures are grouped are *a priori* defined normatively	Individual measures in a given domain may not all reflect the same aspects of quality. If a domain has five measures, of which four relate to aspect of quality Q_1 and one relates to aspect of quality Q_2, then good performance on the four measures of Q_1 may mask poor performance on the single measure of Q_2.	Use exploratory factor analysis to identify domains including only measures that empirically relate to the same aspect of quality. As this ensures measures within a domain are correlated, this reduces the chance that performance on some aspects of quality could be missed.
Standardisation of different measures a mixture of absolute thresholds (for time-based measures) and continuous functions (for proportion-based measures).	Threshold boundaries for time-based measures reduce stability and may distort quality improvement priorities [49]. Reference points set in 2013 may not adequately reflect appropriate performance today, given 7 years of continuous improvement [8], and in particular hospitals with meaningful differences in performance may all have performance above the highest threshold and so look the same.	One option would be to use average hospital performance and the variation in hospital performance to standardise scores, i.e. Z-scoring [83]. This removes threshold boundaries so that similar performances receive similar scores. This partially addresses the issue of improvements in performance removing differences in the standardised measures, so long as performance on the underlying measures is not approaching a natural ceiling (e.g. 100% scores).

4.4 Findings

The July-September 2019 SSNAP clinical audit produced SSNAP scores and levels for 136 hospital trusts that routinely admit stroke patients [131], and this dataset was used as the basis of my analysis. Two hospitals (Leeds General Infirmary and Downe General Hospital) did not submit data to the audit and so were excluded from my analysis. Most hospital trusts received a level of ‘A’ or ‘B’ (48 of 136, 35% and 45, 33%, respectively), but some received Cs (29, 21%), Ds (23, 9%) and Es (2, 1%). In sharp contrast to the CMS Star Ratings scheme where missing domain data is common (see Table 9 – second row, as an example, and many other manifestations of this problem earlier on) in the SSNAP scheme only one hospital – other than the two that did not submit data – had hospital-level missing data: Southport and Formby District General. This hospital was missing information for three individual measures of the five in the ‘Thrombolysis’ domain (‘% eligible patients given thrombolysis’, ‘% patients thrombolysed within one hour’, and ‘Median time until thrombolysis’).

4.4.1 What if domain scores were not banded before being combined into the overall SSNAP score?

The SSNAP methodology applies an unusual preliminary rounding step in its calculation (see Table 14 and Figure 9). SSNAP use hospital scores on the measures in each domain to assign a domain level (A through E), and then assign this level a value that they use to calculate the overall SSNAP score and level. For the Scanning domain, a domain score of 85 is required to achieve a domain level of B [69]. A score of less than 85 (and more than 70) would receive a domain level of C. A domain level of B is valued at 80 points for combining into the overall SSNAP score, while a domain level of C is only valued at 60 points.

This banding or rounding step – beyond its obvious complexity – introduces a potential problem. A hospital that is just above the performance threshold is treated as exactly the same as one that is far above it, yet is treated as meaningfully different from a hospital that is just below the performance threshold. This step function may distort organisational priorities as substantial improvement in quality is required to improve the score of a hospital just above a threshold [49].

Interpolation between performance thresholds provided an alternative that was less problematic than the use of categorical bands. I operationalised this as follows. If S_i was the score assigned for level i and P_i the performance required to achieve level i, then the interpolated score d_j for hospital j with raw domain score p_j, $P_i \leq p_j < P_{i+1}$, was

$$d_j = S_i + (S_{i+1} - S_i) \times \frac{p_j - P_i}{P_{i+1} - P_i}$$

For example, Nottingham University Hospitals NHS Trust had a raw scanning domain score of 79.6 (Table 14). A scanning domain score of 70 was required to be assigned a scanning domain level of C, while a scanning domain score of 85 was required for a scanning domain level of B. The banded scanning domain value used in calculating the overall SSNAP level was hence 60, and the plausible alternative value was

$$60 + (80 - 60) \times \frac{79.6 - 70}{85 - 70} = 72.8$$

Compared with the current approach, my alternative approach that used interpolation to avoid preliminary rounding had a minor impact on organisational comparisons (Figure 10). Organisational performance when the rounding was removed was strongly but not perfectly correlated with organisational performance under the current SSNAP approach (Kendall's Tau 0.86). As might be expected, using the alternative approach increased performance on the SSNAP score and level (Figure 10), with 36 of the 45 hospitals that received a level of B under the current approach receiving an A when interpolation was used (Table 16).

Figure 10. Scatterplot comparing performance when domain scores were not rounded before being combined into the overall summary score against performance under the current approach. The left panel shows scores and the right panel shows ranks. The Kendall's Tau correlation coefficient applies to either plot.

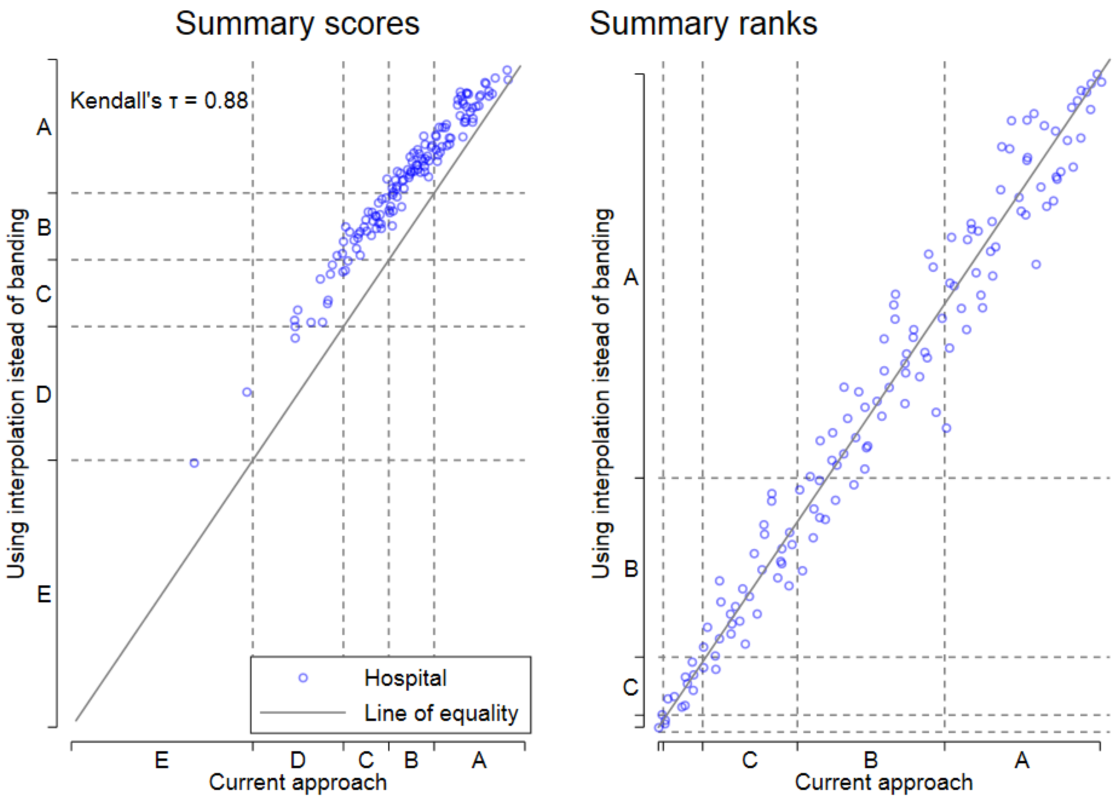


Table 16. Contingency table showing SSNAP level under current approach versus SSNAP level if interpolation was used rather than banding. Zeros have been left blank.

		SSNAP level under current approach					Row total
		A	**B**	**C**	**D**	**E**	
SSNAP level if interpolation was used rather than banding	**A**	48	36				84
	B		9	27	1		37
	C			2	10		12
	D				1	1	2
	E					1	1
Column total		48	45	29	12	2	136

4.4.2 What if different weights were used to combine domain scores into the overall summary score?

SSNAP currently weight domain scores equally when combining them to produce the overall score and grade they present in their audit report [8]. They give no rationale for this choice of weights, nor do they explain why they prefer it over other possible approaches such as deriving weights from some expert assessment of the importance of good performance in each domain. Given this lack of justification for the choice of weights, it was important to examine how robust the current assessment of hospital performance was to different possible choices of weights.

I examined this in two different ways. First, I compared the current SSNAP score, calculated using equal weights, with a single plausible alternative approach. This allowed a more detailed understanding of the impact of this single change, but obscured the potential impact of all the other reasonable weighting schemes that I did not consider. I then addressed this limitation using a Monte Carlo simulation, examining the spread of hospital performances that could be achieved across a wide range of domain weights.

4.4.2.1 Comparison of the current equal domain weights with a single alternative approach

The current SSNAP score and SSNAP level were compared with an alternative approach (somewhat analogous to the current CMS Star Ratings approach outlined in section 0)   that gave 70% of the weight to ‘acute’ domains (those covering acute aspects of the pathway: Scanning; Stroke unit; Thrombolysis; and Specialist assessment) and 30% to ‘recovery’ domains (the other six domains, covering recovery and discharge: Occupational therapy; Physiotherapy; Speech and language therapy; MDT working; Standards by discharge; and Discharge processes). In total, the acute domains included 17 of the 44 measures used in the SSNAP score and level, while the recovery domains included the remaining 27 measures.

Comparing the current equal weights against this alternative set of weights prioritising so-called acute domains suggested the specific choice of weights did not generally have an important impact on the SSNAP score or the SSNAP levels that hospitals received. SSNAP scores under the two different approaches were highly correlated (Kendall’s Tau = 0.8), and the difference in SSNAP score between the two approaches was generally small (Figure 11). The change had more impact on SSNAP levels than SSNAP scores, with 49 of the 136 hospitals (36%) being assigned a different SSNAP level (Table 17); in all except two cases, this level was within one category (i.e. B to C, E to D). For example, 14 of the 48 (29%) hospitals receiving a SSNAP level of A under the current approach receiving a B (13 of 14) or C (one hospital) when using the weights prioritising outcome domains.

Figure 11. Scatterplot comparing performance when acute domains received 70% of the weight and recovery domains 30% of the weight against performance under the current approach (equal weights). The left panel shows scores and the right panel shows ranks. The Kendall's Tau correlation coefficient applies to either plot.

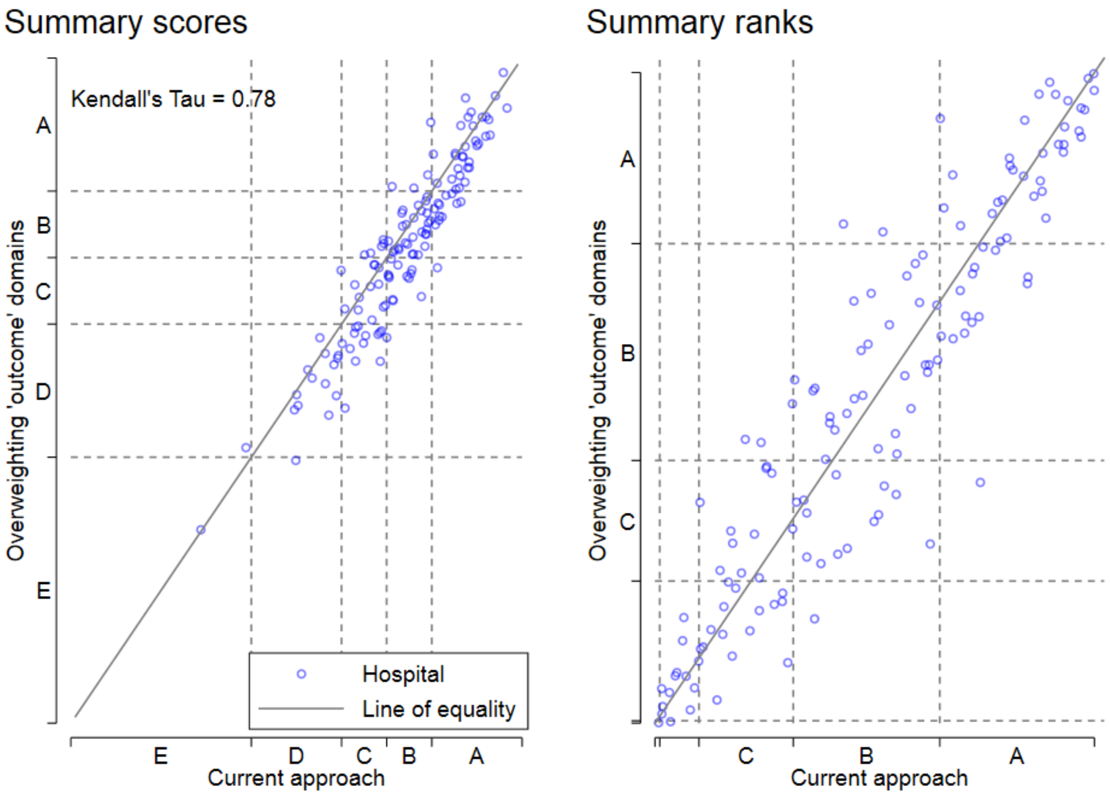


Table 17. Contingency table showing SSNAP level under the current approach versus SSNAP level based on an approach where acute domains were received more weight than recovery domains. Zeros have been left blank.

		Grade under current approach					Row total
		A	**B**	**C**	**D**	**E**	
Grade if acute domains received more weight than recovery domains	**A**	34	2				36
	B	13	29	3			45
	C	1	13	11			25
	D		1	15	12	1	29
	E					1	1
Column total		48	45	29	12	2	136

4.4.2.2 Monte Carlo simulation of approach to domain weighting

The second component of the domain weighting analysis applied Monte Carlo simulation, aiming to show the SSNAP scores and SSNAP levels that could achieved for each hospital by reasonably changing the domain weights alone [79,95]. This simulation suggested that the relatively minor impact observed above when comparing the current equal weights to one specific alternative set of weights was not an artefact of the specific choice of weights (as indicated by the length of the vertical lines in Figure 12 typically covering a narrow range of ranks). Across the 10,000 Monte Carlo draws the lowest (minimum) correlation value observed was a Kendall’s Tau of 0.66 (Figure 13), with a median of 0.85 (IQR 0.82 to 0.87). Given that weights were drawn from a uniform distribution, it was guaranteed that ‘average’ performance across all simulations would approximately match the performance seen using equal weights.

The Kendall’s Tau coefficient on its own does not give the whole picture with regard to SSNAP level reclassification. Looking across all hospitals, 24% of the time (321,123 times out of a possible 1360000), the SSNAP level assigned to individual hospitals were different under randomly chosen weights than under equal weights (Table 18). This particularly appeared to affect hospitals currently judged to have SSNAP level B (30% changing) or C (35% changing), reflecting the narrower width of these performance categories. Consider Chesterfield Royal Hospital, for example. Using the current equal weights, this hospital was judged as SSNAP level of C with a summary score of 68. But with randomly chosen weights, it was only classified as level C in 65% of simulations (6514 of 10000 simulations), being classified as level A in 0.3% of simulations (28 of 10000), as level B in 32% (3173 of 10000), and as level D in 3% (285 of 10000).

Figure 12. Scatterplot comparing hospital performance under randomly-chosen weights against performance under the current approach (equal weights). The blue cross represents the mean across the 10,000 Monte Carlo simulations; the pale blue line shows the range of performance covering the 50% of simulated scores most proximal to either side of the mean value. Under the current specification, multiple hospitals have the same summary score; these have been re-ordered slightly so they can be seen on the plot. Note that this figure is similar in concept to Figure 7 which shows results from the Monte Carlo simulation based on the CMS Star Ratings.

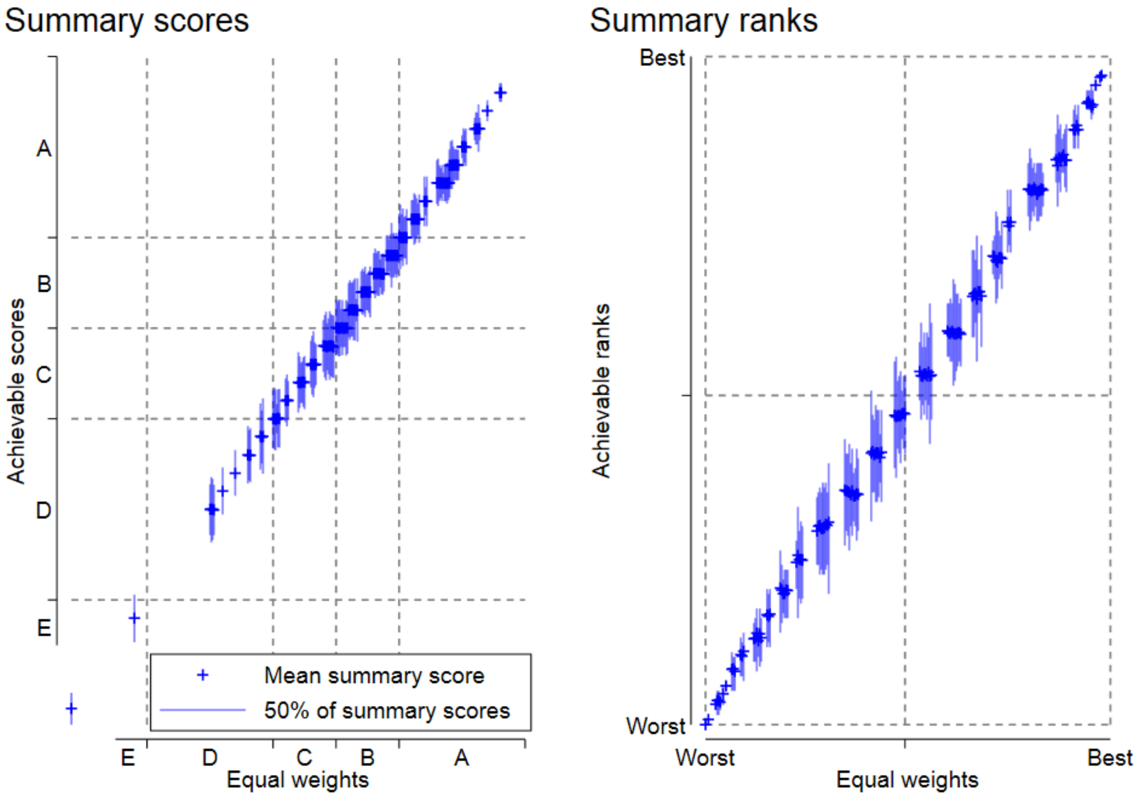


Figure 13. Scatterplots comparing performance on the weighting scheme least correlated with the current approach across the 10,000 simulations, and on eight other randomly chosen simulations, compared with the current approach (equal weights).

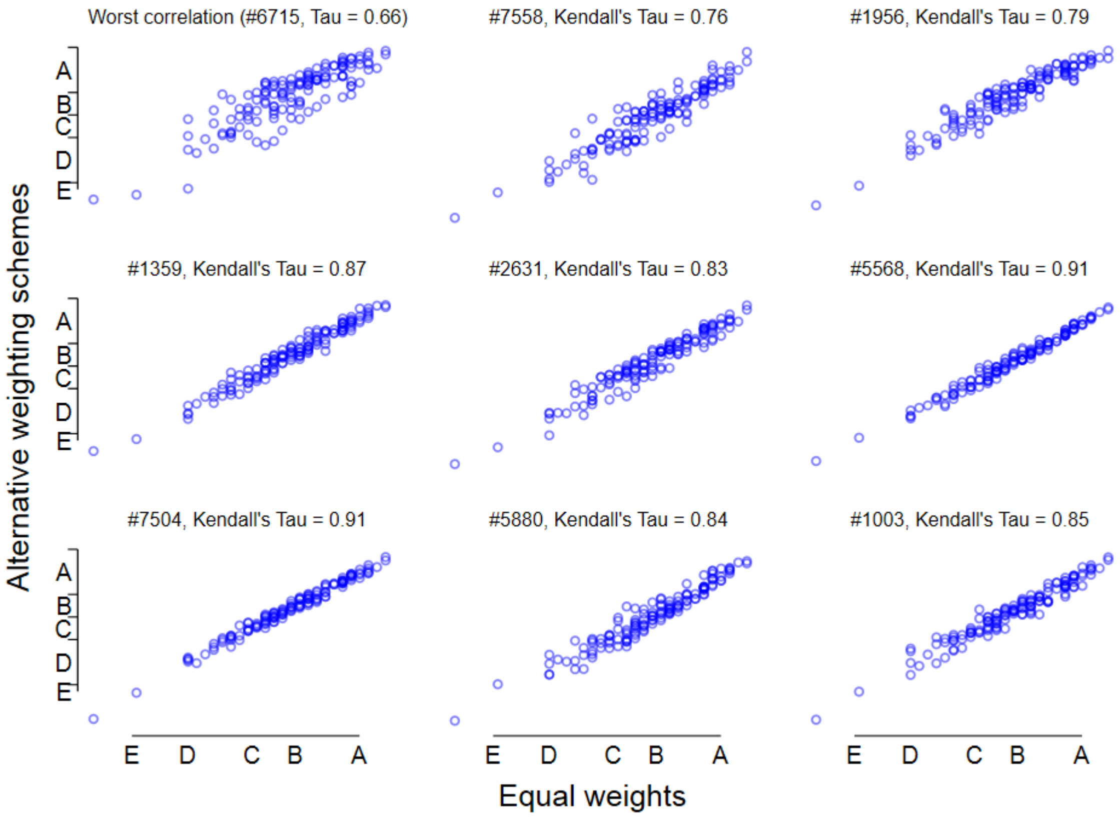


Table 18. Proportion of hospitals changing grade across all simulations, by grade on the current SSNAP composite indicator.

Current SSNAP composite indicator grade	**Total hospitals**	**Total Simulations**	**Number of times a hospital changed grade**	**Percentage of times a hospital changed grade**
Any	136	1,360,000	321,123	23.6
A	48	480,000	60,997	12.7
B	45	450,000	136,602	30.4
C	29	290,000	101,887	35.1
D	12	120,000	18,612	15.5
E	2	20,000	3,025	15.1

4.4.3 What if domains were designed to be empirically distinct?

SSNAP currently assign measures to ‘key indicator’ domains. Each of these domains aims to give an overview of one aspect of acute stroke care, but it is unclear whether these domains are empirically valid. By empirically valid, I mean that measures that are in the same domain should be highly correlated – because if they are not, this suggests that they do not all measure the same aspect of quality. If measures in a single domain measure two different aspects of quality, then the domain score becomes difficult to interpret. An average score in such a scenario could mean that performance on both aspects of quality measured within that domain is average, but it could also mean that good performance on one aspect is being cancelled out by poor performance on another.

I used exploratory factor analysis to develop empirically valid domains [125]. The existing SSNAP composite indicator is based on ten domains. Using exploratory factor analysis, I identified seven domains based on distinct latent factors in the dataset (Figure 14, Table 19). Some of the existing domains were similar to domains identified by the exploratory factor analysis (Figure 14). For example, all three of the ‘Stroke unit’ measures were assigned to the same empirical domain, although this domain also included another five measures from other current domains (Table 19, Figure 14). Overall, eight out of the ten SSNAP domains appeared to split into different empirical domains, suggesting that the domains used in the current SSNAP grade were not empirically distinct. For example, the four measures in the existing ‘Physiotherapy’ domain were assigned to three different empirical domains by the exploratory factor analysis (‘Receipt of occupational and physiotherapy’; ‘Identification of therapy need’; and ‘Time receiving therapy’, Figure 14).

Moving to domains based on empirically distinct latent factors would have some impact on SSNAP scores and levels. While SSNAP scores based on the current approach were clearly correlated with those using empirically-identified domains (Kendall’s Tau 0.71, Figure 15), it was apparent that when empirical domains were used average hospital performance was better than if the current domains were used. When using the current SSNAP domains, 48 hospitals receive a SSNAP level of A, while using the empirical domain this increases to 65 (Table 20). This is largely due to reclassification of hospitals currently receiving a SSNAP level B, with 17 (37%) of the 45 such hospitals reclassified as level A when using the domains from exploratory factor analysis (Table 20).

Figure 14. Sankey diagram showing individual measure flow from the ten domains currently used in SSNAP on the left to the seven empirically-distinct domains (the latter domains, on the right, are labelled arbitrarily).

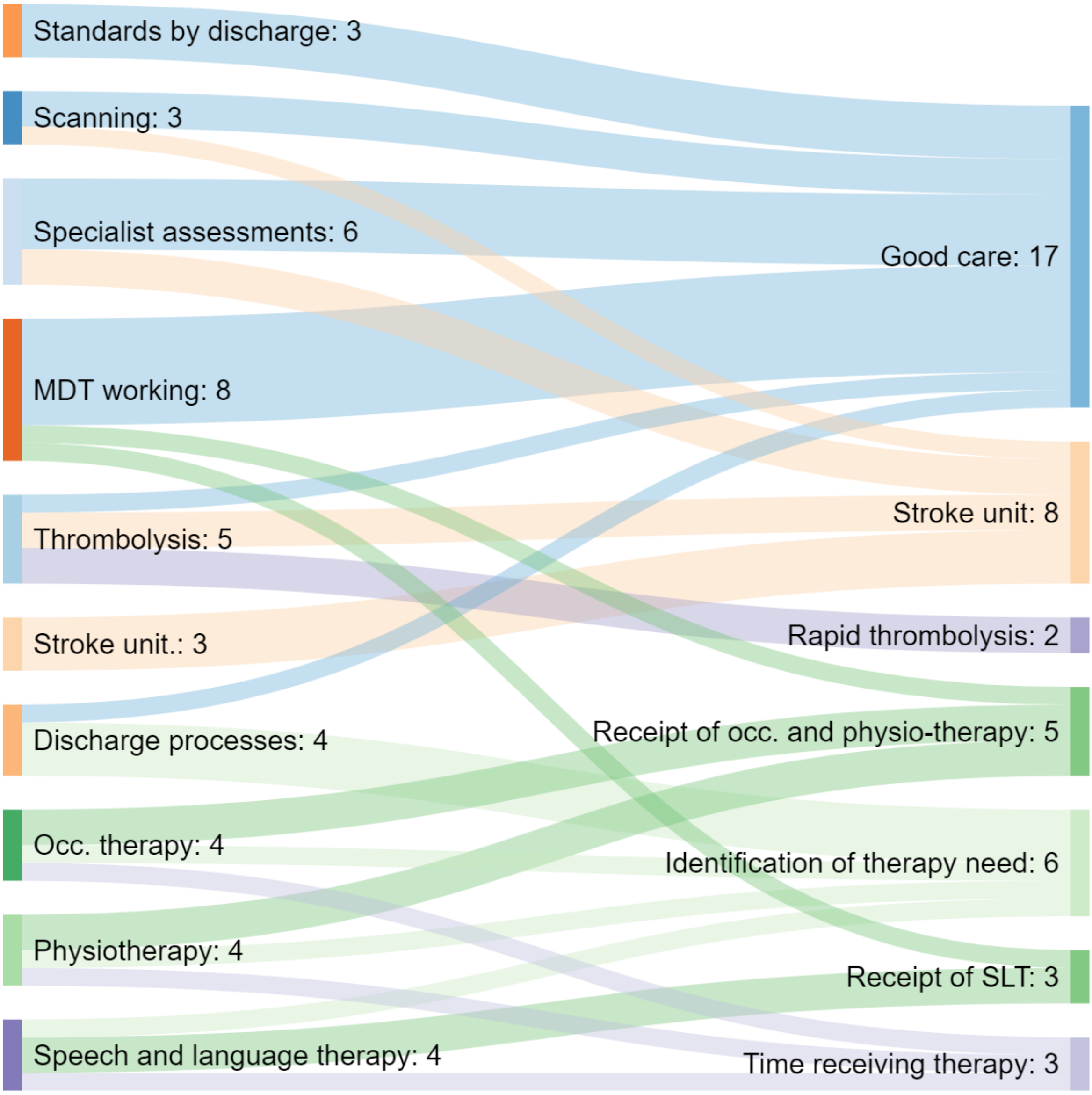


Table 19. Measures included in the SSNAP composite indicator, with current SSNAP domain and domains identified as more empirically coherent using exploratory factor analysis.

Measure	SSNAP domain	Empirical domain
% patients scanned within 1 hour	Scanning	Good care
% patients scanned within 12 hours	Scanning	Good care
Median time until scanned	Scanning	Stroke unit
% patients directly admitted within 4 hours	Stroke unit	Stroke unit
Median time until arrival on stroke unit	Stroke unit	Stroke unit
% patients spending at least 90% of stay on a stroke unit	Stroke unit	Stroke unit
% all stroke patients given thrombolysis	Thrombolysis	Stroke unit
% eligible patients given thrombolysis	Thrombolysis	Good care
% patients thrombolysed within 1 hour	Thrombolysis	Rapid thrombolysis
% applicable patients admitted within 4 hrs AND get thrombolysis	Thrombolysis	Stroke unit
Median time until thrombolysis	Thrombolysis	Rapid thrombolysis
% patients assessed by a stroke specialist within 24 hours	Specialist assessments	Good care
Median time until assessed by a stroke specialist	Specialist assessments	Stroke unit
% patients assessed by a stroke nurse within 24 hours	Specialist assessments	Good care
Median time until assessed by a stroke nurse	Specialist assessments	Stroke unit
% applicable patients given a swallow screen within 24 hours	Specialist assessments	Good care
% applicable patients given a formal swallow assessment within 72 hours	Specialist assessments	Good care
% patients reported as requiring occupational therapy	Occupational therapy	Identification of therapy need
Median minutes per day receiving occupational therapy	Occupational therapy	Time receiving therapy
Median % days on which occupational therapy is received	Occupational therapy	Receipt of occupational and physiotherapy
% compliance against therapy target for occupational therapy	Occupational therapy	Receipt of occupational and physiotherapy
% patients reported as requiring physiotherapy	Physiotherapy	Identification of therapy need

Measure	SSNAP domain	Empirical domain
Median minutes per day receiving physiotherapy	Physiotherapy	Time receiving therapy
Median % days on which physiotherapy is received	Physiotherapy	Receipt of occupational and physiotherapy
% compliance against therapy target for physiotherapy	Physiotherapy	Receipt of occupational and physiotherapy
% patients reported as requiring speech therapy	Speech and language therapy	Identification of therapy need
Median minutes per day receiving speech therapy	Speech and language therapy	Time receiving therapy
Median % days on which speech therapy is received	Speech and language therapy	Receipt of speech and language therapy
% compliance against therapy target for speech therapy	Speech and language therapy	Receipt of speech and language therapy
% applicable patients assessed by occupational therapist within 72 hours	MDT working	Good care
Median time until assessed by occupational therapist	MDT working	Good care
% applicable patients assessed by a physiotherapist within 72 hours	MDT working	Good care
Median time until assessed by physiotherapist	MDT working	Receipt of occupational and physiotherapy
% applicable patients assessed by a speech therapist within 72 hours	MDT working	Good care
Median time until assessed by speech therapist	MDT working	Receipt of speech and language therapy
% applicable patients with rehab goals agreed within 5 days	MDT working	Good care
% applicable patients assessed by all relevant specialists in a timely manner	MDT working	Good care
% applicable patients screened for nutrition and seen by dietician by discharge	Standards by discharge	Good care
% applicable patients with a continence plan drawn up within 3 weeks	Standards by discharge	Good care
% applicable patients who have mood and cognition screening by discharge	Standards by discharge	Good care
% applicable patients receiving a joint health and social care plan on discharge	Discharge processes	Identification of therapy need
% patients treated by a stroke-skilled Early Supported Discharge team	Discharge processes	Identification of therapy need
% applicable patients in atrial fibrillation discharged on anticoagulants	Discharge processes	Identification of therapy need
% patients discharged alive who are given a named person to contact	Discharge processes	Good care

Figure 15. Scatterplot comparing performance when domains were designed to be empirically valid against performance under the current approach. The left panel shows scores and the right panel shows ranks. The Kendall's Tau correlation coefficient applies for either plot.

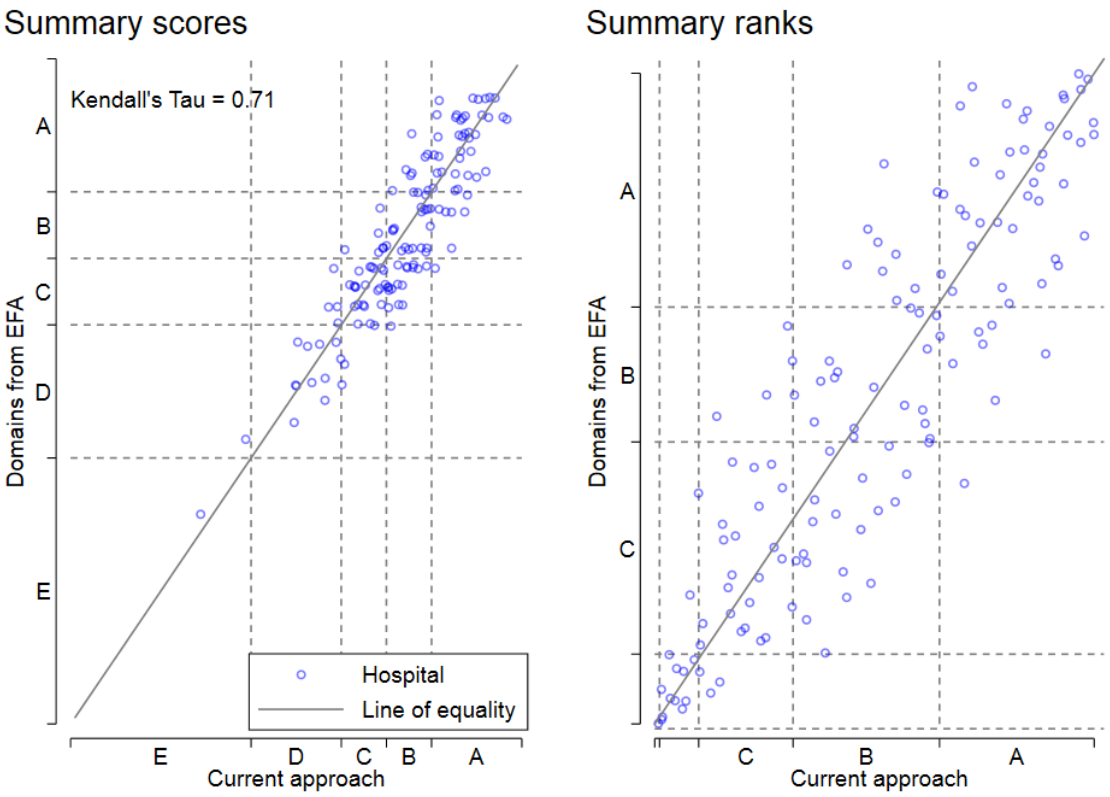


Table 20. Contingency table showing grade under current approach versus grade calculated based on domains from exploratory factor analysis. Zeros have been left blank.

		Grade under current approach					Row total
		A	**B**	**C**	**D**	**E**	
Grade based on domains from exploratory factor analysis	**A**	47	17	1			65
	B	1	25	16	1		43
	C		3	12	9		24
	D				2	2	4
	E						
Column total		48	45	29	12	2	136

4.4.4 What happens if a different standardisation approach is used?

SSNAP currently use an absolute approach to standardisation to ensure all individual measures are on consistent scales [69]. They use fixed reference points adapted from clinical guidelines to standardise measures are standardised to a 0-100 scale based [138]. However, while these reference points have remained unaltered since SSNAP's inception in 2013, the performance of hospitals on the performance measures included in SSNAP has improved substantially over time, resulting in a progressive increase of the number of hospitals with high scores, which are classified as the maximum level of A. When the composite indicator was first reported, for July-September 2013, no hospitals received a SSNAP Level of A, though 43% received the lowest level of E [8]. By July-September 2019, the data examined in this chapter, 71 of the 136 hospitals received an A grade and the majority of other hospitals received a B grade. For 14 of the 44 measures, the average hospital score is over 90 out of 100 (see Table 7 on page 67). It is possible that there are important differences in performance between hospitals in the highest performance category, differences that are masked using this potentially outdated standardisation scheme.

I assessed how adopting an alternative, Z-score-based (analogous to the current approach used by CMS Star Ratings), approach to standardisation affected the rank a hospital receives on the SSNAP score. I compared the ranks of hospitals under each approach graphically and using Kendall's Tau rank correlation coefficient. I did not compare differences in hospital scores per se because in general it is not possible to directly compare hospital scores calculated using different approaches to standardisation. Similarly, while it would be interesting to see how this might affect the SSNAP level, this rating system was built around 0 to 100 scores and cannot be applied to the results standardised using Z-scores.

Hospital SSNAP scores based on measures standardised using Z-scoring were correlated with SSNAP scores calculated under the current approach but for many pairwise comparisons the 'better' hospital would change (Figure 16). The Kendall's Tau correlation coefficient was 0.67, and some hospitals appeared to have large apparent changes in performance.

Comparing Z-score performance against the current SSNAP-assigned levels provided useful context. Performance of most hospitals appeared quite similar on summary scores produced using Z-scoring, with only two hospitals having summary scores outside the -1 to 1 interval (Figure 16). This similarity in Z-score performance between hospitals with different SSNAP levels shows that the distinction between different SSNAP levels is highly sensitive to the approach to standardisation, with performances that are assigned different levels under one approach appearing indistinguishable under an alternative approach.

Yet apparent similarities on the summary score hid differences on individual domains (Figure 17). By their nature, measures standardised using Z-scores will have mean 0 and standard deviation 1, as seen in Table 7. Hence the range of performance on individual measures when converted to Z-scores will tend to cover the range -3 to 3. The narrow range of Z-scored performance shown in Figure 16 occurs because for each hospital better than average performance in some domains tended to be balanced by worse than average performance in other domains. For example, Jersey Health Community Service appeared to be an outlier on both the ‘Standards by discharge’ and ‘Discharge processes’ domains, with Z-scores of -5.2 and -3.9 respectively. Yet even this extreme performance was partially balanced out by relatively average performance on domains such as ‘Physiotherapy’, where performance was near average with a Z-score of -0.1 – giving a final standardised summary score of -1.7 overall.

Figure 16. Scatterplot comparing performance when measures were standardised using Z-scoring against performance under the current approach. The left panel shows scores and the right panel shows ranks. The Kendall's Tau correlation coefficient applies for either plot.

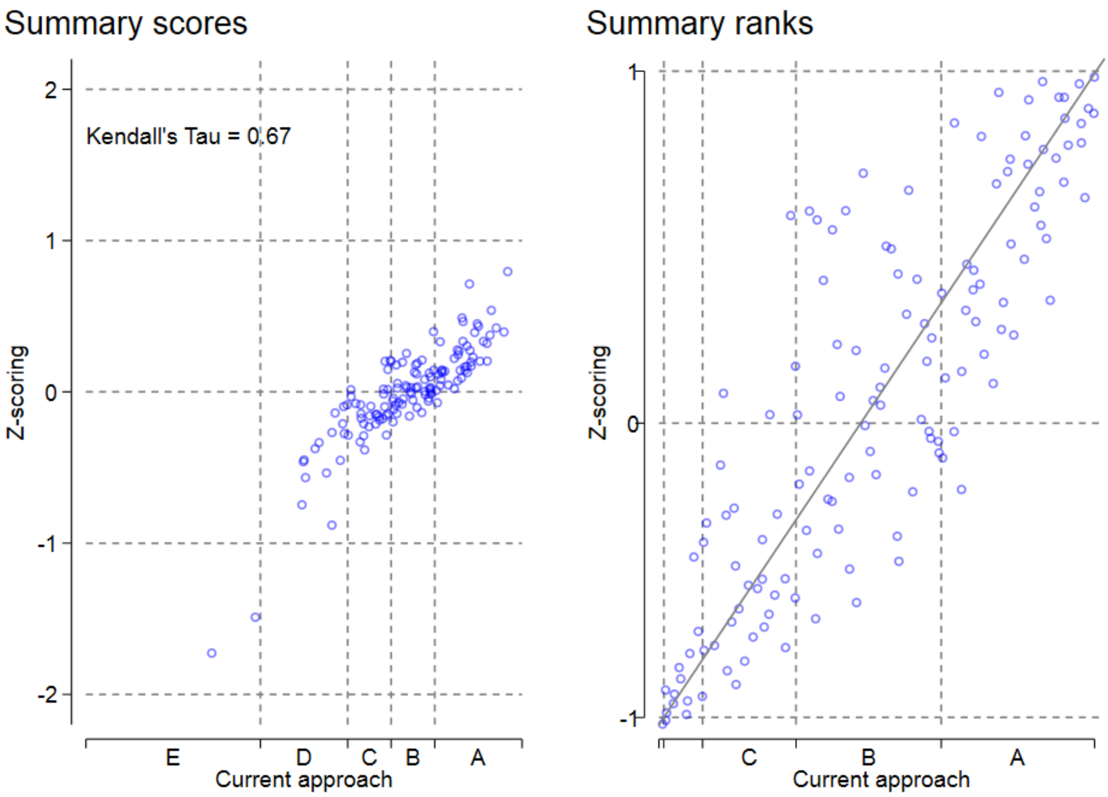
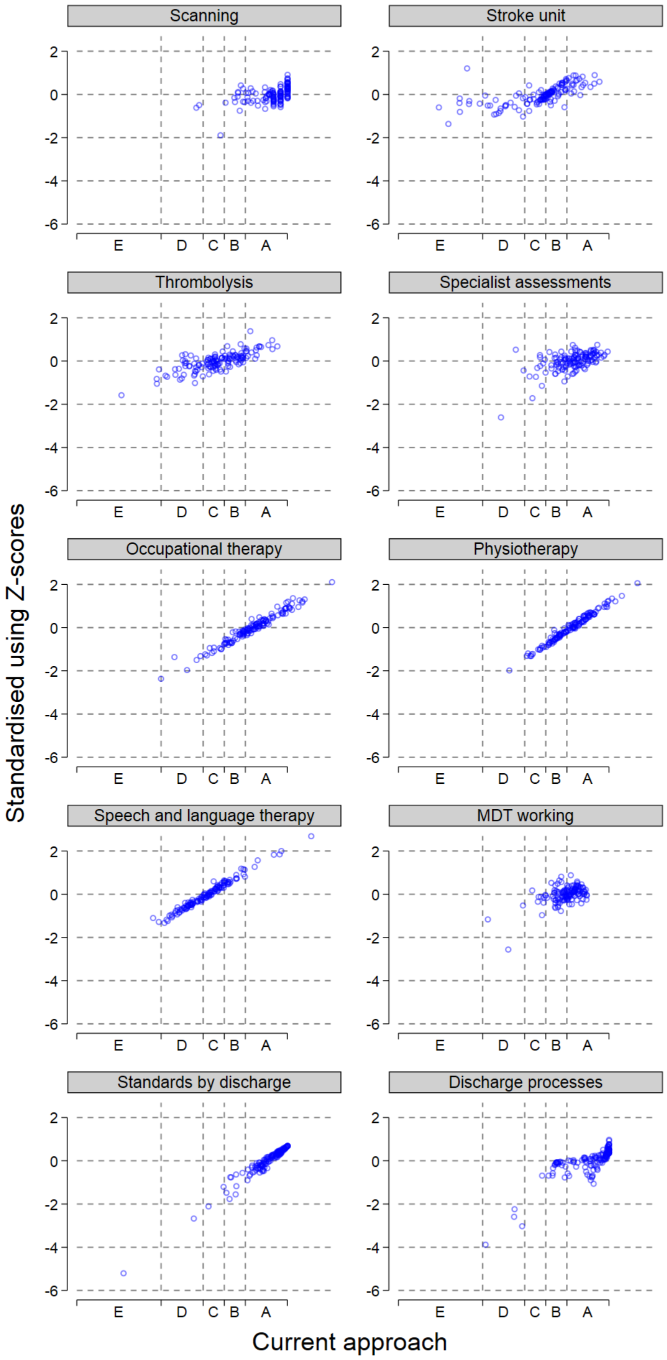


Figure 17. Scatterplot comparing performance on each individual domain when measures were standardised using Z-scoring against performance under the current approach. Axes are consistent between panels, ranging from +2 to -6 (the latter range allows visualisation of an extreme Z-score for the 'Standards by discharge' domain).



Note: For the different therapy domains, scores of over 100 are possible.

4.5 Discussion

This analysis has explored the sensitivity of the SSNAP score and level composite indicators to technical specifications that are plausible alternatives to those currently used. I found that the SSNAP score, the numeric summary of performance across the 10 SSNAP key indicator domains, was relatively robust to specific technical changes, including a range of plausible weight specifications examined through Monte Carlo simulation, with most alternative specifications giving results that were highly correlated with the current approach. However, the SSNAP level, which categorises performance on the SSNAP score into five letter grades ranging from A to E, was less robust. SSNAP levels B and C cover a relatively narrow range of scores (from 70 to 80 and 60 to 70, respectively, while other levels all relate to ranges of width 20-points or more). Hospitals currently receiving a level of B or C frequently could receive a different SSNAP level under plausible alternative specifications.

4.5.1 What happens if a different standardisation approach is used?

Relative performance of hospitals on the SSNAP score is little influenced by the approach to standardisation when this involves a change from a guideline/expertise-based absolute approach to a purely statistical Z-score approach (as for example used in the CMS Star Ratings scheme). Yet the current approach can give different SSNAP levels to hospitals that would receive identical scores if an approach based on Z-scoring was adopted, both overall and for several individual key indicator domains. On the Thrombolysis domain, for example, the lowest domain level under the current approach for a hospital with apparently typical performance (that is, a Z-score-based domain score of 0) was a D, while the highest was an A. This could be perfectly justified clinically, in that the apparent high performance on individual measures that is bringing hospitals that currently receive a D up to a Z-score-based domain score of 0 may not be as important for improving patient outcomes as the measures on which they perform less well. It is unclear whether this is the case, and one possibility is that SSNAP could improve reporting of its indicators by explaining why the current approaches to standardisation of individual measures remain appropriate.

While differences in performance between hospitals appear small, averaging into the overall domain score masks larger differences in some domains. This may in part reflect a ceiling effect, as, for many hospitals, the scope for further improvements on certain domains appears limited (see section 4.4.4, page 132). Though some hospitals have good performance across the board, for others a high overall SSNAP level masks some middling or poor performance on individual domains. Developers may therefore wish to consider whether it is appropriate to exclude domains where all hospitals are performing well in the calculation of the SSNAP score and level, or to devise new domains or scoring systems that distinguish between the performance of hospitals that are currently rated similarly. Such a decision would need to be considered carefully, as clearly it is important not to disincentivise high performance on important measures for the sake of having measures that distinguish between hospitals.

4.5.2 What happens if domains are designed to be empirically distinct?

Using empirically distinct domains leads to greater apparent spread in performance on the SSNAP score, with several hospitals moving from ‘A’ to ‘B’ grades for example. Yet moving to empirically distinct domains does not have a substantial impact on relative performance of organisations. If one hospital appears better on the current SSNAP score, it will tend on average to appear better on a hypothetical score based on empirically distinct domains.

This suggests that, while the impact of domain grouping on the score is relatively small, the grouping of several empirically distinct measures together into one domain means that the composite indicators produced by SSNAP are less effective at distinguishing hospital performance than they could be. It appears that some hospitals may be, in effect, compensating for poor performance in some areas with better performance in other areas. But the fact that these areas are grouped into the same domain means that this is not apparent in the current ratings assigned to hospitals.

4.5.3 What happens if domain scores are not rounded before being combined into the overall summary score?

Preliminary rounding of domain scores before combining into the overall score does not have a substantial impact on the SSNAP Score, although results with and without the rounding are not perfectly correlated. The specific approach to rounding appears to reward hospitals with ‘A’ grade performance in any domain. Removing the rounding increased the number of hospitals that would be graded ‘B’, and at the same time decreased the number that would be graded ‘A’.

The decision to apply preliminary rounding appears unusual and lacks clear justification. Yet similar approaches are seen occasionally in composite indicators, with several UK examples. One such example was the CQC Intelligent Monitoring composite indicator, which was used to summarise a wide range of quality and safety information available to the

English Care Quality Commission [3,178], and was intended as one tool for prioritising hospitals for inspection [47]. In the calculation of the Intelligent Monitoring indicator, individual measures were standardised using Z-scores, but then these Z-scores were converted to points such that hospitals received scores of 1 if they were between two and three standard deviation worse than the mean, a score of 2 if they were over three standard deviations from the mean, and a score of 0 otherwise.

Use of rounding such as this can readily be criticised [16], based both on statistical literature detailing drawbacks of such artificial categorisation (see, for example, [78]) and on a specific discussion of the increased instability in resulting composite indicators that can be caused by such rounding [10]. The exaggeration of extreme performance caused by rounding 'A' grades to a score of 100 (and 'E' grades to a score of 20) is one example of additional instability – improving a domain score from 79 to 81 before rounding leads to a change in the rounded domain score from 70 to 100. Improving individual domains from 'B' to 'A' (increase of 30 points) has more impact on the summary score than changes from 'C' to 'B' (increase of 10 points), 'D' to 'C' (increase of 20 points), or 'E' to 'D' (increase of 20 points). This may encourage hospitals to game the indicator by focusing on turning acceptable performance on some domains into excellent performance, at the expense of domains where performance is truly poor [67,177]. This incentivises perverse behaviour [67,84].

4.5.4 What happens if different domain weights are used?

Changing domain weights influences the apparent performance of hospitals on the SSNAP composite indicator, but in general has little impact on their relative ranking. While different weight specifications were typically highly correlated with scores produced under the current SSNAP specification, changing domain weights frequently led to reclassification of hospital performance (for example, A-grade hospitals being reclassified as B-grade or C-grade).

My use of Monte Carlo simulation, covering a wide range of randomly chosen domain weights, allowed for more complete description of the sensitivity of the reported hospital performance to the choice of weights than single-scenario sensitivity analyses. This reassured that the apparent low sensitivity to the choice of weights highlighted by the individual comparison made was not simply a chance occurrence, but that almost any alternative choice of weights would produce results that are highly correlated with those of the current specification.

4.5.5 Applications of Monte Carlo simulation to composite indicators

There are few descriptions of Monte Carlo simulation in the sensitivity analysis of composite indicators of healthcare quality [49,95,175]. One of these describes the impact of choice of weights [95], similar to the application in this paper. But Monte Carlo simulation could be used more widely to examine uncertainty introduced by other decisions [49], or to examine the impact of simultaneously perturbing multiple technical specifications as I did for the CMS Star Ratings in Chapter 3. Setting up broader simulations may be challenging because of the difficulty of thinking through the appropriate space of decisions to evaluate, but in principle Monte Carlo simulation could be used to propagate uncertainty about any choice involved in turning available data into the final ranks hospitals receive.

Monte Carlo approaches could also be used to produce composite indicators that explicitly describe the uncertainty in the methods used to create them. Hota and colleagues discuss calculating a ‘summary’ composite indicator of hospital quality by finding all existing composite indicators and taking, in some form, the average [107]. Developers of composite indicators could take a similar tack, using Monte Carlo simulation to create hospital rankings that incorporate methodological uncertainty. For example, a representative survey may be used to elicit appropriate weights for each domain. A standard approach would then use the mean weight derived from the survey, but in fact there is sampling uncertainty around this weight. A Monte Carlo simulation could be used to derive the range of plausible ranks for each hospital based on the range of plausible weights for each domain. This is by no means a novel idea: Saisana, Saltelli and Tarantola give an example of the use of Monte Carlo simulation to produce a composite indicator of aspects of economic development that incorporates methodological uncertainty [127], while Schang and colleagues proposed ranking intervals for measuring hospital performance [79]. Such ranking intervals show the best and worst rank that a hospital can achieve under a set of assumptions. While the specific details differ (in that Schang and colleagues consider possible ranks under a known set of possible measure weights), they are analogous to the presentation of the results of the Monte Carlo simulation shown in Figure 12.

4.5.6 Comparison with the CMS Star Ratings

The SSNAP score and level are more robust to the exact technical specification than the CMS Star Ratings. For example, the probabilistic sensitivity analysis of the weights used to combine domains in the CMS Star Ratings showed that this could reasonably lead to substantial changes, with differences on the order of one quarter of possible ranks being relatively typical (see Section 3.4.5). In contrast, for the SSNAP score, typical differences between weight specifications were about one eighth of the possible ranks, i.e. half the size of what was observed for the CMS Star Ratings (Figure 12 on page 125). These smaller differences were despite the analysis of the Star Ratings drawing from distributions centred on the current CMS weights, while the analysis of the SSNAP score drew weights from a uniform random distribution. The SSNAP level was also more robust than the actual star ratings assigned in the CMS Star Ratings, with perturbation of weights leading to a hospital being assigned a different rating one quarter of the time for SSNAP compared with 40% of the time for the Star Ratings.

While the SSNAP score and level was less sensitive to the choice of the weights used to combine domains than the CMS Star Ratings, the picture based on grouping of measures into domains and standardisation of measures was more mixed. It appeared that ranks on the CMS Star Ratings were a closer match to those based on domains derived from exploratory factor analysis than ranks on the SSNAP score, with a Kendall's Tau of 0.8 compared with 0.7; both reflect a high degree of concordance. The alternative approach to standardisation I applied to the CMS Star Ratings had a far greater impact on apparent hospital performance than the alternative approach I applied to SSNAP, with a Kendall's Tau of 0.4 compared with 0.6. But this is not a fair comparison in the technical sense as my alternative standardisation approach for SSNAP was not the same as my alternative approach for the Star Ratings.

4.5.7 Performance rankings vs performance ratings

Recently, Bae, Curtis and Hernandez argued that shifting from performance rankings to performance ratings would be a helpful step for avoiding false precision in composite indicators of hospital quality and safety [179]. My results argue that such a shift would not necessarily be helpful. Rankings on the SSNAP score appeared robust to the precise technical specification, but the performance ratings provided by the SSNAP level did not. While performance ratings may be more actionable, without careful design they may be more misleading than a simple ranking. Technical sensitivity analyses of the type used in this paper should form a key part of designing robust performance ratings.

4.5.8 Conclusion

The SSNAP score, an indication of the quality of stroke care of English hospitals, is relatively robust to technical choices relating to how measures were standardised, grouped and combined – in contrast to the CMS Hospital Compare Star Ratings discussed earlier. Yet while the SSNAP *score* appears robust, the SSNAP *level*, which grades performance based on the SSNAP score, is far more sensitive to specific technical choices, and may not provide a useful guide to underlying hospital performance.

One key difference between the (less robust) CMS Star Ratings and the (more robust) SSNAP composite indicators was the amount of missing data: almost all US hospitals with a CMS Star Rating were missing some performance measures, and many were missing one or more domains of quality entirely. But in the SSNAP dataset I used, only one hospital had any missing measure information, and domain scores could be calculated for all hospitals. My analysis of the CMS Star Ratings showed that hospitals with missing scores for one or more domains of quality often appeared more sensitive to the technical specification, which might suggest that the greater degree of robustness of the SSNAP score was partially due to the higher completeness of the underlying data.

I have shown that the CMS Star Ratings scheme and the SSNAP indicators are both sensitive to plausible alternative choices made in their specification, but the impact of the same choices seems to be different. This suggests that the findings are likely to be indicator-specific, and dependent on both the design features of an indicator and the actual data to which it is applied. The proportion of missing data is an important component of the data structure that typically varies between different contexts/data sources/eras and clinical areas. The findings should be considered as directly relating to the examined indicators, but generalising to other schemes is not prudent, though approach used will be of paradigmatic relevance across different indicators.

Uncertainty and sensitivity analyses of performance ratings based on composite indicator scores are a useful tool for understanding whether hospitals that receive different ratings are likely to have important differences in performance. For those interpreting composite indicators, it is useful to know, for example, that A-rated hospitals tend remain rated as an A under alternative specifications, but that the difference between hospitals rated B and C is easily reversed by changing weights or by regrouping measures. Developers of composite indicators can also apply these tools, using them to design ratings that are robust to differences in indicator specification.

This chapter, and the chapter on the CMS Star Ratings, set out to address my aim of characterising relevant challenges in the design of composite indicators. The understanding of the importance of the technical specification is also helpful toward my second aim, exploring how to improve reporting. The next chapter primarily addresses this second aim, describing the findings of my qualitative interview study with experts on quality measurement.

5 Results 3: Qualitative interview study of experts’ views on good practice in developing composite indicators

5.1 Summary

As the analysis presented thus far in this thesis has demonstrated, the decisions taken when designing a composite indicator have multiple implications and consequences. In particular, taking a different but still reasonable approach can in some cases change the apparent performance of a hospital from among the best to among the worst on the composite indicator. This suggests that much improved reporting of the technical choices involved in the design of indicators is needed to support transparency. However, how to ensure high quality approach to reporting of the development of composite measures is not straightforward. While there is some literature on processes for developing individual performance measures in healthcare, much less addresses how composite measures are developed. To address this gap, I conducted qualitative interviews with purposively chosen experts in quality measurement including clinical experts, commissioners and methodologists. Drawing on my own professional expertise as a statistician and understanding of the literature, I synthesised the findings into a general framework for developing a composite indicator, highlighting the main challenges at different stages and common approaches to address these challenges.

Analysis of the interviews generated nine important themes. Some of these reflected the technical concerns discussed in earlier chapters, but the most important reflected conceptual issues such as the need for development to be purpose-led. The results of this study are one step toward developing reporting guidelines for composite indicators. Yet challenges remain, particularly regarding how to determine a comprehensive view of issues in developing composite indicators and how to develop a reporting guideline for an explicitly iterative

process. Further work is still required to develop reporting guidelines, but the results of the study provide a useful tool for critical appraisal of existing indicators and are a useful starting point for developers planning a new composite indicator.

5.2 Introduction

As discussed throughout this thesis, composite indicators, despite their ubiquity, frequently suffer problems that limit their usefulness [16,65], not least because, as my analyses in the previous chapters have demonstrated, organisational performance on composite indicators is sensitive to the specific technical choices in their design. But the problems with composite indicators go beyond purely technical issues: lack of transparency is a particular problem for many existing composite indicators. One key way of improving transparency and supporting better design involves reporting guidelines [180–182].

Developing a reporting guideline typically requires first identifying and characterising the key issues in a given methodological field, and then using this as the basis of a wider consensus-building process among domain experts [148]. One example is the development of the Template for Intervention Description and Replication (TIDieR) checklist, a reporting guideline intended to support better description of interventions used in medical research. Development of TIDieR began with a review of existing CONSORT checklists [183,184], checklists for specific intervention types, and other relevant literature. The committee produced a list of 34 potential checklist items based on this understanding of existing knowledge. These potential items were then prioritised in a two-round modified Delphi study. Items that scored highly in the Delphi study were included in the draft checklist, and decisions over the inclusion of items with more moderate ratings were made in a consensus meeting with far fewer participants than the Delphi study. The resulting checklist was then piloted to ensure usability, leading to clarification and elaboration for some items.

Understanding the perspectives of different stakeholders is important in identifying and characterising the decisions that need to be made in developing composite indicators. No study of these views is currently reported in the literature. I sought to address this void through a qualitative study. Specifically, I sought to identify the range of choices to be made when developing and reporting composite indicators by interviewing experts in the design of composite quality indicators, and of quality indicators more broadly. I was keen that participants would include clinicians, quality improvement experts, statisticians, epidemiologists and data analysts to ensure that the findings reflect a more complete set of perspectives, rather than being limited to one set of methodological concerns.

5.3 Study aims and objectives

The primary research question was: what are the views of experts on the key decisions in the development and reporting of composite indicators of healthcare quality or safety?

The objective was to identify and characterise the range of choices involved in producing composite indicators, to inform the future development of a reporting guideline.

5.4 Summary of methods

The study design was a qualitative interview study with international experts.

5.4.1 Interview participants and approach

Interview participants were purposively sampled. Initial participants were identified from relevant literature or via professional networks, with subsequent waves identified through additional searches and snowball sampling. Participants were invited to participate if they were known to have had a leading role in healthcare improvement. They could have either a clinical, data analytic, statistical, or managerial background. At least one participant identified themselves as having expertise in each of these areas, with many participants having multiple forms – for example, both a clinical background and managerial experience. All interviews were carried out over the telephone.

Recruitment was planned to include between 12 and 20 participants. In the end, 14 participants were recruited with the 15th participant dropping out due to professional commitments relating to the COVID-19 pandemic. Fieldwork started in August 2019 and ended in February 2020. The sample size was adequate [159], with interviews being dense and consisting of a strong dialogue between an informed interviewer and an expert participant. I did not undertake a formal test for theoretical saturation; I instead used the principle of “information power”, which indicated that I had achieved sufficient range and depth of views [159]. The average interview duration was 48 minutes, ranging between 26 and 87 minutes.

5.4.2 Structure of interviews

Interviews aimed to identify what each participant felt was most important in indicator development. Participants were asked to discuss an ‘example’ area in which they had the relevant expertise to think through what issues were important, talking through developing a composite indicator from their first steps through to reporting the actual indicator.

Interviews used a semi-structured prompt guide (see Appendix 2). These prompts focused on how participants would choose to go about developing a composite indicator, addressing how they would start the process, the issues they might expect to encounter, methods to overcome challenges, and the types of people they would choose to involve in the process. Interviews did not always hew to this guide closely, as I felt it more appropriate to explore interesting statements as they arose than to follow the list of questions to the letter.

5.4.3 Analysis of interview data

The framework method was used to analyse interview data [160]. The initial framework was based on the problems originally identified in chapter one of this thesis. Interviews were initially coded against this framework. The framework was iteratively revised over the course of analysis to produce the final categorisation [163]. This iterative revision involved adding new categories (where important ideas generated by analysis of interviews were not captured in the existing framework), combining existing categories (where distinctions between categories appeared less important), and splitting categories into two (where there appeared to be an important distinction between different themes initially coded into the category).

5.5 Results

Semi-structured interviews were conducted with 14 experts with experience in the development of composite indicators in healthcare. These participants were from a range of different professional backgrounds, and all worked either in the UK or the US. Table 21 briefly summarises each participant’s expertise.

My analysis enabled generation of nine broad themes to organise the interview data. Three of these themes cut across the whole development process for composite indicators:

- Purpose-led development.
- Iterative development.

- Competencies involved in developing a composite indicator.

Six themes covered types of decisions to be made in developing indicators:

- Identifying domains of quality.
- Identifying individual quality measures for each domain of quality.
- Developing final domains of quality and the final set of measures.
- Standardisation of individual measures.
- Combining domain scores into the summary score.
- Reporting of composite indicators.

Here, I present each of these broad themes, discussing the issues raised by the participants.

Table 21. Summary of participants' expertise.

Participant	Expertise
P01	Healthcare informatics and data analysis.
P02	Epidemiology. Clinical medicine.
P03	Health services research and statistics.
P04	Manager and data analysis.
P05	Quality improvement. Clinical medicine.
P06	Data analysis and information.
P07	Health services research. Clinical medicine.
P08	Healthcare informatics and data analysis.
P09	Data analysis. Managing development of indicators.
P10	Quality improvement and health services research.
P11	Health policy.
P12	Clinical audit. Clinical medicine.
P13	Quality improvement. Clinician.
P14	Data analysis.

5.5.1 Purpose-led development

All 14 participants emphasised the need for clarity about the purpose of a composite indicator; eight mentioned specifically that identifying the purpose of a composite indicator was one of the most important steps in the development process. Participants viewed a full understanding of the intended purpose of a composite indicator as critical because it informs many other decisions in the development process. This included closely-related issues such as the relevant aspects of quality, but participants also viewed the understanding the purpose as helpful when addressing many other issues including deciding who to involve in the development process, how best to report the results, and even technical aspects of the design of the indicator. This foundational importance led many to consider the purpose of the indicator as the first question to address when designing a composite indicator:

> "The first thing I'd bear in mind is the question: who's it for and why? What is the ultimate aim of generating some kind of performance scores? Because I think that informs quite a lot of the design […]. If you want to produce […] a new sort of national statistic that's […] mainly aimed at a technical audience […] then that will take you down one particular design pathway. Whereas if you want something that is more practical or simple or more easily understood by people without a quantitative background or a clinical setting, that might take you down another design path." (P02).

> "[One] purpose might be in developing, and evaluating, the care provided by a particular entity, and then often quite relatedly to that, attaching financial incentives for the quality provided. And I think that the use involving summarising information for patients, has somewhat different considerations than what you might do if you were only trying to evaluate, although some of the considerations are common, and what we wind up doing is typically forming a single set of composites for both purposes." (P03).

Participants discussed many aspects of the purpose of composite indicators, ranging from broad issues such as the audience the indicator is intended to serve to the precise details about the aspects of quality it measures and the clinical conditions it includes. My analysis of the interview data suggests that the purpose of a composite indicator can be distinguished into two components: 'Why is a composite indicator being developed?' and 'What is the composite indicator being developed to measure?".

5.5.1.1 Why is a composite indicator being developed?

As I discuss in turn below, participants discussed four primary aims served by composite indicators in healthcare:

- Facilitating quality improvement by clinical teams
- Motivating quality improvement by raising awareness of a healthcare quality or health system problem
- Enabling patients to make more informed choices about their care
- Enabling performance-related financial incentives to organisations

These broad aims were viewed as having an important impact on subsequent design decisions, ranging from the types of individual measure to include

Facilitating quality improvement by clinical teams. One of the main uses of composite indicators described by participants was in summarising performance against a set of quality improvement priorities. Participants tended to discuss the importance, for quality improvement, of working with individual measures when planning local work to improve quality. Yet they raised that, in practice, clinical audits may use so many individual measures that it is challenging to understand overall performance, and discussed the use of composites in quality improvement as providing a useful overview of a complex set of performance measures:

> "It would be sort of maddening to see how you compare to 2000 other hospitals on 60 different measures. Some you're better, some you're worse, some you're average, that is a little crazy. So having a composite in that situation gives you some information about how you're doing as a whole, when you pull together all those measures." (P05).

However, while having a composite indicator to summarize performance was viewed as useful, participants felt that for quality improvement purposes it was vital that this summary could be disaggregated back into the individual performance measures. Without being able to dig into the individual measures, they noted that it was challenging for hospitals to identify the precise areas where quality improvement would be most important:

> "They did have the composite indicators [...]. If you weren't very good at that, if you came out badly on it, the first thing you had to do was unpick it [...]. You couldn't act on them in any way, until you'd unpicked them, and that meant a lot of work for us to

> unpick what they'd done, and to work out which bit of it was creating the problem." (P08).

Enabling patients to make more informed choices about their care. A second common purpose for composite indicators raised by participants was as a tool for helping patients make informed decisions about where to seek care. They viewed this use as motivated by the perceived difficulty of interpreting a wide range of individual performance measures, and of understanding what a particular performance on each measure actually means. Composite indicator were viewed as more useful and more interpretable to those choosing where to seek (typically elective) care than a collection of individual performance measures. But participants also noted that the design of a composite indicator intended for helping patients to make informed choices about their care might differ from the design of an indicator intended for use in other purposes, such as quality improvement:

> "It's important to remember that quality improvement is vitally important, and there are registries and other publicly and private reporting efforts that are useful for that but that if the [...] application is [...] patient decision support then a whole different set of considerations has to be made, and including different measures may make sense [...]. Something like volume, for example, [...] is a powerful predictor of outcomes, independent of the historical outcome performance that a hospital has, and so we include it [...]. From a hospital standpoint it may not be that useful in quality improvement but it's important for patient decision support." (P04).

Motivating quality improvement by raising awareness of a healthcare quality or health system problem. A third reason raised by participants for developing and reporting a composite indicator was achieving impact with policy-makers, local healthcare management, patients and the public. Composite indicators were seen as more able to reach audiences such as the news media and could more easily impact the decision-making process of hospital boards than a collection of individual performance measures. Composite indicators could help focus policy attention on specific issues, leading to quality improvements or an additional national focus on an issue:

> "We are a voluntary agency and we do have enough data to give a credible letter grade, that experts will agree as well, and so we issue the grade. And we do A, B, C, D and F. We do all five categories, and it gets a lot of press attention. We update it

> every six months, and every six months it gets a tonne of attention. So, we have found it to be very successful in getting hospitals to pay more attention to their safety.” (P11).

Participants highlighted that composite indicators with serious technical or conceptual limitations could still be achieving useful policy impacts. Participants felt that providing simple summaries of performance on specific issues could lead to increased attention on the relevant issues, perhaps leading to improvements in the quality of healthcare, regardless of validity. For example, when one organisation stopped producing a composite indicator of hospital quality, they were met with the objection:

> “ ‘Why did you stop doing that? You had our attention. It was really hard to start with, but we all got into it and actually it because a bit of an event for us. We quite liked it and it got new people interested in these kinds of issues in our trust.’ ” (P14).

Enabling performance-related financial incentives to organisations. The final main use of composite indicators discussed by participants was in pay-for-performance schemes. Participants noted that pay-for-performance use led to certain issues becoming more of a concern. ‘Gaming’ of pay-for-performance schemes was a common concern, with participants discussing the need for careful design to avoid perverse outcomes:

> “GPs are paid for doing well on the [Quality and Outcomes Framework] QOF, and they’re not paid for doing well on stuff that isn’t on the QOF. [...] When something’s added into QOF, they all rapidly improve on that. It probably means that something else is being dropped. [...] You can call it gaming, they are trying to achieve their 95 per cent or whatever it is they need, to get their points and get paid. But, in a way, isn’t that the point anyway?” (P08).

As discussed, participants felt that the broad aims of a composite indicator had an important impact on subsequent design decisions and on the desirable features of a composite indicator. But many participants also viewed it as very important for developers of composite indicators to carefully consider what, exactly, the composite indicator was intended to measure.

5.5.1.2 What is the composite indicator being developed to measure?

Participants discussed the importance of identifying exactly what the composite indicator is intended to measure, particularly in terms of the detail underlying high-level concepts such as ‘quality of cardiovascular care’ or ‘quality of surgery’. There were two important aspects to this. The first was being clear about the perspective the composite indicator will take, linking back to the reasons why a composite is being developed. The second was whether concept of the composite itself made sense.

By ‘perspective the composite indicator will take’, I mean identifying the group of people who are expected to take action based on the scores a composite indicator shows. This is closely linked to the reasons why a composite is being developed discussed above. Participants discussed this perspective as important for almost all technical steps in developing a composite, ranging from selection of individual measures to the reporting of the final score, because the composite indicator is intended to be useful to this group.

> “If you’re looking at cardiovascular care from a population point of view, as a sort of performance measure for the commissioning body say, then that’s one perspective which covers a fairly wide range of providers’ performance, which you are responsible for as a commissioner. But, if you’re looking at it from a particular aspect of the service, like primary care or secondary care, you’ve got a bit of a narrower influence. [...] With the individual provider, you’d look at the pathway essentially from what data you’ve got for the patient when they arrive.” (P08).

> “The first question that we always asked was are we interested in understanding the organisational performance directly or are we interested in understanding the effect on patients and populations. And I think [...] it’s possible to do both and then start to combine them, but it actually turns out to be a really critical first question.” (P07).

> “What's the unit we want to look at? What are the factors that might affect quality? How many of them can we narrow down? What's in the gift of a service, and what's external to that gift? And obviously we do that because we're also making judgments [...]. We need to try and align any measurement with that, so that we're...obviously for a provider that we assess they'll say well, that's not a fair judgement of us because that's outside of our gift.” (P14).

A common concern was that a composite could be formed from a set of effectively unrelated measures, leading to a final composite indicator that would not have a meaningful interpretation. This could happen in practice because issues that superficially appear related may not in fact relate to the same issues.

> "We had done this composite of performance of elective surgery for 25 maybe different surgeries, and from looking at all that data it became clear that I didn't feel like it was actually fair or useful to make a composite. So I think you have to start looking at things, is this going to help anybody. Just don't make a composite to make a composite, right? So I do think you have to start wondering whether it's a valid thing to do. Especially you think about surgical lines, you can have hospitals that do really well in certain areas and not well in others, and how is that useful to make into a composite." (P09).

> "With the national expert panel, we spent a lot of time talking about what is the construct that we are hoping to capture with this composite score. And this is where I get myself into trouble. I think others may have not spent as much time being as thoughtful in terms of being very clear on the construct that is trying to be measured." (P10).

Despite the emphasis on purpose, participants often noted that, at the beginning of a development process, the precise details of the purpose might be unclear. This meant that there needed to be flexibility in the development process to accommodate an improved understanding of what was being measured. Being able to iteratively refine the purpose of the indicator was viewed as a key part of a good development process.

5.5.2 Development as an iterative process

Participants emphasised the iterative nature of the development process for composite indicators. Though it is natural to list a series of decision points considered as a sequence of steps, in practice participants explained the process was not so simple.

> Participants' examples of iterative development wereoften were reactive, addressing serious problems that had only become clear when development was already reasonably advanced. "You could talk theoretically about how you would combine things, but then you'll find that you have only an eighth of the data that you really want. So you're missing a bunch of domains, so now you're in trouble because you can't weight the domains the way you want to because you're just missing a ton of data, and then you're hit with the reality that you have maybe, like I said, an eight of the data." (P09).

> “We convened a multi-stakeholder expert panel to give us really a top to bottom review […] to actually ensure ourselves that we still believed the domain framework was even correct. So that type of really stepping back and […] first of all defining what is the purpose of the report card or the composite, and then saying […] do we understand the composite as a concept and can we reliably or at least in some valid way translate a set of indicators into a composite representing that concept. So that’s probably the most difficult part of this.” (P07).

Several participants discussed active approaches to iterative development that aimed to identify and address problems as they arose rather than waiting until problems became so serious that wholesale changes to the composite indicator needed to be made. These active approaches typically involved producing prototype composite indicators. Such prototypes were viewed as helpful for understanding the impact of technical design choices, but they were also seen as a valuable tool in consulting with users. Participants often felt it was easier for patients and other intended users of composites to identify improvements to a prototype that they can see than it was to get detailed feedback over abstract issues:

> “We were asked to […] develop a composite indicator on [the workforce race equality scheme]. […] We developed a set of indicators and we’re using them now, which looks at variation within a trust between staff from BME groups and white staff. […] And we came up with some models where we could combine those. And interestingly, when we presented that people said ‘Oh no, we don’t really want this, because the risk is it’s covering up some variation that we should be worried about.’ […] We didn’t continue with a composite.” (P14).

5.5.3 Identifying domains of quality

Participants often discussed the importance of thinking about the aspects of quality that needed to be captured by the composite indicator. These discussions contained two distinct themes. The first was about identifying all the aspects of quality relevant for the composite indicator, before using these aspects to identify individual performance measures. The second theme was about revisiting these initial domains once individual performance measures had been identified, to ensure the domains were appropriate and that the set of performance measures was balanced.

This section discusses only the first of these themes, the importance of thinking carefully about the aspects of quality that need to be captured prior to setting out to identify individual measures. The second theme is discussed in section 5.5.5; in practice it may entail very different issues and I felt it was helpful to discuss selection of individual performance measures before discussing ways that this set of performance measures may need to be processed.

A frequent concern mentioned by participants about composite indicators was whether all the individual measures included in the composite really added up to give a complete picture of quality. While there were differences in the approach participants tended to take, the typical approach was to draw a distinction between *domains of quality* (the various aspects of quality that are important for the composite indicator) and *the individual performance measures included in the composite indicator.* Drawing this distinction was viewed as helpful because it allowed a more theory-led approach to designing a composite indicator, starting with a conceptual idea of what should be measured rather than with a set of existing performance measures. It also meant that it was easier to see where a composite indicator was missing aspects because there were no measures for one or more domains of quality.

> "What's the underlying construct that you're trying to measure, what are the domains that make up that construct, and then what are the measures that fall into each of those domains. [...] Instead of starting with the measures, it's more how do you start from the beginning and define the construct" (P10).

> "One of the challenges I think we face with the amount of data and the amount of different indicators that exist is [...] are they actually giving you a complete picture or are they kind of missing aspects?" (P06).

5.5.3.1 Conceptual frameworks for identifying domains

Three approaches to conceptualising domains of quality were repeatedly raised by participants. These three approaches were: Donabedian's Structure-Process-Outcome framework; the 'patient pathway' model; and consultation with clinical teams.

Donabedian's *Structure-Process-Outcome* framework was mentioned by every participant. Frequently, this was in the context of measure selection, such as *"our starting position was no process measures"* (P09). But others brought this framework up in terms of identifying domains of quality. For example, in the design process of one composite indicator of hospital safety, one participant explained that the expert team:

> “decided that the grade should be half process and half outcome measures. There was a lot of debate about how that would play out and whether that was right, but that was the consensus that the expert team took.” (P11).

From the fact that every participant raised it, it appeared that the main advantage of using Donabedian’s *Structure-Process-Outcome* framework for identifying domains of quality was its status as an intuitive and established way of categorising performance measures. But participants raised that there were drawbacks of using this framework as a tool for conceptualising aspects of quality, especially for composite indicator aiming to reflect more complex constructs of quality. Because the Donabedian framework is generic, and was not developed for this type of use, some participants who had applied it when developing composite indicators reported challenges around the selection of measures in their composite.

> “The area that we've gotten the most feedback on [...] is why we don't include mortality measures in our patient safety composite. So that's been really interesting to wrestle with. [...] Is that really a measure of patient safety or is that a measure of quality, I don't know. That's where we've struggled a little bit, we've gotten pushback from hospitals about that about why we aren't including mortality measures.” (P10).

Participants viewed alternative, more detailed, ways of conceptualising quality as helpful when developing more complicated composites. Many participants discussed using an idealised *patient pathway* as a tool for conceptualising domains of quality. In summary, care was conceptualised from the view of the patient, using the different stages of the care pathway as a way of understanding the care patients receive. This approach was naturally useful in evaluating care for specific conditions.

> “[We] laid out the trajectory of care starting from first symptoms to either cure or death or some end state, and then thought about the various stages of treatment, so the quality of diagnostic pathology care, the quality of initial treatment phase, the quality of ongoing treatment phase, and the quality of the management of care” (P07).

The third common approach discussed by participants was to use *consultation with clinical teams* to identify measures. This was a particularly common approach among participants involved in quality improvement programmes or clinical audits. The idea of this approach was to identify domains of quality that reflected the care provided by the different clinical teams. In the cancer example above, rather than considering the quality of the initial treatment phase, an approach based around clinical teams might instead consider quality of surgical treatment, the quality of systemic anti-cancer therapy, and the quality of radiotherapy, and might potentially use techniques such as driver diagrams to facilitate the process.

> "If the intent is for clinical quality improvement, we want to work out what the physicians are interested in." (P05).

> "A fairly common thing in the quality improvement world is simply a diagram that has at one side, the outcomes that patients will notice, and then linked into those will have the process measures that they already are using. And it's clear how they then feed into the outcome measures, so anyone can look at a glance and know exactly why that audit is asking for that piece of information, because it is an important measure" (P13).

5.5.4 Identifying individual quality measures for each domain of quality

Participants viewed identifying domains of quality as a valuable tool for identifying individual quality measures that could or should be included, as well as for ensuring that their understanding of the area of quality they intended to measure was appropriate. Yet identifying individual measures took additional effort beyond simply identifying domains. Participants discussed literature reviews, consultation with stakeholders, and expert advice as the main tools for identifying the appropriate individual quality measures corresponding to each domain of quality. Some participants discussed their conscious decision to restrict only to existing items, while others explicitly intended to develop new individual performance measures as part of developing a composite indicator.

> "There's a period of development of actual items, and usually this also involves [...] a literature review [...] to try and understand whether there are existing items that could be tested or adapted, but often it involves trying to write new items, because it could be that [...] you've identified a domain that nobody's really tested before, or items [...] need to accomplish something particular" (P03).

Participants emphasised that selecting appropriate performance measures was usually challenging. In part this reflected the difficulty of finding individual measures that were appropriate to include in the composite indicator, with various criteria discussed by participants set out in section 5.5.4.1. In their experience, participants reported that the ideal measures often simply did not exist. Even where participants were able to create or commission new performance measures, there remained challenges largely because creating suitable performance measures was viewed as a difficult process.

> “The difficult and very much underestimated in terms of time and effort process to at least produce [...] a basket of indicators of acceptable quality” (P01).

> “[We] wanted to [...] use publicly available data to create a composite. And so we do have limitations on what data are available, right. So I think there are some key areas that we wish were....for which we had publicly available data, that we don't. So there's a little bit of we're using the best information we have, but it's not necessarily a perfect set of information.” (P10)

> “If somebody could actually pull together, you know, top ten tips about indicator development, that would be useful. Because I think it is somewhat slightly trial and error, and when you're lucky, you get somebody who's, you know, been around long enough to have a sense of what works with indicators. And at the moment, it's a very unscientific process, in my view.” (P12)

Participants who had created new performance measures for use in a composite also raised various trade-offs, presenting two major concerns. The first was primarily around burden or cost. In participants’ accounts, developers of composite indicators were often very keen for more performance measures and more comprehensive data collection, but this was sometimes viewed to be without clear justification and to be imposing additional burden on clinical teams or hospitals. The second was around timelines. Some performance measures, often those linked to outcomes, may take more time to be realised or may require larger sample sizes to produce precise measurements. Participants raised that there could be an important trade-off between producing an indicator in a timely enough fashion to still be relevant and producing an indicator that included all the relevant information.

> “What data do you have already? Do you have to collect more data for your performance measure? If so, what are the consequences for that, in terms of the timeline and further data collection costs?” (P02).

> "I certainly don't think people deliberately go out…set out to measure things that don't matter. But I think that by its very nature, certainly within the national clinical audit world, there's a tract of people who are very enthused and engaged with their particular disease area, and they like to measure stuff, and that's why they've got involved. And so there can be a tendency […] of taking a view, well: 'If we can measure it, we will measure it. And then we'll think about what we use it for later.' And of course, that really needs to be the other way round." (P13).

5.5.4.1 What are the required properties for individual performance measures to be included in a composite indicator?

Composite indicators, are, by definition, composed of individual measures. These individual measures themselves may be of variable quality. In interviews, participants identified three desirable properties for individual measures: they should be valid; should be reliable; and should be practical to produce.

Participants felt that including measures with poor *validity* risked undermining the entire composite indicator. They raised the importance of multiple aspects of validity, face validity, content validity, and construct validity. These different types of validity were viewed as being important for different reasons. *Face validity* was viewed as important because participants felt that the inclusion of measures that did not immediately seem appropriate to the intended users of the composite indicator increased the risk that the results would simply be dismissed out of hand. Hence an early step in selecting individual measures was to:

> "whittle […] down to things that have face validity." (P03).

Participants also viewed more formal checks of validity as important. Often this involved involvement of stakeholders, ranging from expert panels to members of the public, to formally check the *content validity* of the individual measures identified for use in the indicator. Participants discussed various approaches, but they all came down to:

> "figuring out the relationship between an indicator and a particular composite domain and what's the justification and logic for that and being able to state that for every indicator that's included in the composite" (P07).

Many participants also raised challenges around the *construct validity* of individual performance measures – where a performance measure is included to measure one thing (e.g. the incidence of post-operative venous thromboembolism, where a higher rate indicates worse safety) but may be measuring something else (e.g. the effectiveness of local surveillance efforts for venous thromboembolism, where a higher rate indicators better safety [76]). One common theme was the importance of ensuring individual measures were appropriately case-mix adjusted. But participants reported many other challenges, ranging from the difficulties in translating survey questions between different contexts to differing approaches to recording comorbid conditions.

> "We're always challenged particularly with survey-related indicators to make sure we are using indicators that we don't think are subject to cross-cultural interpretation bias or translation bias, there are a whole set of issues. These actually are not unique to survey data, pretty remarkable differences between the way that countries collect data about their healthcare system. [...] That's kind of at a very macro level, but at every level of the systems, at least in my experience, as we've tried to develop indicators across health plans or across urban areas or across counties or other geographic areas or across primary care clinics. The underlying data variation that's just strictly due to differences in the way people code things, differences in the way the data are collected, the amount of missing data, all of that has to be sort of tested and evaluated." (P07).

Ensuring validity alone was not viewed as enough to make a measure suitable for inclusion. A second common concern among participants was that measures were acceptably precise, effectively that it was possible for a performance measure to detect an important difference between organisations, and that apparent differences were not driven by the random play of chance.

> "You need to have one eye on the natural variation, and small number variation in particular, for whatever indicator it is you're designing. And you need to have someone in the room who can fight that corner, otherwise you'll just end up with indicators driven by clinicians or by... politicians is the wrong word but by managers perhaps, who are naïve to the underlying statistical realities, and generate indicators which are overwhelmed with statistical noise and have very little actual real world significance or meaning." (P01).

However, issues of *practicality* tempered the desire for perfectly valid and highly precise performance measures. Participants often raised issues where the ideal performance measures could not be used in a composite indicator, either due to problems with data availability or because of the need to report within a certain timeframe for results to be actionable. For example, one participant discussed a composite that used measures from a voluntary survey of hospitals, but then wanted to assign reasonable composite ratings to hospitals that did not participate in the survey – leading to obvious trade-offs between coverage and consistency between hospitals. Other participants discussed needing to exclude ideal measures, or use shorter-than-ideal reporting periods, in order to allow the composite indicator to be produced in a timely fashion.

> “[Our survey gives] the best possible data that we could get on safety, just excellent data on safety. But not all hospitals agree to report to us. We get about 2,100 a year, hospitals, so that’s about 70 per cent of the hospitals, but not all of them, so some declined to report to us, and said, well we don’t have that data. So [...] we have a secondary source for that.” (P11).

> “How you’re actually to deploy it, both in terms of generating it in the sort of live environments, because it’s not much use if you just measure it, if you do it just once and publish it. It needs to be used for performance and needs to be updated with a useful frequency, and, I guess, as close to real time as possible” (P02).

> “You either have beautifully curated, perfectly, you know, statistically valid, and nice sample sizes, and you wait a year to get it. Or you have ongoing, you know, result control charts that have much, you know, much smaller datasets, that maybe give you an indication, but by no means are really, you know, black and white.” (P12).

The issues raised by participants around the selection of appropriate individual measures for a composite indicator highlighted many challenges. Part of this underscores the need for iterative development processes, set out in section 5.5.2. But what was clear was that participants felt that the process of checking the validity of the various individual measures was also a process of exploring the adequacy of the various domains of quality that were initially used to select performance measures. Thus, the final domains of quality used in a composite indicator might not match those initially identified when thinking about the purpose of the composite indicator.

5.5.5 Developing final domains of quality and the final set of measures

Participants often mentioned the importance re-assessing and refining the domains of quality, and then finalising the individual measures to use, once a set of potential measures had been identified. They explained that domains might need to be revised to make more sense, or to address characteristics of the individual measures, and often only some of the potential individual measures needed to be included in the composite. In part, this process of refinement was simply the iterative development discussed in section 5.5.2. Yet the emphasis, when discussing the domains of quality, shifted from ‘what aspects of quality do we need to capture?’ to ‘what is the most appropriate way to capture these aspects of quality?’

> “After talking to people, looking at the way data’s collected, what’s available, you can come up with a vague construct and then bring it back to those same people or different people to validate it, to say hey, this is what I heard you say, I went and did this information gathering, this is based on what you said and other people said, this is my general framework, what do you think about that, and get their feedback. The other thing I think you have to think about is what the thing is being used for.” (P09).

Participants split the process of developing final domains into two phases. The first phase was a formal check that the domains and measures made sense in combination. Both expert-led and data-driven approaches were commonly used by participants, with no consensus on which approach was best.

5.5.5.1 Expert-led or data-driven development of the final domains of quality

When combining multiple individual measures into single summaries of quality, participants were concerned about the risk of producing an overall score that failed to be a good guide to performance on individual, important, aspects. Participants discussed the need for thoughtful ways to address this challenge, which sometimes involved stepping away from producing a single composite indicator. But in practice this was viewed as a manageable risk, one that could, with care, be addressed.

> “You could get a negative kind of score in this composite when actually there might be one area that they are doing really well, and that might be the most pertinent area for the care of that patient. And actually therefore understanding the limitations of the indicator and [...] whether that composite could potentially mask some of these things, and how could you allow for that? How could you think of ways of identifying what those are, if you were looking for this kind of overall composite indicator.” (P06).

> "With the [Care Quality Commission] CQC, they do [the overall rating], but then they have their domains, don't they [...]. It's very easy to see where, once you've looked at the overall rating, what that is, the five [...] domains of care. And that, I suspect, is probably a reasonable compromise, so that you have that in front of you, as opposed to just, you know, this organisation is outstanding, or requires improvement, and you leave it at that. I think would be a little bit unfair." (P13).

Participants generally felt that this risk could be managed by being very careful about how individual measures were combined. They discussed two rather different approaches. The first, which I will call the *expert-led approach*, was by being very thoughtful about exactly how measures were combined into the overall composite indicator, so that the average across all (weighted) measures was a justifiable summary of relevant quality. The second, which I will call the *data-driven approach*, was by using multivariate statistical methods to produce data-driven summaries of what I will call *uni-dimensional domains*, that would potentially then need to be combined into the overall composite.

The expert-led approach was conceptually simple. While participants differed in their exact approach, expert-led approaches typically involved restricting to the 'most important' individual measures, and then defining a set of weights that reflected the importance of each measure to the domain of quality it sits in. These domains of quality could then be further weighted to be combined into the overall composite indicator. One participant noted the parallels with RAND Appropriateness Criteria [185]:

> "The appropriateness methodology is also grounded in convening mixed groups of stakeholders who can go through formal ranking and rating processes to understand the complexity of the relationship between indications and treatments, so we kind of had that in our methodological background as we were approaching quality measurement." (P07).

The data-driven approach was conceptually more challenging. The idea underlying this approach was that, instead of needing to carefully balance the measures within a domain, it was possible to define domains based on groups of measures that all appeared empirically to measure the same aspect of quality. Because all measures would then relate to the same construct, the hospital-level performance on these measures would then be correlated. This would make the specific weights used less important, and indeed could be suggested by the statistical methods used to derive these *empirically uni-dimensional* domains.

> “In developing our short-stay measure for nursing homes, which we also used latent modelling for, and we also used only measures that had short-stay patients in the denominator again – so you know, this clear clinical construct – we modelled it with many different candidate measures. And ultimately this billing-centredness measure proved out to be very closely correlated to the latent variable that we considered to be quality. So it ended up in our final model and in our published…publicly available nursing homes ratings for short-stay rehabilitation, the billing-centredness of the story on how much rehab therapy provided is one of the quality indicators that’s in that final model today.” (P04).

> “I find it easiest to aggregate […] building blocks that are uni-dimensional and then think about how you want to assign them weights […]. You can do that in multiple different ways, for example, there are tools that allow patients to describe the priorities of these, say, five different areas, and to create their own customised weighted average, we could have policy makers decide the relative importance of the measures.” (P03).

5.5.5.2 Ensuring that domains of quality make sense to the users of the composite

Participants stressed the need for the domains of quality used in a composite indicator to make sense to the users of the composite. As highlighted earlier, it was seen as key that the included measures appeared reasonable. But participants also felt that users needed to be able to trust that ‘Domain A’, ‘Domain B’, and ‘Domain C’ combine to produce some believable summary of quality. If domains did not appear credible, or if it was not obvious why individual measures fit in the domains they are assigned to, then they felt that users would question whether the composite indicator could provide a useful summary.

The issue of face validity was seen as especially important when the domains were designed using data-driven approaches. This was because the ‘empirical domains’ might include various individual measures that did not seem to fit together. Handling this was seen as requiring sense-checks and work with users of composites to ensure that domains made sense to them.

> “I don’t want to exaggerate the extent of what would happen, often the empirical groups and conceptual groupings correspond, but occasionally an empirical grouping is unsatisfying to patients […] in that they have trouble thinking of it as conceptually uni-dimensional […]. There was one example, there was an empirical domain that

> used to exist in the hospital survey, called Hospital Environments. It included things like cleanliness of the environs […] which were items which struck people initially as questionable, but turned out to validate against hospital specific rates of in-hospital infection, and say the noisiness of the hospital environment, which […] affects some patients sleeping and recovering. But the idea that noisiness and lack of cleanliness might be part of a single dimension was something that patients didn't see and so, no composite was ultimately formed there […] even though there might have been an empirical basis for creating such a grouping." (P03).

While participants agreed on the need for the audience to find the domains valid and believable, they varied on the importance they put on it. Some participants tended to privilege 'expert' views, and others were strong proponents of working with the intended audience to ensure everything was as understandable as possible.

> "When hospitals would say […] this doesn't seem valid, you'd say, well, here's our expert panel, that's what they recommended" (P11).

> "There's the whole method around all of that too, around the types of testing you do and what you test for, how you test for it, what kind of questions you ask. But it's fascinating to watch people look through and tell you what they're seeing and what it means and just moving a button here or there or changing a colour and all the effects it can have. And labelling, oh my gosh. Just spending time on figuring out what to call each domain. You could spend six months trying to optimise that, but that makes a big difference. What you call your composite, what you call the subdomains in people's interpretation." (P09).

A key part of making sense of groups of measures is being able to consistently compare scores between different performance measures. This was viewed as difficult but not particularly problematic, and the next section discusses the common approaches participants used for standardising individual performance measures.

5.5.6 Standardisation of individual measures

Participants noted one common challenge was simply getting all the different performance measures used in a composite indicator onto a common scale. Standardisation was viewed as an important and necessary technical step. Yet in partial contrast to the many challenges around selection of measures and the domains of quality within a composite, the approach to standardisation did not seem to be seen as an issue that would affect whether people trusted the composite indicator or found it useful. Still, there was no consensus over a preferred approach, with participants discussing many different approaches. These could be broadly categorized into either *Z-scores* or *normative scores*.

> "Just how do you get all your measures into some sort of common method of assessment, right. So is it a zero to 100 scale, is it, you know, deviation from the mean. You need some way of sort of standardising scores across different measure types, if you're using different measure types." (P10).

The *Z-score* approaches described by participants were motivated by the idea that the difference between two organisations on a performance measure can be understood in terms of the range of performances across all hospitals. In Z-scoring, the standardised score for each performance measure is based on how many standard deviations from the overall mean the score for that hospital is. In a typical application of Z-scoring, the standardised score on a performance measure is simply the number of standard deviations hospital performance is away from the mean. Some participants also discussed an approach where points were assigned based on thresholds based upon the number of standard deviations a hospital was away from the mean.

> "We basically had a massive composite for each NHS trust then with up to 120 indicators per trust. [...] In that we set the risk level, I think you got two points if you were three standard deviations roughly, and one point if you were at two. And then we basically did a... this is the total score you don't want to get, and this is how many you got out of it." (P14).

In contrast, the *normative* approaches described by participants aimed to produce scores that measured how good the quality of care was in absolute terms, rather than whether the care was better or worse than average. Standardisation in this sense equated to a normative classification of performance or 'standards setting' based on expert opinion and clinical guidelines. The approach was frequently described in the context of clinical audits or quality improvement, where there were evidence-based guidelines for some performance measures.

> "We wanted we wanted it to be ambitious and challenging, and we deliberately designed [...] an A grade [...] to say, okay, what does...like genuinely, world class

> healthcare/care look like, and we'll set it at that. Even though we knew at that point, when we first started it, no NHS organisation was achieving that. [...] It's an absolute measure of performance, it actually measures not relative to the performance of other healthcare organisations." (P02).

> "We used to ask at tender, what are your process and outcomes measures. [...] What we are doing much, much more, is starting with [...], what do you know about the problems that already exist, for your improvement aims what did the standards say in that area." (P12).

Being able to track performance over time and confidence that a good score on the composite actually means that the care provided was truly good, rather than just better than average, might be seen as key advantages over approaches based on Z-scores. Participants, however, explained that it was rarely as straightforward as it seemed. One major challenge was applying the normative approach in the absence of existing clinical guidelines or a clear understanding of what genuinely good performance looked like. Another issue was that, for many composite indicators, there were already a number of trade-offs in terms of the data that were available to produce the composite. In such situations, a normative approach was not particularly appropriate.

> "At least amongst the expert panels, the agreement was that we really, at that junction, didn't know enough to know what 'A' performance looked like. So we designed a relative score similar to the CMS star rating. And so it's really how are you doing relative to the other hospitals. We hope that maybe one day in the future we could move to absolute sort of thresholds of what is 'A' performance, but I don't know we are there yet." (P10).

> "We have favoured the Z-score approach as well, because really at the end of the day, the desire is to understand variability at every level and we know that we're [...] dealing with convenience samples of facilities, providers, patients, there are all sorts of selection bias issues in terms of who gets into care or not get into care. And so because of that uncontrollable sampling at some level, you kind of can impose a certain amount of order on it, but then at the end of the day, you're really just trying to understand the variability of a particular sample that you've drawn and not trying to generalise beyond that typically." (P07).

The different decisions discussed so far primarily relate to measuring individual aspects of quality, that have some form of clear conceptual or empirical basis. Many composite indicators are formed by combining multiple different aspects of quality. To participants, this meant averaging over the different domains.

5.5.7 Choosing weights to combine domain scores into the overall summary score

In every interview there was a point where the participant discussed the challenge of combining several unrelated scores, each representing a single domain or aspect of quality, into an overall composite score. All participants discussed using a weighted arithmetic average to combine such unrelated scores, recognising that the weights used to combine domain scores could be very influential over the apparent performance of different hospitals. Because of this, they found it desirable for the weights used to combine different domains in a composite indicator to be linked to some rationale.

> "The weight shouldn't be arbitrary, there should be some rationale" (P05).

> "Weighting is obviously a big key [issue], are you going to weight each measure the same, are you going to use different relative weights, and what are the criteria you'll use for setting those weights." (P10).

Yet while participants wanted weights to be justified, they also recognised that there might – in general – be no objective way to identify weights for domains of quality within a composite indicator. In participants' accounts, the choice of weights represented some type of outside prioritisation or perspective on the relative importance of domains, rather than any intrinsic relation to overall quality. This meant that identifying appropriate weights was not seeking some form of underlying truth, but instead accepting that domain weights were both subjective and potentially very influential for apparent organisational performance. Justification of the selected weights was more about following a sensible process and being able to explain how the choice of weights was made than about identifying a perfect set of domain weights.

> "There isn't a right way to weight these composites in general, they are constructs that you are looking for a majority of people to believe are real, so it's a mix of a scientific process and a small-p political process." (P07).

> "That's a political issue. I mean, there's no scientific formula for doing this for a composite. It's really about what sounds like the best way to do it, the fairest way [...]

> There's lots of ways to do a composite. It's just like when a professor puts together the requirements for getting the grade of the class, you have to take these two quizzes, you have to turn in two papers, and then there's the final. And we're putting it altogether and here's how, the final's going to count 30 per cent, you know, papers are whatever. I mean, all those things. There's no science to it, but it's kind of the best guess." (P11).

The difficulty of defining appropriate weights led some participants to prefer simpler composites, based on a relatively small number of individual measures grouped into a small number of domains of quality, over more complicated ones. These participants highlighted that with smaller composites, performance on individual domains and individual measures could be reported alongside the overall composite indicator. This made the impact of the weights used to produce the overall score far more obvious, and made it simple for any users who disagreed with the weights used to define their own.

> "We toyed with the idea of allowing people to define their own weights, and in fact one could do that given the data that we've reported, one can make up weights and decide, well, you think mortality is 60 per cent of the weighting as opposed to an equal part of the weighting. But our purpose in reporting it the way we did was to allow people to make those decisions on their own rather than to impose a weighting scheme which has gotten several other groups into trouble, either because they imposed the weighting scheme that people didn't find credible, or they didn't really reveal enough for people to understand that single indicator was getting all the weight in a composite. So that's a very tricky area, but one that I think with sufficient transparency, one can navigate to a reasonable outcome." (P07).

5.5.8 Reporting of composite indicators

Participants proposed three different, but largely complementary, priorities for the reporting strategies used for composite indicators: promoting trust and impact; transparency about technical aspects of the final composite indicator; and fully open code and data. In general, participants were more focused on what to report rather than issues such as visual presentation, but the importance of the ways in which composite indicators were reported was viewed as one important aspect of promoting the impact of the composite indicator. There were differences of opinion among participants over which of these motivations was more important, with disagreements largely about the risks and benefits of being too transparent.

5.5.8.1 Promoting trust and impact

Participants felt that it was key for reporting of composite indicators to promote trust in the composite indicator, supporting the use of the composite in an appropriate way. To many, this meant ensuring that reporting was as easy to understand as possible. In addition to being easy to understand, three characteristics were seen as particularly useful in promoting trust in the composite indicator: highlighting the expertise involved in the design of the indicator, or other specific areas of strength (e.g. dedicated focus on the needs of the audience); endorsement of the composite indicator by a trusted organisation; transparency of reporting, in that it demonstrated that developers of the indicator were not trying to hide any potential limitations.

> "It sounds a bit silly, but one of the reasons why we chose the A, B, C, D score [...] was [...] so people would understand, without having to read any background information, that A means [...] good, and D is not so good" (P02).

> "They're like, very big names in patient safety in the US. And, names that were, what I would call show-stopping. So, when hospitals would say, 'well, who said you should do the composites this way', or, you know, 'this doesn't seem valid', you'd say, well, 'here's our expert panel, that's what they recommended'. [...] I mean it's just an incredible list of experts. They're not going to do this in a way that's not responsible. So, that was a very important part of what we did [...]. For the credibility of the grade when we started, it was absolutely essential that we had a consensus from, again from absolutely top national experts." (P11)

Many of the considerations raised by participants around reporting to promote the impact of a composite indicator focused on the presentation of the composite summary itself, and in making it as easy as possible to trust and use the composite indicator without needing to examine any further documentation. But participants also recognised the importance of having documentation available. Some discussed the importance of methodological transparency, but others reflected that there were often users who wanted to know a little more but for whom the full technical details would be unhelpful. To address this, many participants discussed the value of 'lay summaries', by which I mean short and focused documentation that summarises the use – and any major limitations – of the composite indicator without going into too many technical details.

> “Say what the measure is and why it's important. Why we're looking at it. So, people can go to that level. [...] It's no good if you're only making it clear for the technically minded people who might construct these things themselves. Because we lose a lot of the audience then.” (P14)

> “What we would do for consumers is we would wrap the results up in the stories and then within that story we would explain this is how we made this composite and this is what it’s based on. So if you’re going to then publish the data for the consumer in the consumer end it would always be around a story to let people understand what the data means and how to use it. Not just here’s the data, here’s how to use it, but here’s an entire story about a particular quality area, like maternity quality of care we spent a lot of time on, or safety measures relating to infection, why is it important, but from a story telling point of view.” (P09).

There was little consensus among participants about how this type of documentation should be presented. Some reflected that different composite indicators with different intended audiences should have different types of meta-data. User testing of the documentation was seen as vital to ensure that the intended audience for a composite indicator found the lay summary and the presentation of the composite indicator as useful as possible.

> “There is a danger both…of either giving too little information, or giving too much information that actually the bits of information that someone might need are buried in pages of documentation and therefore no-one is going to read through that fine print in order to kind of find it. [...] Therefore understanding different users, and what is it that’s common to all of them that actually you need to get across in order to say, this indicator is doing this and it’s not doing that” (P06).

> “We’ve spent a ton of time on that, from coming up with 100 different ways to do it and getting a lot of consumer input on what made sense to them and how it would affect how they understood things. So you’ve got the end user you also hear, but really for us was the consumer perspective on what was understandable and meaningful. There is the published literature, the best practice, all the work by Judith Hibbard on how you present cost and quality data. I think that’s critical to start with as well. And a ton of testing, because you are not your own audience. Don’t ever fool yourself. So much testing.” (P09).

Being transparent about the technical details was also viewed as a key part of promoting trust in the composite indicator, but participants generally viewed transparency as critical for other reasons.

5.5.8.2 Transparency about technical aspects of the final composite indicator

Further building on the theme of promoting trust and impact and documentation, participants emphasised that transparency about technical aspects of the final composite indicator was vital. Yet while much of the documentation of composite indicators was viewed as aimed at normal users who wanted to know a little bit more, the detailed technical information was seen as having a different target audience. There was general agreement that only a small subset of people would actually want to read the technical documents, and these were:

> "the five to ten percent of those healthcare professional who are experts, [...] various academics who work in this area, [...] technical experts who want to dig right down to the bones of where an indicator is and where it comes from." (P01).

Participants felt it was important to be fully transparent about the technical aspects, even if they were of interest to relatively few people, for three reasons. The first, as alluded earlier, was that being open was perceived as helpful in encouraging others to trust the results. The second was that it this level of transparency was seen as in keeping with growing norms and expectations around transparency in research and accountability in public administration. The third was that many participants had received helpful methodological feedback following publishing technical information, and so it was viewed as helpful in improving the future design of the composite indicator.

> "We try and publish as much detail as possible about how we construct an indicator. We've published detailed stat papers in the past about stat use in our work. So, we're fully up for as much transparency as possible. And without that you often just lose people straight away. [...] Or we'll just get lots and lots of questions that we need to try and answer." (P14).

> "We obviously have to live by our own principles and so we are very, very transparent about everything we do, meaning that the methodology is detailed on our website, very easily accessible." (P11).

> "When we made the transition from ICD9 codes to ICD10 codes we published our entire, you know, sort of cross lock for the procedures and conditions ratings. And we

> didn’t hear from many people about it but those who did weigh in had a couple of really helpful suggestions, and so we actually made some tweaks to the codes before we published with ICD10 for the first time, based on the feedback that we got” (P04).

5.5.8.3 Fully open code and data used to produce the composite indicator

Most participants felt that true transparency, for a composite indicator, required more than simply documenting the process by which scores were created. The majority view was that domain scores and performance on individual measures should be reported alongside composite indicators, as should the statistical code used to turn the raw measures into the final composite. Yet there was also a counterargument that such high levels of transparency might have risks or unintended consequences.

The main perceived reason for providing the underlying data that composite indicators were based on was to help the rated organisations improve. Participants with experience of working in hospitals often discussed the pressure to dig into each summary rating, to work out why the hospital was receiving the score it was. Yet without the underlying data this was difficult. For composite indicators involved in uses linked to quality improvement, in particular, participants felt that it was necessary to publish underlying data.

> “Transparency is […] being able to disaggregate the composite scoring to the individual components, which are the actionable bits, is important, otherwise it’s very hard to know what it is to improve if it’s just kind of an aggregate score” (P02).

Even for composite indicators used for other purposes, making data and code available was viewed as a helpful step in making results reproducible and allowing others to check that the results were correct – making it possible for others to exercise scrutiny and to look for possible problems. And if a composite indicator had a high profile, then there was a clear incentive for healthcare organisations to look for these problems.

> “We give [hospitals] their numerical score, and we give them how they did on each one of the measures, and we give them a link to where we got the data on them, and we give them three weeks, and a password protected website, to tell us to look at their own data, and tell us, if they have any questions or issues with what we’re reporting on them. So, they have time in other words, to really dig in themselves and see how they did, and where we got it” (P11).

> “The underlying data sources have to be transparent, so people have to be able to go back to the original data and understand how it was collected, what the samples were, what potential sources of bias could have been introduced at the data collection stage, and that’s for every indicator including the composite” (P07).

Yet even proponents of this level of transparency recognised the risk that some healthcare organisations might abuse the availability of data and code to game the score. For high profile composite indicators, perhaps those linked to pay-for-performance or with a clear impact on organisational reputation, participants perceived a strong incentive for organisations to try to improve their score without necessarily improving their actual performance. These challenges led some participants to prefer a more limited approach to code and data sharing, depending primarily on technical documents rather than open data.

> “There’s interest in having the code itself released. And it’s something we’ve considered. I think there’s also enough of a cottage industry around sort of trying to help hospitals figure out how to gain performance metrics around…you know, pay and performance in particular. But I’m not sure of the unintended consequences of releasing, sort of, complete code sets.” (P04).

5.5.9 Competencies required to develop a composite indicator

Participants all recognised the need for different types of expertise in the development of composite indicators. They generally described development teams with a similar breadth of expertise, although the emphasis given to each role varied. But participants generally perceived inclusion of ‘lay’ representatives in the development of composite indicators as posing a number of challenges. Some participants described development processes without direct representation of any patients. Some viewed structured consultation processes as more helpful than solely having patients or members of the general public as direct members of the development team.

5.5.9.1 Necessary roles on the development team

Participants described development teams that brought together multiple forms of expertise. They typically involved a team leader, members with appropriate clinical and domain expertise, a statistician, and people with in-depth knowledge of the relevant datasets. Many participants highlighted that many members of the development team would be able to serve multiple roles at once.

The team leader was viewed as having both an external small-p political role and an internal managerial role. The internal managerial role was mentioned more consistently, and was viewed as necessary to keep the development team focused on the primary goal. The external political role was perhaps not necessary for every indicator, but the participants who brought it up were concerned that developing a composite indicator took a substantial effort and that this effort was largely wasted if a composite indicator was not used or if it was produced once and never again. Thus, the political role of the team leader was seen as promoting use and visibility of the composite indicator, and of ensuring the indicator was sustainable.

> "The initial push [to develop a composite indicator] comes from the manager or politician who is trying to change the world in some sense, so that they have [...] the motivation and the vision." (P01).

> "You need some folks who sort of know the politics of the space and how to navigate that, and who are the key players that you need to engage, and how are you going to roll this out." (P10).

> "Thinking about sustainability and how you're going to keep this thing going and how's it going to get paid for." (P09).

All participants discussed the need for some members of the development team to have relevant clinical or domain-specific expertise. They often highlighted that relevant expertise was not solely limited to clinicians, and that the relevant types of expertise that should be included will vary from composite indicator to composite indicator. Some participants raised other desirable criteria: including domain experts with a national or international reputation was seen as potentially helpful for the face validity of the resulting composite indicator; including clinicians with some statistical or dataset expertise was viewed as potentially helpful for handling some of the technical steps in the development process.

> "I just assumed, as a given, there was no way you could actually do successful quality measurement without involving the multiple stakeholders who participated in the care delivery. Certainly for any measures that are going to be based on guidelines of care or evidence-based treatment, having the guideline developers and the experts in the room was [...] a given, it shocks me that anyone would think you could do this with just analysing data. [...] It does several things, one is that it ensures that the

> priorities are correct for measurable aspects of care, unless you sort of understand the relative importance of different elements of the treatment protocol, it’s very hard to make sure that you’re focusing on those elements that are really critical to the patients’ outcome.” (P07).

> “Hospitals would say, well, ‘who said you should do the composites this way?’, or […] ‘this doesn’t seem valid’, you’d say ‘well, here’s our expert panel, that’s what they recommended’. And it would like, oh, I mean it’s just an incredible list of experts.” (P11).

> “You need a, usually senior, clinician who is willing and has some experience of the underlying data and coding issues and […] knows how the data reflects what is happening in any hidden clinical reality. […] You need to get one of them on board because […] our data expert doesn’t fundamentally understand what is the difference between code A, which is one type of procedure, and code B, which is another type of procedure.” (P01).

Statisticians, or other quantitative methodologists, were also considered a vital part of the development of any composite indicator. Participants gave two sets of reasons for this. First, they explained that the development of composite indicators could involve detailed quantitative analysis, either to produce the final indicator or in checking robustness of results. Second, even where the composite indicator was relatively simple, there was a perceived need for a statistician to recognise the limitations of the underlying data and what could reasonably be done with the available data.

> “Assessing individual items for order effects, assessing them for variation across entities that are being evaluated, assessing their basic associations with […] things that you expect them to be associated with.” (P03).

> “Monte Carlo simulation [to explore] the effect of removing indicators from composites as a way of kind of assuring ourselves that the results are relatively stable under a variety of scenarios.” (P07).

> “[…] have one eye on the natural variation, and small number variation in particular, for whatever indicator it is you’re designing. And you need to have someone in the room who can fight that corner, otherwise you’ll […] generate indicators which are

> overwhelmed with statistical noise and have very little actual real world significance or meaning." (P01).

Some participants additionally raised the need to involve people with specific expertise with the data sources that will be used to produce the composite indicator. The extent to which this was seen as necessary varied a little by context. Participants who discussed the use of large individual-level datasets in producing composite indicators (including survey data, claims data, and datasets such as Hospital Episode Statistics data in England) stressed the value of involving people with plenty of experience working with these data. This was seen as helpful both in avoiding mistakes and also in speeding up the process of producing sensible performance measures. Yet many participants primarily discussed developing composite indicators from existing performance measures. These participants generally did not discuss the need for including experts in the specific datasets these performance measures were drawn from, but did often discuss the importance of at least talking with such experts to identify possible challenges.

> "In any data set there are oddities which are non-apparent to the average person, even if the average person is a competent and experienced analyst in other domains. Each data has its own foibles." (P01).

> "After talking to the measure stewards and got inside information about the problems of the measure we didn't use it. It wasn't so much because stakeholders thought it was terrible, which they did, but it was because we felt like after talking to this measure steward that there were good reasons not to use it. [...] You have these balancing things of stakeholders versus what's been published about it versus what people will tell you about it but hasn't been published." (P09).

When discussing the people involved in developing composite indicators, many participants mentioned patients and members of the public without prompting; others did not. In contrast to other roles on the development team, involvement of 'lay' representatives was viewed as something that, to be most useful, needed very careful handling.

5.5.9.2 Capturing the perspectives of patients and the public

Participants agreed it was important that the design of composite indicators should incorporate perspectives from patients and the public about what aspects of quality were important, how these aspects should be measured, and how best to report results. This was necessarily the case for composite indicators aimed at supporting patients in choosing the most appropriate place to seek care, but participants noted that understanding patient priorities remained important even if the composite indicator had a different primary aim. They reflected that ultimately the point of most applications of composite indicators in healthcare was to improve the care a patient can expect to receive. But while there was agreement on this point, the challenges of actually involving ‘lay’ voices in the development process led to ambivalence about how best to capture these perspectives.

Typically, participants discussed including expertise in the development process by adding members with appropriate experience to the development team. While this was viewed as the appropriate approach for those with professional expertise, participants were less sure about including patients or members of the public in this way. Though seen as a good idea in principle, some participants reflected that in practice the approach had limitations. On the one hand, they were concerned about the possible motivations behind the contributions of some patients (often while noting that the same could be said about ‘expert’ members of the development team too). And on the other hand, they were concerned that at the risk that ‘lay’ participants might not always contribute (though when they did, those contributions might be very valuable) or that operationalising their contributions would be very challenging.

> “Patient involvement is difficult. [...] I kind of feel that we should get a patient view on... I mean most of the data that we look at belongs to the patients essentially. [...] It’s easy to say that but it’s probably more difficult to do in practice. And there’s also a kind of question of a patient can be...come from a very particular point of view, but you can also argue that about other roles as well, about clinicians about the surgeons, who’ll have a very strong viewpoint.” (P06).

> “I’ve been involved in lots of groups in which there’s been a patient representative. I think the majority of them, the patient representative didn’t contribute that much. It’s good to have them there and I support the fact of them being there. Having said that, in a handful of cases the patient representative has intervened very meaningfully, where the informatics or technical professions in the room have been in some form of group think or other, and the patient has asked a pointed question which shakes that up.” (P01).

> “It can be challenging to wrangle the perspectives that patients bring to the table, which obviously are more valid than anyone else’s perspective at the end of the day.

> But to turn them and to convert them into sort of methodological guidance beyond, you know, this measure is important or this measure isn't important to a patient, it has proved challenging for us." (P04).

The challenge of capturing patient perspectives through direct inclusion of 'lay' voices in the development process led some participants to prioritise other sources of knowledge. Some participants noted that the academic literature already includes much relevant information about patient priorities, and that it could be more efficient to work with this pre-processed knowledge about patient perspectives. Others reflected that members of the development team with professional expertise might also have some experience of being a patient, and would bring that perspective to the team.

> "There's a cluster of researchers who've done a lot of work on patient-centred outcome research and patient perspectives on what matters to them in the care they receive, and how they prioritise both in interpreting the information that's publicly reported but also what their priorities are as a patient. And so, I think we've found actually that the literature to do a lot of the work for us, and help make sure that we are keeping the patient perspective in mind, but to begin it's often filtered through the lens of what the academics have published." (P04).

> "At the end of the day, we all are patients, right. What's interesting is, most of us do have experience in that healthcare system as a patient. But that's not to say we shouldn't explicitly have included patients." (P10).

A different group of participants felt that it was not reasonable to produce a composite indicator without working directly with patients and the public to understand their priorities, and discussed approaches to consultation with these key stakeholders that they felt were more helpful than direct involvement in the development team. The main theme in these accounts was the value of a carefully structured approach focusing on the issues where the 'lay' perspective was most critical, and perhaps using prototype composite indicators as a tool to guide discussion. These consultation processes were often described by participants as being a conversation, but the exact formats differed, with some talking about steering groups and others describing a process more akin to qualitative research.

> "It's important to have a conversation as well as set out a vision of what would be the absolute ideal, and it also helps people understand what's possible, so it can help you

actually create something that's actually realistic and feasible. Because as well as it being a waste of time if you do lots of consultation and people design a unicorn, then that also leads to a [...] mismatch of expectations of, well, you asked us to design this, and that's actually not what we got, is an issue. So doing that well, that's the consulting with the users in a sophisticated and not naive way, is important." (P02).

"Start by talking to patients about what the services they received and what the issues were and how they felt about it [...]. One-on-one interviews are way better than focus group type things, only because people are willing to go into a lot more detail and talk about things that they won't talk about in groups." (P09).

5.6 Discussion

The interviews I conducted for this thesis have enabled a detailed exploration of expert views on the development of composite indicators. What emerged was a strong emphasis on the purpose of the indicator and on the need for an explicitly iterative development process. Developing composite indicators was seen as a complex process, involving many technical and conceptual challenges, requiring multiple forms of expertise, and with many decisions involving difficult trade-offs.

Three points raised by participants appear to be underemphasised in existing literature. The first was the importance of defining the purpose of the composite indicator. Designing a composite for use in guiding patient choice might involve a completely different process from designing a composite indicator for summarising performance in a clinical audit, for example. But without making that explicit, it is easy to overlook. The second important point was the value of prototypes and of trying out many alternative designs. This matches some of the thinking behind earlier chapters of this thesis, but this type of active iterative development is rarely so explicitly articulated in the existing literature. The third was tensions and trade-offs about the value and difficulties of involving patients and the public in the development of composite indicators, and indeed in performance measures more widely.

In what follows, I locate the findings of the interview study in the context of current literature on composite indicator development (section 5.6.1) and existing evaluations of composite indicators (section 5.6.2), and I reflect on the difficulties of patient and public involvement indicator development (section 5.6.3). Finally, I comment on my experience of carrying out qualitative research, given my quantitative background (section 5.6.4), and finish with general conclusions from this study.

5.6.1 Contextualising the findings in the existing literature on developing composite indicators

The findings of the interviews share much in common with the existing literature (with some key sources briefly summarised in Table 22), but as I show below, they also expand and deepen our current understanding, in some cases offering some challenge to current practice.

Purpose of the composite indicator. Participants in my study were emphatic about the need for clarity on the goal of a composite indicator. In some descriptions of the development process of a specific composite indicator, such as Noble and colleagues description of the development of the indices of multiple deprivation [186], the emphasis on the goal of the composite indicator was often less obvious, and only partially revealed through the discussion of the theoretical framework used in constructing the indicators (Table 22). However, two well-established and authoritative guides for composite indicator development in healthcare – chapter 8 of Bottle and Aylin's book *Statistical Methods for Healthcare Performance Measurement* and Jacobs, Smith and Goddard's technical report *Measuring performance: An examination of composite performance indicators* — highlight the importance of identifying the purpose of the indicator [18,49]. Despite these calls in the literature, however, there are many examples of existing composite indicators which do not have a clear purpose. In chapter 6 I used the CMS Star Ratings as an example of this issue, and made some recommendations around designing composite indicators with purposes which are more immediately apparent.

Iterative development process. Participants in the interviews identified the importance of an iterative development process, where a cyclical approach to improvement is adopted. While current guides to composite indicator development typically do not stress iterative development (Table 22), these findings suggest an opportunity for evaluating more recursive approaches in the future. The OECD Handbook comes closest to addressing the iterative nature of the development process [1], with the Competence Centre on Composite Indicators and Scorecards (COIN) 10 Step Guide (which is based on the OECD Handbook) being somewhat more explicit about the value of an iterative approach [187]. Both the COIN 10 Step Guide and the OECD Handbook present an apparently linear set of steps. But within each of these steps is a series of suggestions and issues for the developers to address, and some of these issues imply an iterative development process. For example, the third step is multivariate analysis (e.g. using factor analysis), and the OECD Handbook states:

> "The analyst must first decide whether the nested structure of the composite indicator is well-defined (see Step 1 [Developing a theoretical framework]) and if the set of available subindicators is sufficient or appropriate to describe the phenomenon (see Step 2 [Selecting variables]). ... If not, a revision of the sub-indicators might be needed." (OECD Handbook on Constructing Composite Indicators [1], page 13).

Meanwhile, the COIN guide emphasises that:

> “although presented consecutively in the Handbook, the benefit to the developer is in the iterative nature of the steps.” (COIN 10 Step Guide [187]).

How to approach the practical challenge of developing a composite indicator using an iterative approach, and that can accommodate issues such as an inadequate set of individual indicators, needs further investigation.

Identifying individual quality measures for each domain. Existing guides give much emphasis to the need to identify individual quality measures (Table 22), and the fact that it is often challenging to find appropriate quality measures. My interview participants similarly highlighted that it was common to find that there were no appropriate quality measures for domains of quality when producing a composite indicator. In their accounts of developing composite indicators, this was often an area where multiple rounds of iterative development were required in order to identify a more specific purpose for which there were appropriate quality measures, or for which such measures could be developed. Yet this is an area where many existing composite indicators, even those designed in highly principled ways, have problems. One such example is the Baby-MONITOR composite produced by Profit and colleagues [188]:

> “Safety and effectiveness were the primary domains of quality assigned to the selected nine measures. These results imply that in its first iteration, the Baby-Monitor will contain only two rather than all six of the Institute of Medicine's domains of quality.” (Formal selection of measures for a composite index of NICU quality of care: Baby-MONITOR, Profit *et al* 2011 [188]).

Technical aspects of the development process. Findings from my interview study on the more technical aspects of composite indicator development from this study were generally congruent with those from existing guides to indicator development (Table 22), perhaps reflecting the relative underlying ease with which consensus can be reached on some of these. The existing literature and my interview participants generally discussed similar concerns when addressing the desirable properties of individual performance measures, approaches to standardisation of these measures, and combining these measures to produce domain scores and overall composite scores.

While the broad concepts and reasons discussed by participants were very similar to those in the literature, that there are many specific technical approaches that could be used in developing composite indicators that were not covered during interviews. This may have been an artefact of the interviewing format, and in particular that participants were encouraged to discuss how they would approach certain issues. It seems reasonable that participants were more likely to discuss specific approaches that are in common use in designing composite indicators of healthcare quality, rather than those that are rarely used.

Thus, for example, participants discussed three specific approaches to standardisation, while Jacobs, Smith and Goddard discuss nine [49]. Similarly, participants discussed two methods for combining individual measures into composites, namely factor analysis and weighted arithmetic means. But there are many other approaches that could be used to combine measures, but that are rarely used in producing composite indicators in healthcare, including geometric means and other multiplicative approaches to combining domain scores [1,175]; multivariate statistical approaches other than factor analysis or principal components analysis [189–191]; and various multi-criteria decision analysis approaches [1]. See Appendix 3 for details of these alternative approaches.

When using the results of this interview study are used to design or help report composite indicators, how participants focused primarily on common current methods may be a limitation. A reporting guideline that asks how domains were weighted may swiftly become outdated if the alternative approaches to combining domain scores mentioned above become more widespread. Further developments toward a reporting guideline would need to ensure that the focus in the interview study on common current practice did not lead to an overly prescriptive guideline

Table 22. Steps in development of composite indicators implied by this study, and comparison with guidelines to indicator development identified in the literature. Items from the literature may be presented in a different order than they are in the source documents.

This study	**Composite Indicators of Country Performance: A Critical Assessment** [51]	**OECD Handbook on Constructing Composite Indicators** [1] **(and the COIN 10 Step Guide** [187]**)**	**Measuring performance: An examination of composite performance indicators** [49]	**Improving benchmarking by using an explicit framework for the development of composite indicators: an example using pediatric quality of care** [50]	**Statistical methods for healthcare performance monitoring** [18]	**Measuring Multiple Deprivation at the Small-Area Level** [186]
Context: healthcare	Context: economics	Context: economics	Context: healthcare	Context: healthcare	Context: healthcare	Context: public policy
Identifying the purpose of the indicator (see section 5.5.1)	N/A	Developing a theoretical framework	Choosing the organisational objectives to be encompassed in the composite	N/A	Specify the scope and purpose	The spatial scale
			Choosing the entities to be assessed		Choose the unit	
Allowing for iterative development (see section 5.5.2)	N/A	N/A	N/A	N/A	N/A	N/A
Identifying domains of quality (see section 5.5.3	Developing a theoretical framework for the composite	(As part of framework)	N/A	Framework	N/A	Establishing a clear theoretical framework for the measurement of small-area deprivation
Identifying individual quality measures for each domain (see section 5.5.4)	Identifying and developing relevant variables	Selecting variables	Choosing the indicators to be included	Metric selection	Choose the indicators and run descriptive analyses	Domains and indicators
		Imputation of missing data	Adjusting for environmental or other uncontrollable influences on performance	Missing data	Select the data and deal with missing values	The small-numbers problem and the shrinkage technique
			Adjusting for variations in expenditure if a measure of efficiency is required	Initial data analysis		

This study	**Composite Indicators of Country Performance: A Critical Assessment** [51]	**OECD Handbook on Constructing Composite Indicators** [1] **(and the COIN 10 Step Guide** [187]**)**	**Measuring performance: An examination of composite performance indicators** [49]	**Improving benchmarking by using an explicit framework for the development of composite indicators: an example using pediatric quality of care** [50]	**Statistical methods for healthcare performance monitoring** [18]	**Measuring Multiple Deprivation at the Small-Area Level** [186]
Developing final domains of quality and the final set of measures (see section 5.5.5)	N/A	Multivariate analysis	N/A	N/A	N/A	Combining the indicators into domain deprivation measures or domain indices
Standardisation of individual measures (see section 5.5.6)	Standardising variables to allow comparisons	Normalisation of data	Transforming measured performance on individual indicators Combining the individual measures using addition or some other decision rules	Normalisation	Normalise the metrics	Standardising and transforming the domain deprivation measures or domain indices
Combining domain scores into the summary score (see section 5.5.7)	Weighting variables and groups of variables	Weighting and aggregation	Specifying an appropriate set of weights	Weighting and aggregation	Assign weights and aggregate the component indicators	Weighting the domains
Reporting (see section 5.5.8)	N/A	Presentation and dissemination Back to the details (i.e. make individual measures available)	N/A	Presentation and dissemination Deconstruction	Present the results	N/A
Competencies of the development team (see section 5.5.9)	N/A	N/A	N/A	N/A	N/A	N/A
N/A	Conducting sensitivity tests on the robustness of aggregated variables	Robustness and sensitivity	Using sensitivity analysis to test the robustness of the composite to the various methodological choices	Uncertainty analysis	Run sensitivity analyses	N/A
N/A	N/A	Links to other variables	N/A	Links to other metrics	N/A	N/A

5.6.2 Comparison of interview findings with Rating the Raters, an evaluation of several US composite indicators of health quality and/or safety

Bilimoria and colleagues recently rated a series of US composite indicators of hospital quality based on a set of criteria they developed [65]. As the publication by Bilimoria and colleagues represents an important source in this field, I specifically discuss its contents against my findings. They identified six domains against which to evaluate hospital quality rating systems, each containing various criteria:

- Potential for misclassification of hospital performance
- Importance/impact
- Scientific acceptability
- Iterative improvement
- Transparency
- Usability

Five of these domains proposed by Bilimoria and colleagues have natural counterparts in the themes arising from the interviews (Table 23). The interviews did not identify formal pre-release peer review of the methods (under iterative improvement), or providing information on financial conflicts of interest (under transparency). Both might be useful components to explore as part of a future reporting framework.

The importance/impact domain proposed by Bilimoria and colleagues in their ‘Rating the Raters’ study did not have a matching theme in the interview study [65]. They viewed composite indicators as performing well on this domain if they had unique features that were viewed as resonating with patients, referring physicians or hospitals. For example, the use of a “reputation” domain in the US News and World Report Best Hospitals composite indicator was viewed as one such unique feature, even while they noted the potential flaws with the approach. The authors of the ‘Rating the Raters’ study were researchers into performance measurement, and possibly to them part of importance/impact was the impact on the field of performance measurement, hence the emphasis on novel approaches to measurement. In contrast, interview participants in my study discussed the way they would design a composite indicator, rather than ways to move the field of performance measurement forward.

Table 23. Comparison of domains used in "Rating the Raters: an Evaluation of Publicly Reported Hospital Quality Rating Systems" and the major themes from this study.

Domain used in Bilimoria *et al* Rating the Raters: An Evaluation of Publicly Reported Hospital Quality Rating Systems [65]	Example criteria from Bilimoria *et al* Rating the Raters: An Evaluation of Publicly Reported Hospital Quality Rating Systems [65]	Similar themes from this study
Potential for misclassification of hospital performance	Use of known measures that are flawed Hospitals examined (number and types) Risk adjustment Composite methodology Methodological approach Audit mechanism	Identifying individual quality measures for each domain Combining domain scores into the overall score
Importance/impact	Unique features that resonate with patients, referring physicians, and hospitals	
Scientific acceptability	Balanced measurement Hospitals examined (number and types) Distribution/assignment of hospital grades/stars Use of available measures Use of unique data Specific methodological concerns Stability of rankings over time Audit mechanism	Identifying domains of quality Developing final domains of quality
Iterative improvement	Response to stakeholder feedback and scientific advances in measurement science Review of methods prior to release Peer review of methods Expert panel level of involvement	Development as an iterative process Reporting of composite indicators Competencies involved in developing a composite indicator
Transparency	Detailed methods report available (transparency) Clear rationale for methodological decisions Data availability (replicability) Financial conflicts and details regarding how ratings are monetized	Reporting of composite indicators Purpose-led development
Usability	Ease of overall use Ability to compare hospitals easily Attention to varying health literacy and numeracy	Reporting of composite indicators

5.6.3 Reflections on the challenge of patient and public involvement

Participants expressed some ambivalence about the value of patient and public involvement in the development of composite indicators. Gargon and colleagues' analysis of interviews with developers of core outcome sets also found that developers frequently problematized patient participation [157]. One reason given for not including patients in development of core outcome sets was that doing so would have been complicated and challenging – and some of the developers of composite indicators I interviewed raised exactly the same point. Yet, in contrast to the developers of core outcome sets interviewed by Gargon and colleagues [157], the developers of composite indicators that I interviewed were broadly in favour of the principle of involving patients in the development of composite indicators; they just felt it was very challenging to do in a meaningful or straightforwardly operationalisable way.

Patient and public involvement, in general, is motivated by both technocratic and democratic concerns [192–195]. The technocratic motivation for such involvement is the perception that those with direct personal experience of a particular medical condition have a better understanding of what is important for those in that situation than anyone else, and so involving patients will lead to objectively better decisions. The democratic motivation is that these decisions are often highly consequential – in a performance measurement setting, for example, they might affect which types of care are incentivised – and hence it is not appropriate for these decisions to be made without patients in the room.

From participants' discussions of occasions patient representatives made 'meaningful' contributions, and from their reflections on the value of accessing patient views distilled through the academic literature, I infer that most participants saw patient and public involvement as a technocratic endeavour. This explains why some participants were enthusiastic about the value of user testing, interviews, and focus groups in understanding patient views while being cautious about the apparent risks of less structured approaches – including simply having patients on the development team.

5.6.4 Reflections on the differences between doing quantitative research and (basic) qualitative research

I found carrying out qualitative research very different from carrying out quantitative research. Parts of it were fascinating, with interviews and early parts of the analysis being more enjoyable than any aspect of quantitative research. The best part of the interview study was simply talking to people. Due to the recruitment process, all the participants were experts in performance measurement who wanted to talk to someone about composite indicators. This led to many great discussions, enjoyable for both parties. Many interviews ran longer than one hour, and on a few occasions I had to remind participants we were coming to the end of our scheduled meeting to make sure they had no other commitments. One of the reasons this was so enjoyable was that I had enough knowledge to feel comfortable, especially in later interviews, in allowing the interview to proceed organically. While I still referred to my prompt guide, I generally followed up interesting points as they arose and use time at the end to cover topics that were missed. To me, these interviews felt more like the types of discussion one sometimes gets into at a conference rather than a form of research.

Other parts were painful, particularly the process of analysing qualitative data. The amount of time taken to code a transcript, decide that the coding framework was inadequate, and then come back and re-code the transcript was astonishing. It felt similar to data cleaning, in fact, because the approach was all about finding the concepts that did not quite fit together. It was also very unfocused, with interview participants rapidly covering diverse ranges of issues. The write-up process was very challenging for me, and I found myself missing the more structured approach to write-up and analysis typical of quantitative research.

This difficulty of analysis was directly due to the great advantage of qualitative research: the freedom to continually revise the entire analytical framework throughout the entire research process. By doing this in a thoughtful, reflective way, one can end up with a satisfying catalogue of themes that address and interpret the issues raised by participants. This freedom was also deeply unnatural to me, because in an ideal quantitative study the entire analytical framework should be fixed before data are collected, and certainly before analysis begins. This is not, of course, a criticism of the qualitative approach, as the statistical philosophy that leads to analytical flexibility being problematic for quantitative studies is simply not relevant when analyses are not based on statistics.

Using a qualitative approach to examine quantitative methodological questions such as those addressed in my study is uncommon. I know of only three such analyses, all qualitative studies carried out by statisticians, one relating to core outcome sets and two on adaptive trial design and all published in the last few years [155–157]. There is little guidance about how best to carry out qualitative research of this type. I found this an appropriate way of researching a complex methodological question, and I expect similar analyses will become more common as researchers formally engage with increasingly complicated methodological questions.

5.6.5 Conclusions

Participants' accounts enabled a detailed exploration of the way that composite indicators are currently designed in healthcare. This exploration produced areas of apparent consensus, including many technical aspects of individual performance measures, the need for careful attention to how measures and domains are combined, and the need for composite indicators to align with the needs of the people who are meant to be using them. While there was broad agreement over the challenges that needed addressing when developing a composite indicator, there was less agreement over the best technical approaches by which these challenges could be addressed. This accords with existing guides to composite indicator development which typically identify a range of possible methods rather than specifying a preferred approach (e.g. [1,18]).

This interview study is intended to be an early step in the development of a reporting guideline. While the findings of the interviews provide important sensitisation to a range of issues likely to be significant in developing composite indicators, they are unlikely on their own to be sufficiently comprehensive to be used as the basis for a reporting guideline. Future development steps will be discussed in the next chapter.

6 Discussion and reflections: Improving composite indicators of healthcare quality

Composite indicators are technically and conceptually challenging. My thesis has shown that their development process relies on a series of technical choices and decisions that can powerfully influence judgements about which hospital/organisation/unit is better or worse, limit the value of the information for purposes of identifying performance and opportunities for improvement, and produce many unintended consequences. Yet many of the design choices underlying technical decisions in the construction of composite indicators remain obscure and poorly reported. This final chapter reflects on the challenges of composite indicators in healthcare, further building on the consideration of the findings of my quantitative and qualitative studies outlined in chapters 3, 4 and 5, considers approaches to studying complex methodological questions, and concludes with thoughts about how better design and reporting of composite indicators can be best supported in the future.

I set out with two linked objectives. First, to characterise challenges in how composite indicators are currently developed. Second, to explore how reporting of composite indicators may be improved. My thesis has addressed these linked objectives using an innovative multi-method approach, combining novel applications of advanced quantitative methods to examine the impact of different technical choices with a rare application of qualitative research to the study of methodological issues.

The quantitative aspects of this thesis primarily addressed my first aim, by examining how the design of composite indicators affected organisational rankings and performance ratings. The quantitative studies provided three main results. First, I demonstrated that Monte Carlo simulation was a useful and practical way of assessing the sensitivity of composite indicators to design decisions, and not just for measure weights (as has been shown in the past in healthcare [95]) but for multiple technical decisions involved in the design of an indicator – a finding that was also highly relevant to my second objective. Second, I gave practical examples of alternative calculations of existing composite indicators, showing how such alternative specifications could be operationalised as well as the impact of the choice of specification. This made my analysis unusual within health services research, where most assessment of design choices on a composite indicator examines a possible composite indicator rather than one in current use (for example, [49,94,95]). Finally, I demonstrated the importance of carrying out sensitivity analyses when developing composite indicators, by showing that the apparent sensitivity of an indicator to design decisions varied between indicators. The CMS Star Ratings (a very high-profile US measure) appeared far more vulnerable to choices about its specification than the SSNAP score and level (used as part of a national clinical audit of stroke care in the UK), although both had notable limitations.

My qualitative study primarily addressed my second aim, exploring how reporting of composite indicators can be improved. My qualitative results were based on a series of interviews with experts in the design of composite indicators from a series of backgrounds, interviews which focused on how these experts would go about developing a composite indicator. While this had clear relevance for my first objective, identifying key steps in the development of composite indicators is crucial for ensuring these steps are clearly reported. One I feel underappreciated aspect was the importance of the purpose behind the composite indicator. Developing a composite indicator for one purpose – say, quality improvement – but using it for another – perhaps pay-for-performance – may be deeply problematic.

By synthesising my qualitative results with evidence from the relevant literature, and the lessons from my quantitative analyses, I was able to identify current conceptual approaches to developing composite indicators, allowing me to develop a process map (see below) of composite indicator development and a draft prototype checklist for critical appraisal (see below).

6.1 Understanding design and reporting of composite indicators

In this section, I summarise new understanding of how composite indicators are designed and reported in healthcare based on the findings of my thesis. I offer a process model for understanding the key steps in designing and developing composite indicators based on synthesis of my work throughout the thesis, and a prototype critical appraisal tool for composite indicators.

6.1.1 A process model for key steps in designing and developing composite indicators

One achievement of my thesis is that, through my immersion in the literature and my learning from participants in my interview study, I have been able to develop a process model (Figure 18) of the steps involved in designing and developing a composite indicator. This process model allows those intending to develop of a composite indicator to understand the steps involved, and to see how these different steps fit and impact upon each other. It may also assist the creation of supporting technical documentation, in that those reporting a composite indicator can use this model to ensure that they fully report each of the steps captured in the process model. Finally, by understanding the process of developing a composite indicator, it becomes possible to construct principled critical appraisal tools, which I explore in the next section.

Reassuringly, my process model reflects much of the existing guidance on development of composite indicators (e.g. [1,18,49,50]), but with greater emphasis on two dimensions, the intended purpose of the composite indicator, and on the iterative nature of the development process.

In summary, the process model is as follows. Development of a composite indicator should begin with a clear purpose and aim. This aim should be used to identify provisional domains of quality reflecting different aspects of quality relevant to the purpose. An iterative process of identifying and refining domains of quality and individual quality measures, both for statistical properties and acceptability with the audience, then follows:

- Provisional domains of quality are used to identify potential individual measures to include.
- These provisional domains and potential measures are tested with the intended audience of the indicator.
- Audience feedback – and quantitative analysis of validity and reliability – are used to refine provisional domains, and to cut down the individual measures into a set of higher quality and more relevant measures.

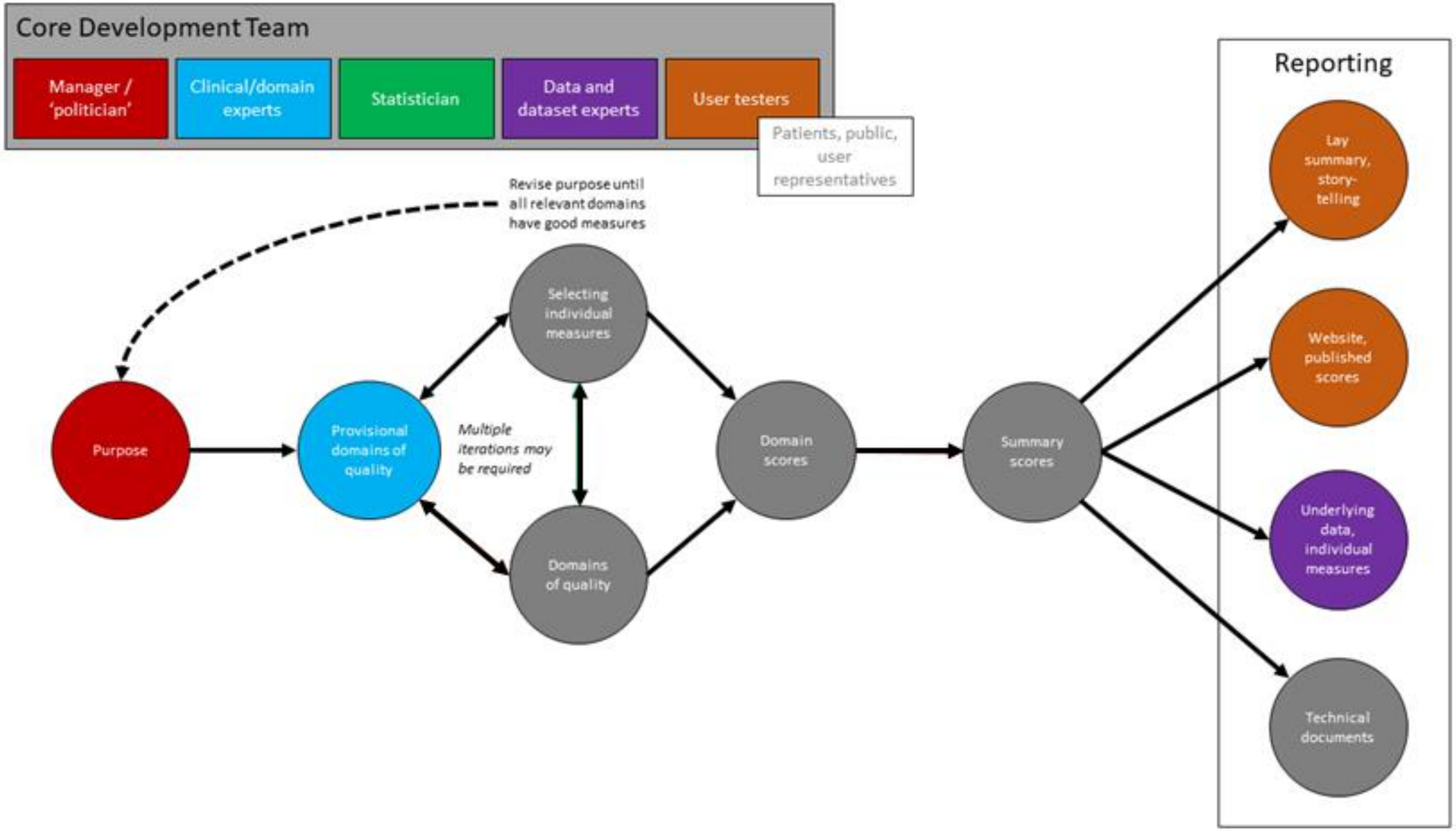
Once a final set of domains and measures is reached, it becomes necessary to check whether the set of domains and measures that are available allow the purpose of the composite indicator to be achieved.

The next step is to combine individual measures to produce scores for each of the domains of quality. These domain scores are then combined to produce the overall summary score.

The final step is reporting. This involves working with the intended audience to develop appropriate lay summaries and a way of publishing and presenting scores, including how underlying data on the different individual measures will be made available to the organisations rated on the composite indicator. It also involves providing technical documentation so that the indicator is, in some sense, reproducible.

Understanding these steps in the construction of a composite indicators helps identify where improvements are needed, and the research needed to support those improvements. This might be quite specific for one composite indicator, a matter of finding or developing better individual measures. In other cases, required research would be more generalizable – for example, concerning how best to display data or the technical documentation of the indicator.

Figure 18. Process model of the development process of composite indicators of healthcare quality, indicating which member of the development team has primary responsibility for each issue.


Core Development Team
Manager / 'politician'
Clinical/domain experts
Statistician
Data and dataset experts
User testers
Patients, public, user representatives
Revise purpose until all relevant domains have good measures
Purpose
Provisional domains of quality
Multiple iterations may be required
Selecting individual measures
Domains of quality
Domain scores
Summary scores
Reporting
Lay summary, story-telling
Website, published scores
Underlying data, individual measures
Technical documents

6.1.2 Critical appraisal of composite indicators

I used the above process map, in combination with the details of the themes of the interview study and my experience of examining technical aspects of composite indicators, as the basis of a prototype critical appraisal tool (Table 24). This was intended to help address a limitation of the existing literature on composite indicators, namely that examinations of existing composite indicators typically focus on a single specific problem, such as the validity of underling measures or the choice of measure weights, [11,81] rather than offering a broader view. There is, for example, no point in improving the selection of measures and the weighting of domains, if the way measures are standardised distorts the information provided by each individual measure [49]. A more holistic approach is necessary.

In developing a framework, one option would be to draw on existing guidelines to composite indicator development, such as the OECD Handbook [1]. Other authors have constructed their own frameworks for critiquing composite indicators, including for example Bilimoria and colleagues' work evaluating a range of US composite indicators of healthcare quality [65]. In both such cases, the derivation of the appraisal framework draws on the accumulated expertise of the authors as opposed to empirical field work (qualitative interviews of experts). A more transparent approach would be to adapt the themes that arose in my interview study, together with contributions from relevant literature (e.g. [1,18,65]) into a framework for critical appraisal (Table 24). For example, Table 24 shows a brief appraisal of the CMS Star Ratings composite indicator, drawing on the published technical methods [2], critiques of the indicator [16,65], and the results of the analysis in chapter 3.

Much of my critical appraisal of the CMS Star Ratings shown in Table 24 focuses the lack of justification for many of the technical choices involved in producing the composite indicator. Addressing such criticisms requires methods for measuring the consequences of specific technical decisions.

Table 24. Critical appraisal of the CMS Star Ratings composite indicator based on the themes identified in this interview study. Row titles are based on section titles in chapter 5, and are similar to the main steps in the process map in Figure 18.

Themes from this study	**Assessment of the CMS Star Ratings against these themes**	**Possible improvements**
Identifying the purpose of the indicator	To summarise information from existing measures on the *Hospital Compare* website in a way that is useful and easy to interpret for patients and consumers.	Make the aim more explicit. What does it mean for a summary to be useful for patients? Is it intended to help patients choose where to seek care? Or for some other purpose?
Allowing for iterative development	Feedback is sought between each release of the CMS Star Ratings, both from through a public consultation process and a technical expert panel.	Be more open to revision to address technical problems identified during the development process (see e.g. 'developing final domains of quality').
Identifying domains of quality	Domains from the *Hospital Compare* website are used.	Evaluate whether aims are suitable for the aim of the indicator.
Identifying individual quality measures for each domain	Measures from the *Hospital Compare* website are used. Structural measures, non-directional measures, measures not used in quality reporting programmes, measures which are missing for the vast majority of hospitals, and measures that overlap substantially with an included measure are excluded.	Consider whether these measures provide an adequate summary of quality for each domain. Develop new measures, or source additional measures from other data sources, so that domains are more completely described.
Developing final domains of quality and the final set of measures	Domains from the *Hospital Compare* website are used. Derivation of domain scores is based on assumption that there is a single latent variable per domain; this is not the case. A conscious decision was taken not to revise the domains.	Revise domains so that the technical assumptions behind the derivation of domain scores are met.
Standardisation of individual measures	Measures are standardised using Z-scoring.	Consider whether alternative approaches to standardisation may be more appropriate.
Combining domain scores into the summary score	Weighted arithmetic average, with fairly arbitrary weights.	Additional justification of choice of weights, perhaps through formal elicitation processes with stakeholders.
Reporting	Usable website allowing easy comparison between hospitals. Detailed technical methods document. Published statistical code and example dataset.	Additional context on meaning of assigned star ratings – should one care if a hospital receives four rather than five stars? May need to be linked to a more explicit purpose.
Competencies of the development team	Mix of clinicians, data analysts and statisticians. Engagement with outside technical and domain expertise through an expert panel, with a patient and patient advocate workgroup, and with a provider leadership workgroup.	Potentially under-represents non-medical clinical staff (although nurse representatives were involved). More structured engagement with patients and members of the public could allow improvements.

6.2 Understanding the consequences of the technical approaches used to create a composite indicator

I set out to characterise relevant challenges in the design of composite indicators. It is clear that developers and critics of composite indicators need ways not only of identifying problems with composite indicators but judging whether those problems are consequential. Many examples in the academic literature point to specific technical problems without describing the impact of these issues [11,16,62,96,196]. For example, Bilimoria and Barnard highlight that measures used in one indicator of safety are vulnerable to surveillance bias [62]: hospitals that look for problems tend to find them, but this is desirable in order to improve safety. Yet no performance measure is ever perfect, and it is unclear how consequential this bias is – simply identifying apparent flaws in a composite does not in itself demonstrate whether such limitations in fact matter.

Understanding consequences of the technical approaches, and any problems these may have, is also a useful part of justifying the technical approach used to construct a composite indicator. In part, as suggested above, this is because it allows developers to demonstrate that an apparent technical limitation has little impact. But it also allows developers to identify steps in their construction of a composite indicator that need the most careful justification.

The question, then, is how best to demonstrate the impact of the technical approach, whether flawed or simply not fully justified, on a composite indicator. There are two ways of doing this. The most common current approach appears to be an *ad hoc* assessment of the consequences of an individual or a few aspects of the technical approach, but, as I discuss below, such limited assessments are not enough. Ideally, developers should undertake a more principled, global, assessment of the technical decisions to support their chosen approach.

6.2.1 Assessing the impact of single technical decisions is often insufficient

Most assessments thus far of the consequences of technical decisions and limitations of composite indicators are one-off evaluations of the consequences of individual technical decisions. Many such assessments solely examine weighting [81,95,175], but there are examples of assessments of the impact of other technical approaches, such as Adelman's comparison of the current CMS Star Ratings with a version where hospital domain scores were derived via an efficient frontier approach [111]. These assessments calculate the composite indicator under two different technical specifications, and compare and contrast the performance of individual hospitals under the two approaches typically through some measure of correlation. Chen and colleagues' investigation into the impact of using so-called harm-based measure weights in the AHRQ PSI-90 composite indicator of hospital safety is a typical example [81], where the original and a proposed improved weighting scheme were compared in order to support the decision to move to using harm-based weights. Sometimes these assessments were broader, perhaps covering many different approaches to the same technical issue, for example Proudlove and colleagues' use of Monte Carlo simulation to examine a wide range of different possible measure weights [95].

Such analyses of the consequences of individual steps in the design of a composite indicator are necessary, and are perhaps sufficient to demonstrate the consequences of problems with the technical design of the composite. But on their own, they have major limitations as a way of comparing different, reasonable, design choices.

1. Comparing the importance of the consequences of two different technical steps is conceptually challenging. If there is uncertainty about both standardisation and weighting, for example, then the standardisation approach applied when assessing the impact of weighting will affect the results.
2. It often leads to separate technical issues being conflated in a single change. For example, the assessment of the impact of harm-based weights conflated the issue of what the weights should be based on and how those weights should be derived.
3. Each technical step can give many different technical specifications, but many of these analysis compare just two different indicator specifications.

6.2.2 A proposal for a more principled approach to examining the technical specification of composite indicators

Addressing the limitations of a typical examination of the consequences of the specification of composite indicators requires adopting a wider view of such sensitivity analysis. The analyses described in Chapter 3 and 4 represent a step toward this objective, bridging the gap between the single-step analyses that form the majority of the literature and the type of global sensitivity analysis that would be ideal. Yet my analyses were still quite limited in scope, in that I examined just two different approaches to grouping of measures and standardisation of measures.

Developers of composite indicators are in a position to carry out comprehensive global sensitivity analyses, and I briefly outline what that might look like below. Such sensitivity analysis requires an understanding of the current constraints on the design of the indicator – for example, due to a policy choice to limit to a certain set of technical approaches, or due to the preferences of the intended audience. Once current constraints are understood, developers are left with a (wide) range of technical specifications that all appear reasonable. The idea behind a global sensitivity analysis is to simultaneously compare this entire range of reasonable technical specifications, and it should:

- Allow the importance of multiple different decisions to be directly compared, so that inconsequential issues can be justifiably ignored and consequential issues can be examined in more detail.
- Allow separate decisions (e.g. Choice of approach and the implementation of that approach) to be examined separately.
- Allow simultaneous assessment of many different approaches.

The Monte Carlo simulations of multiple different technical specifications I carried out for the CMS Star Ratings in Chapter 3 addressed the second two points, but did not make it easy to directly compare the importance of different decisions. Proposals for 'multiverse analyses' in psychological research have a similar motivation [197,198]. But there is already an approach that satisfies all three criteria: variance-based sensitivity analysis [199,200], also known as the Sobol method. There is even an example of this method being applied to composite indicators [127]. Section 6.2.2.1 gives a brief description of this approach.

Wider application of global sensitivity analysis in development of composite indicators would mean that developers identified the most consequential questions around the technical specification during the development process. This would give developers clear options for defending the design of the composite indicator. In some cases, they would be able to demonstrate that the uncertainty about how best to derive the composite indicator had little impact on judgements about hospital quality. In other cases, they would be able to identify the most influential aspects of the design, and carry out further work to find the best way to handle it. Finally, in some cases there may be multiple sensible approaches with no realistic way to choose between them, and in such cases they could either average hospital performance over the different specifications (similar to Hota and colleagues' suggestion of averaging over multiple composite indicators [107]) or present interval summaries of hospital performance (similar to Schang and colleagues suggestion of ranking intervals [79]).

6.2.2.1 A technical aside: Variance-based sensitivity analysis

When examining multiple technical decisions, it is often desirable to understand which of these decisions has the most impact on the scores and ranks of a composite indicator. The analyses presented in chapter 3 potentially allows this for a small number of decisions, but would clearly be challenging if a wider number of decisions were considered. In this short technical aside I briefly explain one approach that allows simultaneous estimation of the sensitivity of scores on a composite indicator to multiple different technical decisions: variance-based sensitivity analysis.

Variance-based sensitivity analysis was initially developed to show the sensitivity of mathematical functions to their input values [200]. But at face value a composite indicator is not a mathematical function; it is a series of technical choices chained together. From a different perspective, however, this chain of technical choices is just a more complicated function.

Consider a simple composite indicator with just a single subdomains. There are a set of candidate measures $\boldsymbol{M}_0$ of which a subset $\boldsymbol{M}_1$ are used in the composite. These measures are transformed to a set of measures $\boldsymbol{T}_1$ on a consistent scale, and then some aggregation function is applied to get the final composite score Y. Hence there is a selection function $f_{select}(\boldsymbol{M}_0) = \boldsymbol{M}_1$, a standardisation function $f_{stand}(\boldsymbol{M}1) = \boldsymbol{T}_1$, and an aggregation function $f_{combine}(\boldsymbol{T}_1) = Y$. But these can all be chained together to get the single function:

$$g(\boldsymbol{M}_0) \equiv f_{combine}\left(f_{standard}\left(f_{select}(\boldsymbol{M}_0)\right)\right) = Y$$

The choice of selection, standardisation and aggregation functions are input values for the higher-level function $g(\boldsymbol{M}_0)$, so variance-based sensitivity analysis can be applied.

Think about the composite score of a single hospital, y_{ijk}, when calculated under one of i different approaches to selection (f_{select_i}), j different approaches to standardisation (f_{stand_j}) and k different approaches to aggregation ($f_{combine_k}$). So,

$$g_{ijk}(\boldsymbol{M}_0) \equiv f_{combine_k}\left(f_{standard_j}\left(f_{select_i}(\boldsymbol{M}_0)\right)\right) = y_{ijk}$$

It is straightforward to calculate the variance V of the hospital score across the choices of selection functions, standardisation functions and combination functions. But also, because this variance is coming from these independent functions, it can be decomposed into:

- First order variances $V_{select}, V_{standard}$ and $V_{combine}$ coming solely from uncertainty about selection, standardisation, or combination of measures.
- Higher-order variances coming from uncertainty in multiple different processes, e.g. $V_{select,standard}$

Sensitivity to the selection of measures, choice of standardisation functions, and approach to aggregating measures can then be summarised in terms of the first-order and total sensitivity indices. For the selection of measures, the first order sensitivity index S_{select} and total sensitivity index $S_{T\ select}$ are given by

$$S_{select} = {}^{V_{select}}/_{V}$$

$$S_{T\ select} = S_{select} + S_{select,stand} + S_{select,combine} + S_{select,stand,combine}$$

The mathematical details of the calculation of these various variances are set out elsewhere [127,201]. In practice, if the different options are discrete, the variances can be estimated directly from a Monte Carlo simulation: the variance in the score can be calculated by selecting different measures for each combination of standardisation functions and aggregation approaches, and then taking the mean across each of these approaches. Often there will be a mix of continuous options – such as choice of weights – and discrete options – the grouping of measures into domains, perhaps. This makes the process more challenging, but approximate answers can still be found – for example using Sobol sequences [202].

Variance-based sensitivity analyses, and broadly similar approaches including the Monte Carlo simulations presented in this thesis, allow inferences about the importance of different technical decisions. But such quantitative methods can not tell us which technical approach should be used. Quantitative tools are only one part of the answer; understanding complex methodological questions such as the design of composite indicators requires a broader perspective.

6.3 Researching complex methodological questions and the interplay between quantitative and qualitative approaches

Composite indicators are naturally quantitative measures and, as set out above, advanced quantitative methods are useful when examining their design. But many of the problems with composite indicators arise in relation to how they are understood, interpreted, and acted upon – all issues closer to the social sciences than the statistics [203]. As my thesis has shown, a multi-method approach applying qualitative methods informed by detailed quantitative studies addresses the limitations of either quantitative or qualitative approaches alone. Much research on composite indicators tends to focus in on specific technical issues – similar to my own early results chapters. Such research is vital, but its narrow perspective means it often misses potentially far more fundamental and consequential flaws. Qualitative research is therefore a natural contender to broaden and deepen understanding. However, if it is not informed by an understanding of the technical issues, it may fail to explore the full technical challenge.

6.3.1 Limitations of (solely) quantitative research on composite indicators

Producing a composite indicator involves making multiple different technical choices, but in quantitative research it is common to focus on individual problems in isolation. Many useful quantitative methods papers relating to composite indicators suffer from such one-sided focus. Austin, Lee and Leckie describe an interesting approach to using multivariate Bayesian random-effects logistic regression models for hospital profiling [189], but (among various issues) it is only applicable if the measures it is based on form an appropriate basis for such a profile. Longford discusses the application of decision theory in rating institutional performance [190], but this relies on eliciting preferences about the potential repercussions of different performance classifications which is by no means straightforward. Landrum and colleagues discuss construction of composite indicators reflecting latent variables in sets of performance measures [191], but do not discuss how to ensure this latent variable is actually measuring a relevant construct.

Similarly, many critiques of composite indicators focus on specific quantitative aspects but miss important theoretical points. One set of examples involves discussions about ‘disagreement’ between different composite indicators (e.g. [40,107,110]), without noting that these different composites are trying to measure different things. For example, Hota and colleagues measured the agreement between the US News Best Hospitals and the Leapfrog Safety Grade [107], among other composites. But one of these aims to measure overall quality, while the other aims to measure overall safety – two related but conceptually distinct constructs.

Both types of quantitative research are valuable, but on their own are not enough to help improve how composite indicators are designed. Quantitative assessments of the impact of different quantitative approaches and of the differences between different composite indicators help to understand the different methods and can point to possible flaws, but do not on their own provide a guide to development. And while new methods for approaching specific technical steps are needed, because they allow the field to move toward composite indicators that are as interpretable and actionable as possible, they may not resolve the challenges of designing or reporting a composite from start to finish.

Drawing broader conclusions about how composite indicators should be designed and reported requires engagement with both quantitative and social science aspects of composite indicators. This points toward a multi-method approach.

6.3.2 Multi-method research from a quantitative perspective

Applied quantitative research, including the development of a composite indicator, needs a theoretical basis. That creates an opportunity for what might be termed applied qualitative research, aimed at generating micro-theory about how the research can be approached [204]. Indeed, Kuc-Czarnecka, Lo Piano and Saltelli set out to identify ingredients of a possible theory of composite indicators [203], conceptualising composite indicators as a form of quantitative story-telling.

In conducting qualitative research about research, there are advantages to having quantitative methodologists closely involved, and perhaps ideally carrying out the research themselves. In this way, my interview study with experts in performance measurement drew on my statistical expertise and familiarity with quantitative research. Such applied qualitative research might be considered a form of multi-method research, because while it uses qualitative methods, it builds directly on the results of relevant quantitative analyses.

The first reason to involve quantitative methodologists directly in carrying out this research is to avoid misunderstandings about technical aspects. Concepts from quantitative research are technically difficult and hard to understand. Even concepts ubiquitous as the *p*-value are commonly misunderstood [205], and misunderstandings about the rationale behind randomisation lead to enormous amounts of confusion and debate [206]. Asking a quantitative methodologist to carry out interviews and qualitative analysis is more reasonable than asking a qualitative researcher to become an expert in quantitative methods, although the project will require substantial input from researchers with true qualitative expertise. These concerns may explain the small number of qualitative studies carried out by quantitative methodologists, for example on adaptive trial designs [155,156].

Another reason for having a quantitative methodologist to carry out this type of research is that it leads to better interviews, that flow more freely. I found that shared technical terminology (‘jargon’) and understanding of terms made it easier understand what the interview participant was saying, and easier to ask appropriate follow-up questions. The fact that professional roles impact on interactions with participants and what participants say has been noted before [207], and when researching methodological issues I believe this is an advantage.

One of my main aims in this thesis was to characterise challenges in the design of composite indicators, but the discussion so far has addressed general problems and strategies for researching complex problems. The next section discusses specific challenges, and sets out possible methods to address these issues.

6.4 Common problems with existing composite indicators – and potential strategies for mitigation

Composite indicators promise a simple, interpretable overview of complex sets of healthcare quality information [1]. But that may be an empty promise unless the problems described in this thesis are addressed. Though clamour about flawed composite indicators and their role in comparing organisations persists [3,11,12,42,62,67,208], they continue to be widely deployed. Implementing improvements to the design and reporting of composite indicators and other performance measures will require higher levels of scrutiny of decisions about individual measures of quality, their related technical specification and standards. Building on my previous exposition of problems with composite indicators earlier in the thesis and prior work (section 1.6 and [16]), summarised in Table 25, and existing principled frameworks for developing composites (e.g. Bottle and Aylin's book [18], the OECD Handbook [1], Profit and colleagues work [50]), in this section I reflect on how challenges relating to methodological transparency, purpose-led design, good statistical practice, and data visualisation can be addressed.

Table 25. Requirements, steps forward and remaining challenges for robust and useful composite indicators. Taken from Barclay, Lyratzopoulos and Dixon-Woods, The Problem with Composite Indicators [16].

Requirement	Steps forward	Remaining challenges
Transparency *The principles and theory underlying the composite indicator must be clear*	Being clear about who is involved in making decisions in developing the composite indicator.	Many stakeholders may be involved. The design may evolve in unexpected ways over time.
	Fully describing the decision-making process, reporting the reasons and justifications for the decisions made.	
Purpose-led design *The composite indicator must plausibly measure what it sets out to measure*	Selecting individual measures to cover the full range of services intended to be measured by the composite.	Identifying appropriate individual measures. Appropriate measures may not exist for all areas included in the composite.
	Choosing weights that reflect the relative importance of the different quality measures.	Balancing the weighting system against competing priorities.
Technical reproducibility *The composite indicator must be reproducible using the raw data and the published methodology*	Providing clear and comprehensive technical documentation.	
	Reporting full definitions of the individual underlying measures and how they are combined.	Individual measures may only be available from sources that do not fully document the details, but these measures should not be used in the composite.
	Publishing the code used in data processing and statistical analysis.	
Statistical fitness *Individual measures must be adequately adjusted for case-mix, have acceptable statistical reliability, and be appropriately standardised to consistent scales*	Performing appropriate statistical case-mix adjustment.	Accurate patient-level data may not exist for important case-mix factors. Adequate statistical case-mix adjustment may not be possible Interpretable results may require further processing.
	Using reporting periods long enough to give acceptable reliability.	Longer reporting periods may be necessary to increase reliability, but impedes use in driving quality improvement.
	Standardising measures to consistent scales in a principled way that preserves the useful information in the underlying measures.	Understanding what good and bad performance in the real world looks like on each measure.

6.4.1 Lack of methodological transparency

My research suggests that transparency about methods, including design choices about technical issues, is key to addressing many current problems with composite measures. The existing literature does not offer clear directions regarding the guideline development process for composite indicators, not least because of the fragmented and siloed way it has developed.

At present, for example, guides to the development of composite indicators are often written by specific professional groups (e.g. [1,18,50]), and may reflect the singular perspective of that profession. Papers by statisticians on development of composite indicators often focus solely on methods of combining a pre-existing set of performance measures into the composite (e.g. [189–191,209]), but do not address the challenges of identifying these measures in the first place. On the other hand, papers written from a quality improvement perspective may primarily address measure selection (e.g. [97,188]), but may ignore some of the more technical challenges of combining these measures into an appropriate summary measure. Health economists may focus on the problem of eliciting appropriate weights [49,79,80], and policy researchers may debate whether flawed measurement can still lead to improvement in performance [73,210]. The extent to which guides from economics and education [1,51,211,212], where composite indicators are extensively used, may be applied to quality healthcare is also unclear. Similarly, appropriate methods for displaying or accounting for uncertainty in final composites are not obvious; good statistical practice for developing individual performance measures is relatively well-understood [83], but there is little or no guidance in the literature on handling the complex, multilevel, missing data problems encountered when developing composites; the implications of measure standardisation seem poorly understood; and the visualisation of results often appears unhelpful.

An important element of transparency is that composite indicators should be presented with accompanying displays of statistical uncertainty [83]. However, this is rarely done and indeed is challenging to conceptualise. Uncertainty in composite indicators arises both from statistical noise and from the way individual measures are chosen, standardised and aggregated. Sensitivity analyses should investigate whether reasonable alternative methods would substantially alter organizational rankings [84], and the results of these analyses should be reported [79]. This may require addressing the current lack of scientific consensus about how best to represent uncertainty for star-ratings and other categorical performance classifications. Interval estimates, such as confidence intervals, are the typical way of representing uncertainty and can certainly be calculated for ranks and scores on composite indicators. They may be less useful for indicators presented as star-ratings; it may be better to discuss the probability that a rating is correct, or too high or low, drawing on Bayesian approaches to ranking hospital performance on individual measures [213]. One alternative is to build a formal decision model based on the harm caused by misclassifying a hospital as better or worse than it is [190,214], but in practice this may raise further problems relating to how harms are judged.

6.4.1.1 Components of methodological transparency

A key theme in my thesis is the need, in order to report a composite indicator with true methodological transparency, for justification of the technical approach used. In practice such justification is often lacking in the documentation of composite indicators. In producing justifications for technical choices, transparency is likely to involve three distinct components.

1. Identifying the possible technical approaches that could be used in calculating a composite.
2. Explaining the implications of different technical choices. The types of analyses I report in Chapters 3 and 4 exemplify this.
3. Identifying how choices might be made between the different technical approaches. This would be a bespoke process for each composite indicator, following one of three broad families of approaches.
    a. The different technical choices considered in the second component may turn out to have little influence over apparent hospital quality, which was for example the case for the domain weights used for BABY-MONITOR [175]. Arguably, in such cases it is reasonable to make an arbitrary choice between the different options, although some justification may be required.
    b. Relevant external information may be helpful for choosing the most appropriate technical design. For example, for a composite based on process measures such as the SSNAP indicators, it may be desirable to choose the technical choices that lead to a composite with the strongest associations with costs and outcomes.
    c. Sometimes decisions are influential and there is no relevant set of external measures to help guide the design. In these cases, as discussed by my

interview participants in Chapter 5, the design of the indicator should be guided by the preferences of the users. This may involve allowing users of the composite to specify portions of the design themselves [174].

6.4.2 Lack of clarity about the purpose of the composite indicator

Throughout my research, I have identified the importance of clarity of purpose in addressing many of the challenges currently affecting composite indicators. At minimum, the aims and limitations of composite indicators should be presented alongside ratings to aid understanding of where scores and ratings come from, what they mean, and what limits their usefulness or interpretability. Clear explanation is needed of the logic underlying the development of each composite indicator, including the choice of measures, any compromises between different goals, whose views have been taken into account in producing the indicator, and how. In contrast, this does not constitute common practice in the field. Many composite indicators would be improved by reflecting the aims and preferences of the relevant stakeholders in the choice and weighting of individual measures using a clear process and explicit theory-of-change [97,215–217]. Methodological information should be readily available and clearly linked to the indicator. Yet much deeper, broader, understanding is required to turn this superficial summary into something that easily and widely usable in practice.

The CMS Star Ratings is an excellent example of a composite indicator lacking clarity of purpose, to the extent that its technical documentation and lay summary describe different purposes. Presentation of the Star Ratings on the care-compare tool describes them as *"based on how well a hospital performs across different areas of quality"* [218] (see Figure 19 for an example). But the technical methods describe the primary purpose of the Star Ratings as *"summarizing information from the existing measures on Hospital Compare in a way that is useful and easy to interpret for patients and consumers."* ([2], page 6).

The root of many of the issues with the CMS Star Ratings, including many of those I describe Table 24, chapter 1.6 and elsewhere [16], is that the existing measures on Hospital Compare are not chosen to provide a good overview of overall hospital quality. Measures on Hospital Compare are there because they are used in one of four public-reporting or pay-for-performance programmes [219]. Yet performance on these programmes may not reflect overall care quality, even if the performance measures do give some information about certain aspects of quality at each hospital. The developers are trying to produce something meaningful based on information that is simply not adequate for the task.

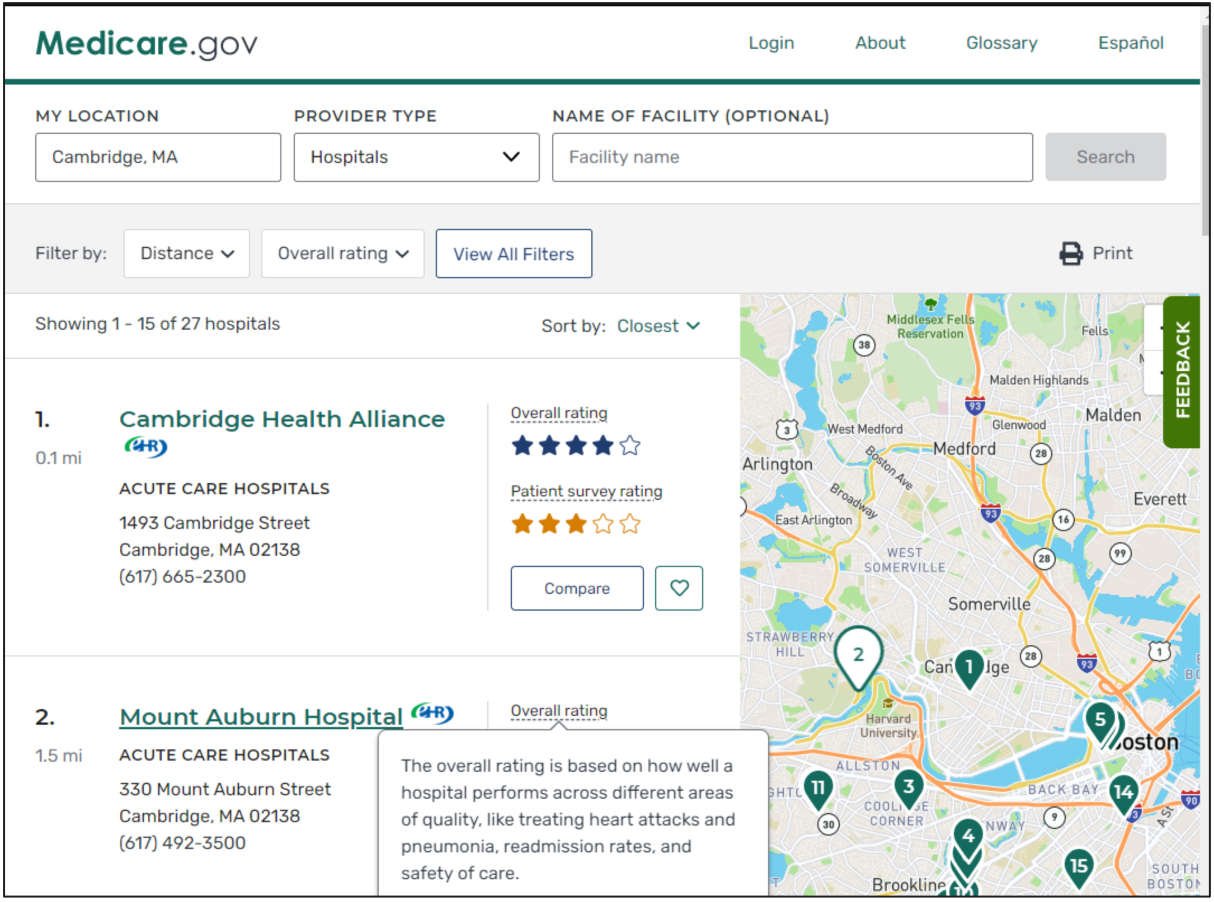
Clarity about abstract constructs such as quality of care is more challenging to achieve than for more concrete, measurable, constructs such as mortality rates. As quality itself is not directly measurable, the only way to measure it is to see if hospitals have characteristics that are expected to reflect high quality care, such as good patient experience scores or, indeed, low mortality rates. But the breadth of a construct like the quality of an entire hospital makes identifying an appropriate set of markers of quality difficult. Clarity of purpose about broad composites that aim to measure constructs such as the quality of care of an entire hospitals may not be practical. Instead, composite indicators may be most appropriate for more limited aspects of quality or more bounded lines of service.

Consider the two exemplar composite indicators I examine in detail in chapters 3 and 4. The CMS Star Ratings aim to measure overall quality, classifying hospitals as one, two, three, four or five stars. The SSNAP score and level aim to measure the quality of acute stroke care, classifying hospitals as A, B, C, D or E.

The CMS Star Ratings raise the question: what does it mean for a hospital to be good quality? This is not an easy question to answer. And so knowing simply that one hospital receives three stars and another receives five stars does not give much useful information about the quality of care that any given patient would expect to receive. It is hard to know whether this difference matters.

In contrast, the SSNAP score and level are an answer to the question: what does it mean for an acute stroke service to be good? It is easier to see how this question can be answered, starting from clinical guidelines and moving up. There are still challenges, and it is still difficult to understand whether a difference between an A and a C is important. But it is far easier to intuitively understand what this indicator is measuring.

Figure 19. Screencap of the CMS Hospital Compare Star Ratings for hospitals in Cambridge, Massachusetts, as presented on the medicare.gov care-compare tool.



Medicare.gov Login About Glossary Español

MY LOCATION: Cambridge, MA
PROVIDER TYPE: Hospitals
NAME OF FACILITY (OPTIONAL): Facility name
Search

Filter by: Distance | Overall rating | View All Filters | Print

Showing 1 - 15 of 27 hospitals
Sort by: Closest

1. Cambridge Health Alliance
0.1 mi
ACUTE CARE HOSPITALS
1493 Cambridge Street
Cambridge, MA 02138
(617) 665-2300
Overall rating ★★★★☆
Patient survey rating ★★★☆☆
Compare ♡

2. Mount Auburn Hospital
1.5 mi
ACUTE CARE HOSPITALS
330 Mount Auburn Street
Cambridge, MA 02138
(617) 492-3500
Overall rating

The overall rating is based on how well a hospital performs across different areas of quality, like treating heart attacks and pneumonia, readmission rates, and safety of care.

6.4.3 Sub-optimal statistical practice

My research has repeatedly identified that, to the extent that good statistical practice is well understood, composite indicators need to be compliant. Underlying measures should, at minimum, be appropriately adjusted for case-mix, assessed for possible sources of bias, and meet basic standards of inter-unit reliability [84,220,221]. The reasons for missing data should be explored, and principled approaches should be adopted to address missing data, although there is certainly scope for a detailed exploration of how missing data should be handled when producing composite indicators. Entirely missing measures (e.g. a hospital has no thrombolysis time information at all) may sometimes be handled using statistical approaches to identify common factors between measures based on the observed hospital-level correlations [166,191,222]. Missing data in individual measures (e.g. 30% of patients at a given hospital have missing thrombolysis time) may sometimes be handled using multiple imputation to predict what missing values should have been based on the available information [121,223]. The likely best solution is to refine inclusion criteria and improve data collection so that the proportion of missing data becomes negligible.

Methods of standardisation that preserve the information in individual measures as far as possible are another area where careful investigation could lead to practical improvements: clear guidance on and examples of principled standardisation, and elicitation of appropriate thresholds, may well be very valuable. Individual measures must be on the same scale before they can meaningfully be combined into an overall composite, and there are many methods of standardising collections of measures. It appears obvious that methodological choices need to be guided by an understanding of clinical best practice and the meaning of differences in performance on the individual scales. Often, it may simply be that ‘higher is better’, and so default approaches may be optimal. One default option is to standardise against the observed standard deviation (‘Z-scoring’ [83]), with the standardised measure describing how far a given hospital’s performance is from the average hospital, relative to variation across all hospitals. Another option is to standardise against the possible range of measure scores, so the standardised value describes how close a hospital is to achieving the theoretical maximum performance. But it is often possible to modify these defaults to produce a more meaningful composite, perhaps by measuring performance relative to targets or by incorporating information about the importance of achieving particular levels. In particular, it may be possible for some measures to identify clear thresholds for acceptable, good and excellent performance on a measure, as for example for some component measures of the MyNHS Overall Stroke Care Rating [69]. Interpolation between thresholds allows standardisation to a meaningful scale without the use of cliff-edge decision rules.

6.4.4 Lack of appropriate data visualisation

Appropriate data visualisation techniques may help make composite indicators more informative and useful in healthcare, perhaps building on emerging examples of composite measures and rankings outside of healthcare where the user can interactively specify measure weights on a web page and immediately see the impact on results [224]. This may allow users to make composites that reflect their own priorities, and to explore uncertainty due to the way measures are aggregated. But poorly designed visualisation may mislead users, or require more effort to understand than less attractive options. Research focused on the benefits and harms of different data visualisation strategies for performance measurement is vital.

6.5 Limitations of this thesis

The research presented in this thesis has several limitations. These range from strategic issues that became clear following completion of the study to more specific limitations with the quantitative and qualitative analyses.

6.5.1 Strategic limitations

In this thesis I adopted a multi-method approach, where independent quantitative and qualitative studies were synthesised in this overall discussion chapter [225]. This was a pragmatic choice, as it meant each of my studies could proceed independently rather than the start of one study needing to be delayed until another had finished.

A mixed-method approach, where the results arising from one methodological approach informed the design and interpretation of the next study, would have had some benefits. This would have been a valuable strategy: the interview study could have been used to identify the range of decisions involved in producing a composite indicator, potentially allowing my quantitative studies to explore a wider range of plausible options than the chosen specific subset of technical approaches. This was not possible in the time I had available.

6.5.2 Limitations of the quantitative analyses

The limitations of my quantitative analyses were primarily driven by a pragmatic choice to focus on a manageable number of composite indicators, technical choices, and possible options for each technical choice.

I examined only two composite indicators which were sampled from a much greater number in current or recent use. My specific results on the sensitivity of these composite indicators to the technical approaches I considered may not apply to other composite indicators. Yet the paradigmatic implications (e.g. regarding the importance of methodological transparency about choices and assumptions made by the composite indicator developers; the role of empirical examination of consequences of key decisions by comparing to alternative choices; and the role of sensitivity analysis), apply more widely.

I examined the consequences of only three aspects of composite indicator specification (domain weights, individual measure standardisation, and the grouping of measures into domains) of several other such aspects. My results demonstrated that the examined technical decisions can be important, but studies of the impact of other technical decisions may additionally be worthwhile.

An issue that deserves reflection is the handling of missing or unreported measure information. I did describe the extent of this issue, particularly for the CMS Star Ratings. My results suggested that the amount of missing or unreported measure information may influence the apparent performance of hospitals. This highlights the need for comparative studies examining the impact of different ways of handling missing domain data when deriving the composite and, additionally, comparing the current approach with methods based on proxy information or methods such as multiple imputation. However principled statistical approaches to handling of missing data such as multiple imputation would require access to patient-level data. Such data were not available for the case studies chosen for my thesis, but would be a valuable focus for future work.

A further limitation was that, for the technical decisions on how measures were grouped into domains and for the standardisation of measures, I only considered two out of many possible options. For the grouping of measures into domains, options such as a consulting with a clinical reference group on appropriate grouping of the performance measures could in principle be justified. When considering standardising individual measures I compared Z-scores and reference-based standardisation approaches, but several other approaches could have been considered: For example, Jacobs and colleagues discuss nine different approaches to standardisation [49], of which Z-scores and reference-based standardisation are two. Another approach might rely on funnel plots, which are also frequently used in scoring hospital performance [83]. In this setting, use of funnel plots in measure standardisation represents an extension to a Z-score approach, additionally accounting for variation in hospital performance introduced by chance.

My quantitative analyses explicitly ignored chance variation. This can have an important impact on composite indicator ratings, with for example Venkatesh and colleagues reporting that chance alone could lead to between one in five and one in three hospitals changing Star Rating category in the CMS Star Ratings [2], although changes beyond neighbouring ratings were rare. While uncertainty due to the technical approach is not impacted by chance variation, and so my analyses are not affected by failing to examine chance variation, the converse is not true [49]: certain technical approaches are more robust to chance variation.

Where possible, future analyses should aim to examine uncertainty introduced both by technical choices and by random chance.

Finally, I only examined composite indicators in current use which tend to be driven by policy-makers as opposed to methodologists. Another potential route of inquiry would involve examining methodologically-innovative approaches such as Austin *et al*'s multivariate Bayesian approach or Longford's decision theory approach [189,190]. Both multivariate Bayesian and formal decision-theoretic approaches can account for chance variation when producing composite scores. The multivariate Bayesian approach combines separate measures via a single statistical model, allowing more informative composite summaries that directly incorporate the impact of chance variation, such as the probability of a hospital being better than average on all individual quality measures. The decision theory approach aims to incorporate information on the costs and benefits of giving hospitals specific ratings, such that the risks of chance misclassification are accounted for when assigning ratings. Analyses of more innovative approaches would be valuable, but comparing results from these more innovative methods with those from more traditional approaches may be challenging.

6.5.3 Limitations of the qualitative analysis

My qualitative study focused on technical aspects of the design of composite indicators, and accordingly engaged only with informers with professional expertise. Patient and public involvement in what was effectively a methodological study seemed at risk of being tokenistic [226], although might have had some advantages including providing views on the types of ratings patients use: a recent discussion of patient involvement in methodological research highlighted that one valuable contribution was in selecting organisations for inclusion in the sample [227]. It has been proposed that indicator schemes are rarely used by patients in practice [228]. There remains a need for qualitative studies focusing on the types of information patients and users of composite indicators find useful, and on the appropriate format for that information. Beyond patients and members of the public, such studies should ideally include clinicians, board and executive-level stakeholders, and system stakeholders.

An additional limitation of my qualitative study was that there was no double coding or other formal checks of my application of the coding framework. Yet my data analysis was informed by extensive discussion of themes and data between myself and my supervisors, including reviews of data extracts and summaries. While I did not undertake a formal double-coding procedure, thanks to these steps I am confident my application of the coding framework was correct.

6.6 Conclusions: routes toward better designed and reported composite indicators

My thesis characterises technical and conceptual challenges in the design of composite indicators, and explores how reporting of composite indicators could be improved. My quantitative analyses provide a detailed assessment of the consequences of alternative technical specifications on apparent hospital performance based on two existing composite indicators. This has demonstrated that technical choices can sometimes be highly consequential, and some existing, high profile schemes may not be a good guide to underlying quality – especially for hospitals which do not report on all of the performance measures used to produce the composite. My analysis also demonstrated the converse: sometimes variations in technical choices may not matter too much. A practical implication of my thesis is the need for careful sensitivity analysis of each composite indicator.

My qualitative interview study explored the way that experts in performance measurement would approach the development of a composite indicator. The results pointed to many areas of apparent consensus, especially around the overall way that participants conceptualised the process of designing a composite indicator. This, in combination with evidence from the literature and insights derived from my quantitative studies, allowed me to produce a process map for typical development of a composite indicator in healthcare. I adapted this map into a prototype critical appraisal tool and showed how it might be used in practice.

My research suggests two key recommendations for improvements in current practice in the design of composite indicators in healthcare. As noted above, the first is the need for sensitivity analysis; the methods I used in my quantitative studies could be used as a template for this kind of work. As my quantitative studies demonstrated, it allows understanding of the stability of composite indicators under alternative specifications. In combination with the emphasis raised by participants in my interview study on iterative development, perhaps involving prototype composite indicators, sensitivity analysis may allow for production of more robust composite indicators. Failing that, it allows for reporting of composite indicators that can clearly identify the importance of certain assumptions.

My second key recommendation is the need for clarity of purpose. In part, as I set out in section 6.4.2, this may mean accepting that it is impossible to produce composite indicators that represent certain constructs. Broad constructs including overall quality of care do not appear possible to understand in a way that allows the development of an appropriate composite indicator. Composite performance measurement in healthcare would be easier to understand and produce if it focused on more specific or bounded issues, such as the quality of specific aspects of clinical services. The Sentinel Stroke National Audit Programme composites of quality of acute stroke care are one such example [8], as are the various service-line quality composites produced by US News & World Report [39].

Improving many aspects of the design and reporting of composite indicators requires further research. One key aspect relates to good statistical practice in developing a composite, and particularly how to handle the types of missing data problems that arise with routine performance measurement. Typical statistical approaches to handling missing data rely on assumptions about the reasons for missing data – essentially, that it is happening at random, or that any non-random part can be accounted for with known data [121] – that rarely appear sensible with healthcare performance measures. There is a need for guidance on when and how developers of composite indicators can account for typical missing data issues, and in which cases the challenges are so severe that the missing data makes performance measurement impossible.

An additional area for further research is the visualisation of composite indicators. Interview participants highlighted the value of user testing in identifying how best to report performance measures. Yet despite the proliferation of modern, interactive dashboards of performance measures and composite indicators (e.g. [218,229,230]), there is little consistency of presentation and scant evidence about the best way to present healthcare performance information.

A key motivation for my research was to improve the transparency with which composite indicators are reported. There are technical and pragmatic aspects to this. The technical aspect is identifying ways of presenting uncertainty that are interpretable to the users of the composite, and identifying good statistical practice as discussed above. The pragmatic aspect is understanding how indicators are currently developed and reaching consensus over how to report this complicated and iterative process – and thereby developing a reporting guideline. The results of the interview study, and the process model in Figure 18, provide a starting point for such a guideline, but are not on their own sufficient. Developing a reporting guideline usually requires reaching consensus over its contents [148] – often via a Delphi study [152], a multi-round research study carried out with a panel of experts. There are many applications of Delphi studies and there little standardisation in how the technique is applied [231–233]. But in a guideline context this would typically mean identifying the issues that the panel as a whole agreed were both relevant and important [148]. The results of my interview study make a natural starting point for this Delphi process, and the next steps are to seek wider input via a Delphi study, with the aim of reaching a broader consensus over the most important aspects of technical reporting.

References

1 Nardo M, Saisana M, Saltelli A, *et al.* Handbook on Constructing Composite Indicators. Published Online First: 2005. doi:https://doi.org/10.1787/18152031

2 Venkatesh AK, Bernheim SM, Hsieh A, *et al.* Overall Hospital Quality Star Ratings on Hospital Compare Methodology Report (v2.0). 2016.internal-pdf://202.33.59.0/Star_Rtngs_CompMthdlgy_052016.pdf

3 Griffiths A, Beaussier A-L, Demeritt D, *et al.* Intelligent Monitoring? Assessing the ability of the Care Quality Commission's statistical surveillance tool to predict quality and prioritise NHS hospital inspections. *BMJ Qual Saf* Published Online First: 2016. doi:10.1136/bmjqs-2015-004687

4 NHS England Analytical Team. Statement of Methodology for the Overall Patient Experience Scores (Statistics). 2014.internal-pdf://65.255.172.213/Methods-statement_20150420.pdf

5 AD. Zo komt de Top 100 tot stand | Ziekenhuis Top 100. AD.nl. 2019.https://www.ad.nl/binnenland/zo-komt-de-top-100-tot-stand~a25ddbe9/ (accessed 29 Oct 2020).

6 Olmsted MG, Geisen E, Murphy J, *et al.* Methodology U.S. News & World Report 2017-18 Best Hospitals: Specialty Rankings. 2017.internal-pdf://0597828054/BH_Methodology_2017-18.pdf

7 The Leapfrog Group. Leapfrog Hospital Safety Grade Scoring Methodology Spring 2017. 2017.http://www.hospitalsafetygrade.org/media/file/HospitalSafetyGrade_ScoringMethodology_Spring2017_Final2.pdf

8 Sentinel Stroke National Audit Programme. Moving the dial of stroke care: the sixth SSNAP annual report. 2019. https://www.strokeaudit.org/Documents/National/Clinical/Apr2018Mar2019/Apr2018Mar2019-AnnualReport.aspx

9 Staff. The effects of CMS' star ratings: 5 things hospital leaders should know. Becker's Hosp. Rev. 2016.https://www.beckershospitalreview.com/hospital-management-administration/the-effects-of-cms-star-ratings-5-things-hospital-leaders-should-know.html (accessed 19 Aug 2020).

10 Goddard M, Jacobs R, Leatherman S. Using composite indicators to measure performance in health care. In: Smith PC, Mossialos E, Papanicolas I, eds. *Performance measurement for health system improvement*. Cambridge: : Cambridge University Press 2010. 339–68. doi:10.1017/CBO9780511711800.013

11 Bilimoria KY, Barnard C. The new cms hospital quality star ratings: The stars are not aligned. *JAMA* 2016;**316**:1761–2. doi:10.1001/jama.2016.13679

12 Shekelle PG. The English star rating system - failure of theory or practice? *J Health Serv Res Policy* 2005;**10**:3–4. doi:10.1177/135581960501000102

13 Castellucci M. CMS star ratings disproportionately benefit specialty hospitals, data show. Mod. Healthc.

2018.https://www.modernhealthcare.com/article/20180314/NEWS/180319952/cms-star-ratings-disproportionately-benefit-specialty-hospitals-data-show (accessed 19 Aug 2020).

14 Spiegelhalter D. The mystery of the lost star: A statistical detective story. *Significance* 2005;**2**:150–3. doi:10.1111/j.1740-9713.2005.00126.x

15 Shahian DM, Wolf RE, Iezzoni LI, *et al.* Variability in the Measurement of Hospital-wide Mortality Rates. *N Engl J Med* 2010;**363**:2530–9. doi:10.1056/NEJMsa1006396

16 Barclay ME, Dixon-Woods M, Lyratzopoulos G. The problem with composite indicators. *BMJ Qual Saf* Published Online First: 2018. doi:10.1136/bmjqs-2018-007798

17 Spiegel AD, Springer CR. History: Babylonian medicine, managed care and codex hammurabi, Circa 1700 B.C. *J Community Health* 1997;**22**:69–89. doi:10.1023/A:1025151008571

18 Bottle A, Aylin P. *Statistical Methods for Healthcare Performance Monitoring*. CRC Press 2016.

19 Nightingale F. *Notes on Hospitals*. London: : Longman, Green, Longman, Roberts, and Green 1863. https://archive.org/details/notesonhospital01nighgoog/page/n10/mode/2up (accessed 29 Oct 2020).

20 Nightingale F. *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army*. Harrison and sons 1858. https://books.google.co.uk/books/about/Notes_on_Matters_Affecting_the_Health_Ef.html?id=IHLfQAAACAAJ&redir_esc=y (accessed 29 Oct 2020).

21 Neuhauser D. Ernest Amory Codman MD. *BMJ Qual Saf* 2002;**11**:104–5. doi:10.1136/QHC.11.1.104

22 The Joint Commission. The Joint Commission: Over a century of quality and safety. 1990.

23 McKenzie M, Weir R, Richardson T, *et al. Further studies in hospital and community*. The Nuffield Trust 1962. https://www.nuffieldtrust.org.uk/research/further-studies-in-hospital-and-community (accessed 29 Oct 2020).

24 Donabedian A. Evaluating the Quality of Medical Care. *Milbank Mem Fund Q* 1966;**44**:166–206. doi:10.2307/3348969

25 Lowry S. Focus on performance indicators. *Br Med J* 1988;**296**:992–4. doi:10.1136/bmj.296.6627.992

26 Yates JM, Davidge MG. Can you measure performance? *Br Med J* 1984;**288**:1935–6. doi:10.1136/bmj.288.6434.1935

27 Goldstein H, Spiegelhalter D. League tables and their limitations: Statistical issues in comparisons of institutional performance. *J R Stat Soc Ser A Stat Soc* 1996;**159**:385–443.http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/statistical-issues-for-league-tables.pdf

28 Epstein A. Performance Reports on Quality -- Prototypes, Problems, and Prospects. *N Engl J Med* 1995;**333**:57–61. doi:10.1056/NEJM199507063330114

29 Hofer TP, Hayward RA, Greenfield S, *et al.* The unreliability of individual physician 'report cards' for assessing the costs and quality of care of a chronic disease. *JAMA* 1999;**281**:2098–105. doi:10.1001/jama.281.22.2098

30 Green J, Wintfeld N. Report Cards on Cardiac Surgeons — Assessing New York State's Approach. *N Engl J Med* 1995;**332**:1229–33. doi:10.1056/nejm199505043321812

31 Kassirer JP. The Use and Abuse of Practice Profiles. *N Engl J Med* 1994;**330**:634–6. doi:10.1056/nejm199403033300910

32 Hill CA, Winfrey KL, Rudolph BA. 'Best Hospitals': A Description of the Methodology for the Index of Hospital Quality. *Inquiry* 1997;**34**:80–90.https://www.jstor.org/stable/29772672 (accessed 29 Oct 2020).

33 Comarow A. In search of the best: The rankings explained. US News World Rep. 1998;:65.https://web-b-ebscohost-com.ezp.lib.cam.ac.uk/ehost/detail/detail?vid=0&sid=b0ec0d96-8dfd-4d30-adc8-4e07ff975075%40pdc-v-sessmgr01&bdata=JnNpdGU9ZWhvc3QtbGl2ZSZzY29wZT1zaXRl#AN=854662&db=bsu (accessed 29 Oct 2020).

34 Hutman RF. IBM Watson Health Announces 100 Top Hospitals. IBM News Room. 2019.https://newsroom.ibm.com/2019-03-04-IBM-Watson-Health-Announces-100-Top-Hospitals (accessed 29 Oct 2020).

35 AHRQ QI Composite Measure Workgroup. Patient Safety Quality Indicators Composite Measure Workgroup Final Report. 2008.https://www.qualityindicators.ahrq.gov/Downloads/Modules/PSI/PSI_Composite_Development.pdf

36 Austin JM, D'Andrea G, Birkmeyer JD, *et al.* Safety in Numbers: The Development of Leapfrog's Composite Patient Safety Score for U.S. Hospitals. *J Patient Saf* 2014;**10**:64–71. doi:10.1097/PTS.0b013e3182952644

37 Jha A. Hospital Rankings Get Serious. An Ounce Evid. | Heal. Policy. 2012.https://blogs.sph.harvard.edu/ashish-jha/2012/08/14/hospital-rankings-get-serious/ (accessed 29 Oct 2020).

38 Consumer Reports. How We Rate Hospitals. 2017.http://article.images.consumerreports.org/prod/content/dam/cro/news_articles/health/PDFs/Hospital_Ratings_Technical_Report.pdf

39 Binger T, Martin G, Majumder A, *et al.* Methodology: US News & World Report 2020-2021 Best Hospitals Procedures & Conditions Ratings. US News World Rep. 2020.https://health.usnews.com/media/best-hospitals/BHPC_Methodology_2020-21 (accessed 29 Oct 2020).

40 Wang DE, Tsugawa Y, Figueroa JF, *et al.* Association between the centers for medicare and medicaid services hospital star rating and patient outcomes. *JAMA Intern Med* 2016;**176**:848–50. doi:10.1001/jamainternmed.2016.0784

41 Jha AK. The Stars of Hospital Care: Useful or a Distraction? *JAMA* 2016;**315**:2265. doi:10.1001/jama.2016.5638

42 Mannion R, Davies H, Marshall M. Impact of star performance ratings in English acute

hospital trusts. *J Health Serv Res Policy* 2005;**10**:18–24. doi:10.1177/135581960501000106

43 Roland M, Guthrie B. Quality and Outcomes Framework: What have we learnt? *BMJ* 2016;**354**. doi:10.1136/bmj.i4060

44 NHS England Analytical Team. CCG IAF Methodology Manual. 2017.https://www.england.nhs.uk/wp-content/uploads/2017/07/Methodology-Manual-CCG-IAF.pdf

45 NHS England. STP Progress Dashboard – Methodology. 2017.https://www.england.nhs.uk/wp-content/uploads/2017/07/stp-progress-dashboard-methods-2017.pdf

46 Centers for Medicare & Medicaid Services. Guide to Choosing a Hospital. 2017. https://www.medicare.gov/Pubs/pdf/10181-Guide-Choosing-Hospital.pdf (accessed 26 Oct 2020).

47 Bardsley M. Learning how to make routinely available data useful in guiding regulatory oversight of hospital care. BMJ Qual. Saf. 2017;**26**:90–2. doi:10.1136/bmjqs-2016-005311

48 NHS England. myNHS retirement. https://www.nhs.uk/mynhs/index.html (accessed 15 Dec 2020).

49 Jacobs R, Smith PC, Goddard M. Measuring performance: An examination of composite performance indicators. 2004;**29**.internal-pdf://101.90.51.24/tp29.pdf

50 Profit J, Typpo K V, Hysong SJ, *et al.* Improving benchmarking by using an explicit framework for the development of composite indicators: an example using pediatric quality of care. *Implement Sci* 2010;**5**:13. doi:10.1186/1748-5908-5-13

51 Freudenberg M. *Composite Indicators of Country Performance*. OECD Publishing 2003. file:///content/workingpaper/405566708255http://dx.doi.org/10.1787/405566708255

52 Busweiler LAD, Schouwenburg MG, van Berge Henegouwen MI, *et al.* Textbook outcome as a composite measure in oesophagogastric cancer surgery. *Br J Surg* 2017;**104**:742–50. doi:10.1002/bjs.10486

53 Kolfschoten NE, Kievit J, Gooiker GA, *et al.* Focusing on desired outcomes of care after colon cancer resections; hospital variations in ‘textbook outcome’. *Eur J Surg Oncol* 2013;**39**:156–63. doi:10.1016/j.ejso.2012.10.007

54 Follmann D, Fay MP, Hamasaki T, *et al.* Analysis of ordered composite endpoints. *Stat Med* Published Online First: 2019. doi:10.1002/sim.8431

55 Manja V, AlBashir S, Guyatt G. Criteria for use of composite end points for competing risks—a systematic survey of the literature with recommendations. *J Clin Epidemiol* 2017;**82**:4–11. doi:http://dx.doi.org/10.1016/j.jclinepi.2016.12.001

56 Adams JL, Mehrotra A, Thomas JW, *et al.* Physician Cost Profiling — Reliability and Risk of Misclassification. *N Engl J Med* 2010;**362**:1014–21. doi:10.1056/NEJMsa0906323

57 Dimick JB, Welch H, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: The problem with small sample size. *JAMA* 2004;**292**:847–51.

doi:10.1001/jama.292.7.847

58 Howell V, Schwartz AE, O'Leary JD, *et al.* The effect of the SQUIRE (Standards of QUality Improvement Reporting Excellence) guidelines on reporting standards in the quality improvement literature: a before-and-after study. *BMJ Qual Saf* 2015;**24**:400–6. doi:10.1136/bmjqs-2014-003737

59 Walker K, Neuburger J, Groene O, *et al.* Public reporting of surgeon outcomes: low numbers of procedures lead to false complacency. *Lancet* 2013;**382**:1674–7. doi:10.1016/s0140-6736(13)61491-9

60 Dimick JB, Staiger DO, Birkmeyer JD. Ranking Hospitals on Surgical Mortality: The Importance of Reliability Adjustment. *Health Serv Res* 2010;**45**:1614–29. doi:10.1111/j.1475-6773.2010.01158.x

61 Dimick JB, Staiger DO, Osborne NH, *et al.* Composite Measures for Rating Hospital Quality with Major Surgery. *Health Serv Res* 2012;**47**:1861–79. doi:10.1111/j.1475-6773.2012.01407.x

62 Rajaram R, Barnard C, Bilimoria KY. Concerns about using the patient safety indicator-90 composite in pay-for-performance programs. *JAMA* 2015;**313**:897–8. doi:10.1001/jama.2015.52

63 Dixon-Woods M, Cavers D, Agarwal S, *et al.* Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. BMC Med. Res. Methodol. 2006;**6**:35. doi:10.1186/1471-2288-6-35

64 Dixon-Woods M. Systematic Reviews and Qualitative Studies. In: Silverman D, ed. *Qualitative Research*. SAGE Publications 2016. 379.

65 Bilimoria KY, Birkmeyer JD, Burstin H, *et al.* Rating the Raters: An Evaluation of Publicly Reported Hospital Quality Rating Systems. *NEJM Catal* Published Online First: 2019.https://catalyst.nejm.org/evaluation-hospital-quality-rating-systems/

66 The Leapfrog Group. Explanation of Hospital Safety Grades SPRING 2020. 2016. www.HospitalSafetyGrade.org. (accessed 25 Aug 2020).

67 Bevan G, Hood C. What's measured is what matters: Targets and gaming in the English public health care system. *Public Adm* 2006;**84**:517–38. doi:10.1111/j.1467-9299.2006.00600.x

68 Monitor, NHS Trust Development Authority. Learning from mistakes league. 2016.https://www.gov.uk/government/publications/learning-from-mistakes-league

69 Sentinel Stroke National Audit Programme. SSNAP Summary Report for December 2016 - March 2017 admissions and discharges. Royal College of Physicians 2017. internal-pdf://0361761859/DecMar2017-SummaryReport.xls (accessed 2 Jul 2020).

70 MyNHS: Data for better services. Performance of stroke services in England. https://www.nhs.uk/service-search/performance-indicators/organisations/hospital-specialties-stroke

71 Medicare.gov. Hospital Compare overall rating: Measures included in measure categories (December 2017). https://www.medicare.gov/hospitalcompare/Data/Measure-groups.html

72 Rowan K, Harrison D, Brady A, *et al.* Hospitals' star ratings and clinical outcomes:

ecological study. *BMJ* 2004;**328**:924–5. doi:10.1136/bmj.38007.694745.F7

73 Bevan G, Hood C. Have targets improved performance in the English NHS? *BMJ* 2006;**332**:419–22. doi:10.1136/bmj.332.7538.419

74 DeLancey JO, Softcheck J, Chung JW, *et al.* Associations between hospital characteristics, measure reporting, and the centers for medicare & medicaid services overall hospital quality star ratings. *JAMA* 2017;**317**:2015–7. doi:10.1001/jama.2017.3148

75 Analytics TH, Health IBMW. 100 Top Hospitals Study, 2017. 2017.http://100tophospitals.com/Portals/2/assets/TOP-17558-0217-100TopMethodology.pdf

76 Bilimoria KY, Chung J, Ju MH, *et al.* Evaluation of Surveillance Bias and the Validity of the Venous Thromboembolism Quality Measure. *JAMA* 2013;**310**:1482. doi:10.1001/jama.2013.280048

77 Barclay ME, Lyratzopoulos G, Greenberg D, *et al.* Missing data and chance variation in public reporting of cancer stage at diagnosis: Cross-sectional analysis of population-based data in England. *Cancer Epidemiol* 2018;**52**:28–42. doi:10.1016/j.canep.2017.11.005

78 Collins GS, Ogundimu EO, Cook JA, *et al.* Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Stat Med* 2016;:n/a-n/a. doi:10.1002/sim.6986

79 Schang L, Hynninen Y, Morton A, *et al.* Developing robust composite measures of healthcare quality – Ranking intervals and dominance relations for Scottish Health Boards. *Soc Sci Med* 2016;**162**:59–67. doi:http://dx.doi.org/10.1016/j.socscimed.2016.06.026

80 Gutacker N, Street A. Multidimensional performance assessment of public sector organisations using dominance criteria. *Health Econ* 2018;**27**:e13–27. doi:10.1002/hec.3554

81 Chen Q, Rosen AK, Borzecki A, *et al.* Using Harm-Based Weights for the AHRQ Patient Safety for Selected Indicators Composite (PSI-90): Does It Affect Assessment of Hospital Performance and Financial Penalties in Veterans Health Administration Hospitals? *Health Serv Res* 2016;**51**:2140–57. doi:10.1111/1475-6773.12596

82 Agency for Healthcare Research and Quality. PSI 90 Fact Sheet. 2016.https://www.qualityindicators.ahrq.gov/News/PSI90_Factsheet_FAQ_v1.pdf

83 Spiegelhalter D, Sherlaw-Johnson C, Bardsley M, *et al.* Statistical methods for healthcare regulation: rating, screening and surveillance. *J R Stat Soc Ser A (Statistics Soc* 2012;**175**:1–47. doi:10.1111/j.1467-985X.2011.01010.x

84 Bird SM, Cox SD, Farewell VT, *et al.* Performance indicators: good, bad, and ugly. *J R Stat Soc Ser A (Statistics Soc* 2005;**168**:1–27. doi:10.1111/j.1467-985X.2004.00333.x

85 Health and Social Care Information Centre. Criteria and considerations used to determine a quality indicator. 2015.http://content.digital.nhs.uk/media/14624/Criteria-and-considerations-used-to-determine-a-quality-indicator/pdf/Criteria_and_considerations_used_to_determine_a_quality_indicator_v3.

pdf

86 Finkelstein A, Gentzkow M, Hull P, *et al.* Adjusting Risk Adjustment — Accounting for Variation in Diagnostic Intensity. *N Engl J Med* 2017;**376**:608–10. doi:doi:10.1056/NEJMp1613238

87 Song Y, Skinner J, Bynum J, *et al.* Regional Variations in Diagnostic Practices. *N Engl J Med* 2010;**363**:45–53.https://www.nejm.org/doi/full/10.1056/nejmsa0910881 (accessed 18 Aug 2020).

88 Meacock R, Anselmi L, Kristensen SR, *et al.* Higher mortality rates amongst emergency patients admitted to hospital at weekends reflect a lower probability of admission. *J Health Serv Res Policy* Published Online First: 2016. doi:10.1177/1355819616649630

89 Ryan AM, Doran T. The effect of improving processes of care on patient outcomes: Evidence from the United Kingdom's quality and outcomes framework. *Med Care* 2012;**50**:191–9. doi:10.1097/MLR.0b013e318244e6b5

90 Llanwarne NR, Abel GA, Elliott MN, *et al.* Relationship between clinical quality and patient experience: Analysis of data from the English quality and outcomes framework and the national GP patient survey. *Ann Fam Med* 2013;**11**:467–72. doi:10.1370/afm.1514

91 Ashworth M, Seed P, Armstrong D, *et al.* The relationship between social deprivation and the quality of primary care: A national survey using indicators from the UK Quality and Outcomes Framework. *Br J Gen Pract* 2007;**57**:441–8./pmc/articles/PMC2078188/?report=abstract (accessed 5 Nov 2020).

92 Minchin M, Roland M, Richardson J, *et al.* Quality of Care in the United Kingdom after Removal of Financial Incentives. *N Engl J Med* 2018;**379**:948–57. doi:10.1056/nejmsa1801495

93 Ryan AM, Krinsky S, Maurer KA, *et al.* Changes in Hospital Quality Associated with Hospital Value-Based Purchasing. *N Engl J Med* 2017;**376**:2358–66. doi:10.1056/nejmsa1613412

94 Samuel CA, Zaslavsky AM, Landrum MB, *et al.* Developing and Evaluating Composite Measures of Cancer Care Quality. *Med Care* 2015;**53**:54–64. doi:10.1097/mlr.0000000000000257

95 Proudlove NC, Goff M, Walshe K, *et al.* The signal in the noise: Robust detection of performance "outliers" in health services. *J Oper Res Soc* 2018;:1–13. doi:10.1080/01605682.2018.1487816

96 Cefalu MS, Elliott MN, Setodji CM, *et al.* Hospital quality indicators are not unidimensional: A reanalysis of Lieberthal and Comer. *Health Serv Res* 2019;**0**. doi:doi:10.1111/1475-6773.13056

97 Shekelle PG. Quality indicators and performance measures: methods for development need more standardization. *J Clin Epidemiol* 2013;**66**:1338–9. doi:10.1016/j.jclinepi.2013.06.012

98 medicare.gov. How are hospital overall ratings calculated? https://www.medicare.gov/hospitalcompare/Data/Hospital-overall-ratings-calculation.html (accessed 26 Oct 2020).

99 IBM Watson Health. Methodology: Watson Health 100 Top Hospitals Study, 2020. Cambridge, MA: 2020. https://www.ibm.com/downloads/cas/MZRA5RBV (accessed 16 Oct 2020).

100 Sutter Health. Eight Sutter Hospital Campuses Earn Five-Star Rating from CMS. 2020.https://www.sutterhealth.org/newsroom/eight-sutter-hospital-campuses-earn-five-star-rating-from-cms (accessed 26 Oct 2020).

101 Rose Medical Center. Rose Medical Center Maintains Five-star Rating from CMS. 2020.https://rosemed.com/about/newsroom/rose-medical-center-maintains-five-star-rating-from-cms (accessed 26 Oct 2020).

102 Adventist Health Glendale. Adventist Health Glendale Earns Five Stars from CMS. 2020.https://www.adventisthealth.org/blog/2020/january/adventist-health-glendale-earns-five-stars-from-/ (accessed 26 Oct 2020).

103 Koh CY, Inaba CS, Sujatha-Bhaskar S, *et al.* Association of centers for medicare & medicaid services overall hospital quality star rating with outcomes in advanced laparoscopic abdominal surgery. *JAMA Surg* Published Online First: 2017. doi:10.1001/jamasurg.2017.2212

104 Papageorge M V, Resio BJ, Monsalve AF, *et al.* Navigating by Stars: Using CMS Star Ratings to Choose Hospitals for Complex Cancer Surgery. *JNCI Cancer Spectr* 2020;**4**. doi:10.1093/jncics/pkaa059

105 Wan W, Liang CJ, Duszak R, *et al.* Impact of Teaching Intensity and Sociodemographic Characteristics on CMS Hospital Compare Quality Ratings. *J Gen Intern Med* 2018;**33**:1221–3. doi:10.1007/s11606-018-4442-6

106 Shi B, King CJ, Huang SS. Relationship of Hospital Star Ratings to Race, Education, and Community Income. *J Hosp Med* 2020;**15**:588–93. doi:10.12788/jhm.3393

107 Hota B, Webb T, Chatrathi A, *et al.* Disagreement Between Hospital Rating Systems: Measuring the Correlation of Multiple Benchmarks and Developing a Quality Composite Rank. *Am J Med Qual* 2020;**35**:222–30. doi:10.1177/1062860619860250

108 Halasyamani LK, Davis MM. Conflicting measures of hospital quality: Ratings from ‘Hospital Compare’ versus ‘Best Hospitals’. *J Hosp Med* 2007;**2**:128–34. doi:10.1002/jhm.176

109 Smith SN, Reichert HA, Ameling JM, *et al.* Dissecting Leapfrog: How Well Do Leapfrog Safe Practices Scores Correlate With Hospital Compare Ratings and Penalties, and How Much Do They Matter? *Med Care* 2017;**55**:606–14. doi:10.1097/mlr.0000000000000716

110 Austin MM, Jha AK, Romano PS, *et al.* National hospital ratings systems share few common scores and may generate confusion instead of clarity. *Health Aff* 2015;**34**:423–30. doi:10.1377/hlthaff.2014.0201

111 Adelman D. The CMS Hospital Star Rating System: Fixing A Flawed Algorithm. Health Aff. 2020.https://www.healthaffairs.org/do/10.1377/hblog20200318.632395/full/ (accessed 27 Oct 2020).

112 Castellucci M. How the CMS picks winners of valuable quality measure contracts. Mod. Healthc. 2019.https://www.modernhealthcare.com/article/20190119/NEWS/190119903

113 Castellucci M. CMS, Yale New Haven Health on hot seat over design of quality measures. Mod. Healthc. 2019.https://www.modernhealthcare.com/article/20190119/NEWS/190119904

114 Zimlich R. Three top criticisms against CMS' overall hospital star ratings. Manag. Healthc. Exec. 2017.https://www.managedhealthcareexecutive.com/view/three-top-criticisms-against-cms-overall-hospital-star-ratings (accessed 19 Aug 2020).

115 Centers for Medicare and Medicaid Services. Hospital Compare Overall ratings SAS Package - January 2020. 2020.https://www.qualitynet.org/inpatient/public-reporting/overall-ratings/sas (accessed 19 Jun 2020).

116 Dixon WJ. Simplified Estimation from Censored Normal Samples. *Ann Math Stat* 1960;**31**:385–91. doi:10.1214/AOMS/1177705900

117 Shwartz M, Restuccia JD, Rosen AK. Composite Measures of Health Care Provider Performance: A Description of Approaches. *Milbank Q* 2015;**93**:788–825. doi:10.1111/1468-0009.12165

118 Centers for Medicare and Medicaid Services. Hospital Value-Based Purchasing Fact Sheet. 2017. https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/Hospital_VBPurchasing_Fact_Sheet_ICN907664.pdf (accessed 19 Jun 2020).

119 Lawley D. The Estimation of Factor Loadings by the Method of Maximum Likelihood. *Proceeding R Soc Edinburgh* 1940;**60**:64–82. doi:10.1017/S037016460002006X

120 Graham JW. Missing data analysis: Making it work in the real world. Annu. Rev. Psychol. 2009;**60**:549–76. doi:10.1146/annurev.psych.58.110405.085530

121 Sterne JAC, White IR, Carlin JB, *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;**338**:b2393. doi:10.1136/bmj.b2393

122 UCLA Institute for Digital Research & Education. How can I do factor analysis with missing data in Stata? | Stata FAQ. https://stats.idre.ucla.edu/stata/faq/how-can-i-do-factor-analysis-with-missing-data-in-stata/ (accessed 15 Nov 2020).

123 Dempster AP, Laird NM, Rubin DB. Maximum Lifelihood from Incomplete Data via the EM Algorithm. *J R Stat Soc Ser B* 1977;**39**:1–38.www.jstor.org/stable/2984875

124 Cattell RB. The scree test for the number of factors. *Multivariate Behav Res* 1966;**1**:245–76. doi:10.1207/s15327906mbr0102_10

125 Harris CW, Kaiser HF. Oblique factor analytic solutions by orthogonal transformations. *Psychometrika* 1964;**29**:347–62. doi:10.1007/BF02289601

126 Leonardi MJ, McGory ML, Ko CY. Publicly available hospital comparison web sites: Determination of useful, valid, and appropriate information for comparing surgical quality. *Arch Surg* 2007;**142**:863–8. doi:10.1001/archsurg.142.9.863

127 Saisana M, Saltelli A, Tarantola S. Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *J R Stat Soc Ser A (Statistics Soc* 2005;**168**:307–23. doi:10.1111/j.1467-985X.2005.00350.x

128 Ryan M, Scott DA, Reeves C, *et al.* Eliciting public preferences for healthcare: a

systematic review of techniques. *Heal Technol Assess* 2001;**5**:1–186.

129 Barclay ME. Impact of methodological choices on CMS Hospital Compare Star Ratings - final code and data. 2020.https://bitbucket.org/mattebarclay/cms-star-ratings-methodology-final-code-and-data/src (accessed 20 Jun 2020).

130 SSNAP. Sentinel Stroke National Audit Programme (SSNAP) UPCARE-tool. 2018.

131 Sentinel Stroke National Audit Programme. SSNAP clinical audit results portfolio, Jul-Sep 2019. 2019.https://www.strokeaudit.org/Documents/National/Clinical/JulSep2019/JulSep2019-FullResultsPortfolio.aspx (accessed 23 Aug 2020).

132 Jasmine Rapson. Wycombe Hospital’s stroke unit awarded top rating by the Royal College of Physicians. Bucks Free Press. 2017.https://www.bucksfreepress.co.uk/news/15356850.wycombe-hospitals-stroke-unit-awarded-top-rating-by-the-royal-college-of-physicians/ (accessed 27 Oct 2020).

133 Cambridge Network. Addenbrooke’s stroke team scores an A. 2020.https://www.cambridgenetwork.co.uk/news/addenbrooke’s-stroke-team-scores (accessed 27 Oct 2020).

134 Colm Bradley. SWAH stroke unit remains best in Northern Ireland. Impartial Report. 2019.https://www.impartialreporter.com/news/17773702.swah-stroke-unit-remains-best-northern-ireland/ (accessed 27 Oct 2020).

135 Sentinel Stroke National Audit Programme. SSNAP - National Results - Clinical. https://www.strokeaudit.org/results/Clinical-audit/National-Results.aspx (accessed 18 Aug 2020).

136 Sentinel Stroke National Audit Programme. SSNAP - SSNAP Reporting. https://www.strokeaudit.org/About-SSNAP/SSNAP-Clinical-Audit/SSNAP-Reporting.aspx (accessed 24 Aug 2020).

137 Intercollegiate Stroke Working Party. National clinical guideline for stroke. 2016. https://www.strokeaudit.org/SupportFiles/Documents/Guidelines/2016-National-Clinical-Guideline-for-Stroke-5t-(1).aspx (accessed 24 Aug 2020).

138 Sentinel Stroke National Audit Programme. SSNAP - Dependencies and evidence based standards. https://www.strokeaudit.org/About-SSNAP/Dependencies-and-evidence-based-standards.aspx (accessed 24 Aug 2020).

139 National Institute for Health and Care Excellence. Overview | Stroke in adults | Quality standards. 2016. https://www.nice.org.uk/guidance/qs2 (accessed 24 Aug 2020).

140 Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* 1960;**20**:37–46. doi:10.1177/001316446002000104

141 Barclay M. SSNAP technical methodology examination. 2020.https://bitbucket.org/mattebarclay/ssnap-methodology-final-code-and-data/src/master/ (accessed 26 Nov 2020).

142 Spearman C. ‘General Intelligence’ Objectively Determined and Measured. *Am J Psychol* 1904;**15**:201–93. doi:10.1037/11491-006

143 Jöreskog KG. A general method for analysis of covariance structures. *Biometrika* 1970;**57**:239–51. doi:10.1093/biomet/57.2.239

144 Jöreskog KG. A GENERAL APPROACH TO CONFIRMATORY MAXIMUM LIKELIHOOD FACTOR ANALYSIS. *ETS Res Bull Ser* 1967;**1967**:183–202. doi:10.1002/j.2333-8504.1967.tb00991.x

145 Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983;**67**:361–70.

146 Mehta RH, Liang L, Karve AM, *et al.* Association of patient case-mix adjustment, hospital process performance rankings, and eligibility for financial incentives. *JAMA - J Am Med Assoc* 2008;**300**:1897–903. doi:10.1001/jama.300.16.1897

147 Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl Stat* 1979;**28**:100. doi:10.2307/2346830

148 Moher D, Schulz KF, Simera I, *et al.* Guidance for Developers of Health Research Reporting Guidelines. *PLOS Med* 2010;**7**:e1000217. doi:10.1371/journal.pmed.1000217

149 Nikolakopoulou A, Trelle S, Sutton AJ, *et al.* Synthesizing existing evidence to design future trials: Survey of methodologists from European institutions. *Trials* 2019;**20**:334. doi:10.1186/s13063-019-3449-6

150 Kadam R, Borde S, Madas S, *et al.* Opinions and perceptions regarding the impact of new regulatory guidelines: A survey in Indian Clinical Trial Investigators. *Perspect Clin Res* 2016;**7**:81. doi:10.4103/2229-3485.179437

151 Colombo C, Roberto A, Krleza-Jeric K, *et al.* Sharing individual participant data from clinical studies: A cross-sectional online survey among Italian patient and citizen groups. *BMJ Open* 2019;**9**:24863. doi:10.1136/bmjopen-2018-024863

152 Dalkey NC. The Delphi method: an experimental study of group opinion. Santa Monica, CA: 1969. http://www.rand.org/pubs/research_memoranda/RM5888.html

153 Sinha IP, Smyth RL, Williamson PR. Using the Delphi technique to determine which outcomes to measure in clinical trials: recommendations for the future based on a systematic review of existing studies. *PLoS Med* 2011;**8**:e1000393. doi:10.1371/journal.pmed.1000393

154 Bennett WL, Robbins CW, Bayliss EA, *et al.* Engaging Stakeholders to Inform Clinical Practice Guidelines That Address Multiple Chronic Conditions. *J Gen Intern Med* 2017;**32**:883–90. doi:10.1007/s11606-017-4039-5

155 Flight L, Julious S, Brennan A, *et al.* How can health economics be used in the design and analysis of adaptive clinical trials? A qualitative analysis. *Trials* 2020;**21**:252. doi:10.1186/s13063-020-4137-2

156 Dimairo M, Boote J, Julious SA, *et al.* Missing steps in a staircase: A qualitative study of the perspectives of key stakeholders on the use of adaptive designs in confirmatory trials. *Trials* 2015;**16**:430. doi:10.1186/s13063-015-0958-9

157 Gargon E, Williamson PR, Young B. Improving core outcome set development: qualitative interviews with developers provided pointers to inform guidance. *J Clin Epidemiol* 2017;**86**:140–52. doi:10.1016/j.jclinepi.2017.04.024

158 Palinkas LA, Horwitz SM, Green CA, *et al.* Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Adm Policy Ment*

Heal 2015;**42**:533–44. doi:10.1007/s10488-013-0528-y

159 Malterud K, Siersma VD, Guassora AD. Sample Size in Qualitative Interview Studies. *Qual Health Res* 2016;**26**:1753–60. doi:10.1177/1049732315617444

160 Gale NK, Heath G, Cameron E, *et al.* Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Med Res Methodol* 2013;**13**:117. doi:10.1186/1471-2288-13-117

161 Heckathorn DD. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Soc Probl* 1997;**44**:174–99. doi:10.2307/3096941

162 Goodman LA. Snowball Sampling. *Ann Math Stat* 1961;**32**:148–70. doi:10.1214/aoms/1177705148

163 Fram SM. The Constant Comparative Analysis Method Outside of Grounded Theory. 2013. http://www.nova.edu/ssss/QR/QR18/fram1.pdf (accessed 16 Sep 2020).

164 StataCorp. Stata Statistical Software: Release 15. 2013.

165 Reeves D, Campbell SM, Adams J, *et al.* Combining Multiple Indicators of Clinical Quality. *Med Care* 2007;**45**:489–96. doi:10.1097/MLR.0b013e31803bb479

166 Shwartz M, Ren J, Pekoz EA, *et al.* Estimating a composite measure of hospital quality from the Hospital Compare database: differences when using a Bayesian hierarchical latent variable model versus denominator-based weights. *Med Care* 2008;**46**:778–85. doi:10.1097/MLR.0b013e31817893dc

167 Lingsma HF, Bottle A, Middleton S, *et al.* Evaluation of hospital outcomes: the relation between length-of-stay, readmission, and mortality in a large international administrative database. *BMC Health Serv Res* 2018;**18**:116. doi:10.1186/s12913-018-2916-1

168 Abel GA, Saunders CL, Lyratzopoulos G. Cancer patient experience, hospital performance and case mix: evidence from England. *Futur Oncol* 2013;**10**:1589–98. doi:10.2217/fon.13.266

169 Sebastian RM, Kumar D, Alappat BJ. Uncertainty and sensitivity analyses of incinerability index. *Environ Prog Sustain Energy* 2019;**38**. doi:10.1002/ep.13250

170 Şalap-Ayça S, Jankowski P. Analysis of the influence of parameter and scale uncertainties on a local multi-criteria land use evaluation model. *Stoch Environ Res Risk Assess* 2018;**32**:2699–719. doi:10.1007/s00477-018-1535-z

171 Anderson CC, Hagenlocher M, Renaud FG, *et al.* Comparing index-based vulnerability assessments in the Mississippi Delta: Implications of contrasting theories, indicators, and aggregation methodologies. *Int J Disaster Risk Reduct* 2019;**39**:101128. doi:10.1016/j.ijdrr.2019.101128

172 Adelman D. An Efficient Frontier Approach to Scoring and Ranking Hospital Performance. *Oper Res* 2020;**68**:762–92. doi:10.1287/opre.2019.1972

173 Adelman D. Efficient Frontier Hospital Ratings. Chicago Booth Rev. 2020.https://review.chicagobooth.edu/content/efficient-frontier-hospital-ratings (accessed 28 Oct 2020).

174 Rumball-Smith J, Gurvey J, Friedberg MW. Personalized Hospital Ratings —

Transparency for the Internet Age. *N Engl J Med* 2018;**379**:806–7. doi:10.1056/NEJMp1805000

175 Profit J, Kowalkowski MA, Zupancic JAF, *et al.* Baby-MONITOR: A Composite Indicator of NICU Quality. *Pediatrics* 2014;**134**:74–82. doi:10.1542/peds.2013-3552

176 Fisher RJ, Byrne A, Chouliara N, *et al.* Effectiveness of Stroke Early Supported Discharge: Analysis from a National Stroke Registry. *Circ Cardiovasc Qual Outcomes* 2020;**13**:571–9. doi:10.1161/CIRCOUTCOMES.119.006395

177 Lilford R, Mohammed MA, Spiegelhalter D, *et al.* Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004;**363**:1147–54. doi:https://doi.org/10.1016/S0140-6736(04)15901-1

178 Care Quality Commission. Intelligent Monitoring NHS acute hospitals: Statistical methodology. 2015.https://www.cqc.org.uk/sites/default/files/20150615_acute_im_v5_statistical_methodology.pdf

179 Bae JA, Curtis LH, Hernandez AF. National Hospital Quality Rankings: Improving the Value of Information in Hospital Rating Systems. *JAMA* Published Online First: 2020. doi:10.1001/jama.2020.11165

180 Elwood M. Forward projection--using critical appraisal in the design of studies. *Int J Epidemiol* 2002;**31**:1071–3.

181 Ivers NM, Taljaard M, Dixon S, *et al.* Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *BMJ* 2011;**343**. doi:10.1136/bmj.d5886

182 Bastuji-Garin S, Sbidian E, Gaudy-Marqueste C, *et al.* Impact of STROBE Statement Publication on Quality of Observational Study Reporting: Interrupted Time Series versus Before-After Analysis. *PLoS One* 2013;**8**:e64733. doi:10.1371/journal.pone.0064733

183 Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Bmj* 2010;**340**:c332. doi:10.1136/bmj.c332

184 Hoffmann TC, Glasziou PP, Boutron I, *et al.* Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 2014;**348**:g1687–g1687. doi:10.1136/bmj.g1687

185 Fitch K, Bernstein María SJ, Aguilar D, *et al.* The RAND/UCLA Appropriateness Method User’s Manual. 2001. http://www.rand.org (accessed 11 Nov 2020).

186 Noble M, Wright G, Smith G, *et al.* Measuring multiple deprivation at the small-area level. *Environ Plan A* 2006;**38**:169–85. doi:10.1068/a37168

187 Competence Centre on Composite Indicators and Scoreboards. 10 Step Guide | COIN. https://composite-indicators.jrc.ec.europa.eu/?q=10-step-guide (accessed 9 Dec 2020).

188 Profit J, Gould JB, Zupancic JAF, *et al.* Formal selection of measures for a composite index of NICU quality of care: Baby-MONITOR. *J Perinatol* 2011;**31**:702–10. doi:10.1038/jp.2011.12

189 Austin PC, Lee DS, Leckie G. Comparing a multivariate response Bayesian random effects logistic regression model with a latent variable item response theory model for provider profiling on multiple binary indicators simultaneously. *Stat Med* 2020;**39**:1390–406. doi:10.1002/sim.8484

190 Longford NT. Decision theory for comparing institutions. *Stat Med* 2018;**37**:457–72. doi:10.1002/sim.7525

191 Landrum MB, Bronskill SE, Normand S-LT. Analytic Methods for Constructing Cross-Sectional Profiles of Health Care Providers. *Heal Serv Outcomes Res Methodol* 2000;**1**:23–47. doi:10.1023/a:1010093701870

192 Knaapen L, Lehoux P. Science as Culture Three Conceptual Models of Patient and Public Involvement in Standard-setting: From Abstract Principles to Complex Practice Loes Knaapen & Pascale Lehoux. Published Online First: 2015. doi:10.1080/09505431.2015.1125875

193 Ives J, Damery S, Redwod S. PPI, paradoxes and Plato: who’s sailing the ship? *J Med Ethics* 2013;**39**:186–7. doi:10.1136/medethics-2012-100512

194 Williams O, Robert G, Martin GP, *et al.* Is Co-production Just Really Good PPI? Making Sense of Patient and Public Involvement and Co-production Networks. In: Bevir M, Waring J, eds. *Decentring Heath and Care Networks, Organizational Beha*. Springer International Publishing 2020. 213–37. doi:10.1007/978-3-030-40889-3_10

195 Williams O, Sarre S, Papoulias SC, *et al.* Lost in the shadows: reflections on the dark side of co-production. *Heal Res Policy Syst* 2020;**18**:43. doi:10.1186/s12961-020-00558-0

196 Shahian DM, Normand S-LT, Friedberg MW, *et al.* Rating the Raters: The Inconsistent Quality of Health Care Performance Measurement. *Ann Surg* 2016;**34**:36–8. doi:10.1097/sla.0000000000001631

197 Liu Y, Kale A, Althoff T, *et al.* Boba: Authoring and Visualizing Multiverse Analyses. Published Online First: 10 July 2020.http://arxiv.org/abs/2007.05551 (accessed 19 Jul 2020).

198 Steegen S, Tuerlinckx F, Gelman A, *et al.* Increasing Transparency Through a Multiverse Analysis. *Perspect Psychol Sci* 2016;**11**:702–12. doi:10.1177/1745691616658637

199 Sobol′ I. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math Comput Simul* 2001;**55**:271–80. doi:10.1016/S0378-4754(00)00270-6

200 Sobol IM. Sensitivity analysis for non-linear mathematical models. *Math Model Comput Exp (Engl Transl)* 1993;**1**:407–14.

201 Homma T, Saltelli A. Importance measures in global sensitivity analysis of nonlinear models. 1996.

202 Sobol’ IM. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput Math Math Phys* 1967;**7**:86–112. doi:10.1016/0041-5553(67)90144-9

203 Kuc-Czarnecka M, Lo Piano S, Saltelli A. Quantitative Storytelling in the Making of a

Composite Indicator. *Soc Indic Res* 2020;**149**:775–802. doi:10.1007/s11205-020-02276-0

204 Reeves S, Albert M, Kuper A, *et al.* Why use theories in qualitative research? *BMJ* 2008;**337**:a949–a949. doi:10.1136/bmj.a949

205 Greenland S, Senn SJ, Rothman KJ, *et al.* Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;**31**:337–50. doi:10.1007/s10654-016-0149-3

206 Senn S. Seven myths of randomisation in clinical trials. *Stat Med* 2013;**32**:1439–50. doi:10.1002/sim.5713

207 Richards H, Emslie C. The ‘doctor’ or the ‘girl from the University’? Considering the influence of professional roles on qualitative interviewing. *Fam Pract* 2000;**17**:71–5. doi:10.1093/fampra/17.1.71

208 Black N. To do the service no harm: the dangers of quality assessment. *J Health Serv Res Policy* 2015;**20**:65–6. doi:10.1177/1355819615570922

209 Teixeira-Pinto A, Normand S-LT. Statistical methodology for classifying units on the basis of multiple-related measures. *Stat Med* 2008;**27**:1329–50. doi:10.1002/sim.3187

210 Bevan G, Hamblin R. Hitting and missing targets by ambulance services for emergency calls: effects of different systems of performance measurement within the UK. *J R Stat Soc Ser A Stat Soc* 2009;**172**:161–90. doi:10.1111/j.1467-985X.2008.00557.x

211 Leckie G, Goldstein H. The evolution of school league tables in England 1992–2016: ‘Contextual value-added’, ‘expected progress’ and ‘progress 8’. *Br Educ Res J* 2017;**43**:193–212. doi:10.1002/berj.3264

212 Leckie G, Goldstein H. Understanding Uncertainty in School League Tables*. *Fisc Stud* 2011;**32**:207–24. doi:10.1111/j.1475-5890.2011.00133.x

213 Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *BMJ* 1998;**316**:1701–5. doi:10.1136/bmj.316.7146.1701

214 Austin PC. Bayes rules for optimally using Bayesian hierarchical regression models in provider profiling to identify high-mortality hospitals. *BMC Med Res Methodol* 2008;**8**:30. doi:10.1186/1471-2288-8-30

215 Stelfox HT, Straus SE. Measuring quality of care: considering measurement frameworks and needs assessment to guide quality indicator development. *J Clin Epidemiol* 2013;**66**:1320–7. doi:10.1016/j.jclinepi.2013.05.018

216 Stelfox HT, Straus SE. Measuring quality of care: considering conceptual approaches to quality indicator development and evaluation. *J Clin Epidemiol* 2013;**66**:1328–37. doi:10.1016/j.jclinepi.2013.05.017

217 Smith PC, Street A. Measuring the efficiency of public services: the limits of analysis. *J R Stat Soc Ser A (Statistics Soc* 2005;**168**:401–17. doi:10.1111/j.1467-985X.2005.00355.x

218 Centers for Medicare and Medicaid Services. Find Healthcare Providers: Compare Care Near You | Medicare. https://www.medicare.gov/care-

compare/results?searchType=Hospital&page=1&city=Cambridge&state=MA&zipcode=&radius=10&sort=closest (accessed 14 Sep 2020).

219 Centers for Medicare and Medicaid Services. Hospital Compare Public Reporting Overview. https://www.qualitynet.org/inpatient/public-reporting/public-reporting (accessed 14 Sep 2020).

220 National Quality Forum. Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement. 2016.http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=83123

221 Institute of Medicine. *Performance Measurement Accelerating Improvement*. Washington, DC: : The National Academies Press 2006. doi:10.17226/11517

222 Landrum MB, Normand S-LT, Rosenheck RA. Selection of Related Multivariate Means. *J Am Stat Assoc* 2003;**98**:7–16. doi:10.1198/016214503388619049

223 Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: : John Wiley and Sons 1987. doi:10.1002/9780470316696 LB - Rubin1987

224 Parker T, Knox C. New Zealand’s best place to retire. 2018.http://insights.nzherald.co.nz/article/best-retirement-area/

225 Johnson RB, Onwuegbuzie AJ. Toward a Definition of Mixed Methods Research. *J Mix Methods Res* 2007;**1**:112–33. doi:10.1177/1558689806298224

226 Ocloo J, Matthews R. From tokenism to empowerment: progressing patient and public involvement in healthcare improvement. *BMJ Qual Saf* 2016;**25**:626. doi:10.1136/BMJQS-2015-004839

227 Al-Janabi H, Coles J, Copping J, *et al.* Patient and Public Involvement (PPI) in Health Economics Methodology Research: Reflections and Recommendations. *Patient - Patient-Centered Outcomes Res 2020 144* 2020;**14**:421–7. doi:10.1007/S40271-020-00445-4

228 Metcalfe D, Rios Diaz AJ, Olufajo OA, *et al.* Impact of public release of performance data on the behaviour of healthcare consumers and providers. Cochrane Database Syst. Rev. 2018;**2018**. doi:10.1002/14651858.CD004538.pub3

229 Public Health England. Cancer Services Profiles. 2015.http://fingertips.phe.org.uk/profile/cancerservices

230 National Congenital Heart Disease Audit. Child Heart Surgery: All hospitals. 2020.https://childrensheartsurgery.info/data/table (accessed 18 Dec 2020).

231 Hsu C-C, Sandford BA. The Delphi Technique: Making Sense of Consensus. *Prat Assessment, Res Eval* 2007;**12**.https://pareonline.net/pdf/v12n10.pdf

232 MacLennan S, Kirkham J, Lam TBL, *et al.* A randomized trial comparing three Delphi feedback strategies found no evidence of a difference in a setting with high initial agreement. *J Clin Epidemiol* 2018;**93**:1–8. doi:10.1016/j.jclinepi.2017.09.024

233 Humphrey-Murto S, de Wit M. The Delphi Method - More Research Please. *J Clin Epidemiol* Published Online First: 23 October 2018. doi:10.1016/j.jclinepi.2018.10.011

234 Mardani A, Jusoh A, MD Nor K, *et al.* Multiple criteria decision-making techniques and

their applications – a review of the literature from 2000 to 2014. *Econ Res Istraživanja* 2015;**28**:516–71. doi:10.1080/1331677X.2015.1075139

235 Saaty RW. The analytic hierarchy process—what it is and how it is used. *Math Model* 1987;**9**:161–76. doi:10.1016/0270-0255(87)90473-8

Appendices

Appendix 1. Current approaches to individual measure standardisation for each of the measures used in the SSNAP score and level

Domain 1. Scanning

Scanning measure 1: % patients scanned within 1 hour.

Raw measure ranges from 0% to 100%.

$$\text{Standardised measure } = \begin{cases} 100 & \text{if \% patients scanned within 1 hour} > 95\% \\ 90 & \text{if \% patients scanned within 1 hour} > 90\% \text{ and } \leq 95\% \\ 80 & \text{if \% patients scanned within 1 hour} > 85\% \text{ and } \leq 90\% \\ 70 & \text{if \% patients scanned within 1 hour} > 80\% \text{ and } \leq 85\% \\ 60 & \text{if \% patients scanned within 1 hour} > 75\% \text{ and } \leq 80\% \\ 50 & \text{if \% patients scanned within 1 hour} > 70\% \text{ and } \leq 75\% \\ 40 & \text{if \% patients scanned within 1 hour} > 65\% \text{ and } \leq 70\% \\ 30 & \text{if \% patients scanned within 1 hour} > 60\% \text{ and } \leq 65\% \\ 20 & \text{if \% patients scanned within 1 hour} > 55\% \text{ and } \leq 60\% \\ 10 & \text{if \% patients scanned within 1 hour} > 50\% \text{ and } \leq 55\% \\ 0 & \text{if \% patients scanned within 1 hour} \leq 50\% \end{cases}$$

Scanning measure 2: % patients scanned within 12 hours.

Raw measure ranges from 0% to 100%.

$$\text{Standardised measure } = \begin{cases} 100 & \text{if \% patients scanned within 12 hours} > 95\% \\ 90 & \text{if \% patients scanned within 12 hours} > 90\% \text{ and } \leq 95\% \\ 80 & \text{if \% patients scanned within 12 hours} > 85\% \text{ and } \leq 90\% \\ 70 & \text{if \% patients scanned within 12 hours} > 80\% \text{ and } \leq 85\% \\ 60 & \text{if \% patients scanned within 12 hours} > 75\% \text{ and } \leq 80\% \\ 50 & \text{if \% patients scanned within 12 hours} > 70\% \text{ and } \leq 75\% \\ 40 & \text{if \% patients scanned within 12 hours} > 65\% \text{ and } \leq 70\% \\ 30 & \text{if \% patients scanned within 12 hours} > 60\% \text{ and } \leq 65\% \\ 20 & \text{if \% patients scanned within 12 hours} > 55\% \text{ and } \leq 60\% \\ 10 & \text{if \% patients scanned within 12 hours} > 50\% \text{ and } \leq 55\% \\ 0 & \text{if \% patients scanned within 12 hours} \leq 50\% \end{cases}$$

Scanning measure 3: Median time until scanned

Raw measure is time-to-event, and theoretically can take any positive value.

$$\text{Standardised measure} = \begin{cases} 100 & \text{if median time until scanned} <45 \text{ minutes} \\ 90 & \text{if median time until scanned} <60 \text{ and } \geq 45 \text{ minutes} \\ 80 & \text{if median time until scanned} <75 \text{ and } \geq 60 \text{ minutes} \\ 70 & \text{if median time until scanned} <90 \text{ and } \geq 75 \text{ minutes} \\ 60 & \text{if median time until scanned} <120 \text{ and } \geq 90 \text{ minutes} \\ 50 & \text{if median time until scanned} <180 \text{ and } \geq 120 \text{ minutes} \\ 40 & \text{if median time until scanned} <240 \text{ and } \geq 180 \text{ minutes} \\ 30 & \text{if median time until scanned} <300 \text{ and } \geq 240 \text{ minutes} \\ 20 & \text{if median time until scanned} <360 \text{ and } \geq 300 \text{ minutes} \\ 10 & \text{if median time until scanned} <480 \text{ and } \geq 360 \text{ minutes} \\ 0 & \text{if median time until scanned} \geq 480 \text{ minutes} \end{cases}$$

Domain 2. Stroke unit

Stroke unit measure 1. % patients directly admitted within 4 hours

Raw measure ranges from 0% to 100%.

Standardised measure = % patients directly admitted within 4 hours (i.e. Raw measure used as-is).

Stroke unit measure 2. Median time until arrival on stroke unit

Raw measure is time-to-event, and theoretically can take any positive value.

$$\text{Standardised measure} = \begin{cases} 100 & \text{if median time until scanned} <60 \text{ minutes} \\ 90 & \text{if median time until scanned} <120 \text{ and } \geq 60 \text{ minutes} \\ 80 & \text{if median time until scanned} <180 \text{ and } \geq 120 \text{ minutes} \\ 70 & \text{if median time until scanned} <240 \text{ and } \geq 180 \text{ minutes} \\ 60 & \text{if median time until scanned} <270 \text{ and } \geq 240 \text{ minutes} \\ 50 & \text{if median time until scanned} <300 \text{ and } \geq 270 \text{ minutes} \\ 40 & \text{if median time until scanned} <330 \text{ and } \geq 300 \text{ minutes} \\ 30 & \text{if median time until scanned} <360 \text{ and } \geq 330 \text{ minutes} \\ 20 & \text{if median time until scanned} <420 \text{ and } \geq 360 \text{ minutes} \\ 10 & \text{if median time until scanned} <480 \text{ and } \geq 420 \text{ minutes} \\ 0 & \text{if median time until scanned} \geq 480 \text{ minutes} \end{cases}$$

Stroke unit measure 3. % patients spending at least 90% of stay on a stroke unit

Raw measure ranges from 0% to 100%.

Standardised measure = % patients spending at least 90% of stay on a stroke unit (i.e. Raw measure used as-is).

Domain 3. Thrombolysis

Thrombolysis measure 1. % all stroke patients given thrombolysis

Raw measure ranges from 0% to 100%.

$$\text{Standardised measure} = \begin{cases} 5 \times (\%\ \text{raw measure}) & \text{if raw measure} < 20\% \\ 100 & \text{otherwise} \end{cases}$$

Thrombolysis measure 2. % eligible patients given thrombolysis

Raw measure ranges from 0% to 100%.

Standardised measure = % eligible patients given thrombolysis (i.e. Raw measure used as-is).

Thrombolysis measure 3. % patients thrombolysed within 1 hour

Raw measure ranges from 0% to 100%.

Standardised measure = % patients thrombolysed within 1 hour (i.e. Raw measure used as-is).

Thrombolysis measure 4. % applicable patients admitted within 4 hrs AND receiving thrombolysis

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Thrombolysis measure 5. Median time until thrombolysis

Raw measure is time-to-event, and theoretically can take any positive value.

$$\text{Standardised measure} = \begin{cases} 100 & \text{if median time until thrombolysis} <30 \text{ minutes} \\ 90 & \text{if median time until thrombolysis} <40 \text{ and } \geq 30 \text{ minutes} \\ 80 & \text{if median time until thrombolysis} <50 \text{ and } \geq 40 \text{ minutes} \\ 70 & \text{if median time until thrombolysis} <60 \text{ and } \geq 50 \text{ minutes} \\ 60 & \text{if median time until thrombolysis} <70 \text{ and } \geq 60 \text{ minutes} \\ 50 & \text{if median time until thrombolysis} <80 \text{ and } \geq 70 \text{ minutes} \\ 40 & \text{if median time until thrombolysis} <90 \text{ and } \geq 80 \text{ minutes} \\ 30 & \text{if median time until thrombolysis} <100 \text{ and } \geq 90 \text{ minutes} \\ 20 & \text{if median time until thrombolysis} <110 \text{ and } \geq 100 \text{ minutes} \\ 10 & \text{if median time until thrombolysis} <120 \text{ and } \geq 110 \text{ minutes} \\ 0 & \text{if median time until thrombolysis} \geq 120 \text{ minutes} \end{cases}$$

Domain 4. Specialist assessment

Specialist assessment measure 1. % patients assessed by a stroke specialist within 24 hours

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Specialist assessment measure 2. Median time until assessed by a stroke specialist

Raw measure is time-to-event, and theoretically can take any positive value.

$$\text{Standardised measure} \\ = \begin{cases} 100 & \text{if median time until assessed by a stroke specialist} <3 \text{ hours} \\ 90 & \text{if median time until assessed by a stroke specialist} <6 \text{ and } \geq 3 \text{ hrs} \\ 80 & \text{if median time until assessed by a stroke specialist} <9 \text{ and } \geq 6 \text{ hrs} \\ 70 & \text{if median time until assessed by a stroke specialist} <12 \text{ and } \geq 9 \text{ hrs} \\ 60 & \text{if median time until assessed by a stroke specialist} <15 \text{ and } \geq 12 \text{ hrs} \\ 50 & \text{if median time until assessed by a stroke specialist} <18 \text{ and } \geq 15 \text{ hrs} \\ 40 & \text{if median time until assessed by a stroke specialist} <21 \text{ and } \geq 18 \text{ hrs} \\ 30 & \text{if median time until assessed by a stroke specialist} <24 \text{ and } \geq 21 \text{ hrs} \\ 20 & \text{if median time until assessed by a stroke specialist} <36 \text{ and } \geq 24 \text{ hrs} \\ 10 & \text{if median time until assessed by a stroke specialist} <48 \text{ and } \geq 36 \text{ hrs} \\ 0 & \text{if median time until assessed by a stroke specialist} \geq 48 \text{ hrs} \end{cases}$$

Specialist assessment measure 3. % patients assessed by a stroke nurse within 24 hours

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Specialist assessment measure 4. Median time until assessed by a stroke nurse

Raw measure is time-to-event, and theoretically can take any positive value.

Standardised measure

$$= \begin{cases} 100 & \text{if median time until assessed by a stroke specialist } <0.5 \text{ hours} \\ 90 & \text{if median time until assessed by a stroke specialist } <1 \text{ and } \geq 0.5 \text{ hrs} \\ 80 & \text{if median time until assessed by a stroke specialist } <2 \text{ and } \geq 1 \text{ hrs} \\ 70 & \text{if median time until assessed by a stroke specialist } <3 \text{ and } \geq 2 \text{ hrs} \\ 60 & \text{if median time until assessed by a stroke specialist } <6 \text{ and } \geq 3 \text{ hrs} \\ 50 & \text{if median time until assessed by a stroke specialist } <9 \text{ and } \geq 6 \text{ hrs} \\ 40 & \text{if median time until assessed by a stroke specialist } <12 \text{ and } \geq 9 \text{ hrs} \\ 30 & \text{if median time until assessed by a stroke specialist } <15 \text{ and } \geq 12 \text{ hrs} \\ 20 & \text{if median time until assessed by a stroke specialist } <18 \text{ and } \geq 15 \text{ hrs} \\ 10 & \text{if median time until assessed by a stroke specialist } <21 \text{ and } \geq 18 \text{ hrs} \\ 0 & \text{if median time until assessed by a stroke specialist } \geq 21 \text{ hrs} \end{cases}$$

Specialist assessment measure 5. % applicable patients given a swallow screen within 24 hours

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Specialist assessment measure 6. % applicable patients given a formal swallow assessment within 72 hours

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Domain 5. Occupational therapy

Occupational therapy measure 1. % patients reported as requiring occupational therapy

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Occupational therapy measure 2. Median minutes per day receiving occupational therapy

Raw measure is elapsed time, and theoretically can take any positive value up to 1440 minutes. There are 1440 minutes in one day.

$$\text{Standardised measure} \\ = \begin{cases} 100 & \text{if median minutes per day receiving occ. therapy} > 40 \text{ mins} \\ 90 & \text{if median minutes per day receiving occ. therapy} > 32 \text{ and } \leq 40 \text{ mins} \\ 80 & \text{if median minutes per day receiving occ. therapy} > 28 \text{ and } \leq 32 \text{ mins} \\ 70 & \text{if median minutes per day receiving occ. therapy} > 24 \text{ and } \leq 28 \text{ mins} \\ 60 & \text{if median minutes per day receiving occ. therapy} > 20 \text{ and } \leq 24 \text{ mins} \\ 50 & \text{if median minutes per day receiving occ. therapy} > 16 \text{ and } \leq 20 \text{ mins} \\ 40 & \text{if median minutes per day receiving occ. therapy} > 12 \text{ and } \leq 16 \text{ mins} \\ 30 & \text{if median minutes per day receiving occ. therapy} > 8 \text{ and } \leq 12 \text{ mins} \\ 20 & \text{if median minutes per day receiving occ. therapy} > 4 \text{ and } \leq 8 \text{ mins} \\ 10 & \text{if median minutes per day receiving occ. therapy} > 0 \text{ and } \leq 4 \text{ mins} \\ 0 & \text{if median minutes per day receiving occ. therapy} = 0 \text{ mins} \end{cases}$$

Occupational therapy measure 3. Median % days on which occupational therapy is received

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Occupational therapy measure 4. % compliance against therapy target for occupational therapy

Raw measure ranges from 0% up. Note that hospitals can overachieve on this measure, so scores can go above 100%.

Standardised measure = Raw measure.

Domain 6. Physiotherapy

Physiotherapy measure 1. % patients reported as requiring physiotherapy

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Physiotherapy measure 2. Median minutes per day receiving physiotherapy

Raw measure is elapsed time, and theoretically can take any positive value up to 1440 minutes. There are 1440 minutes in one day.

Standardised measure

$$= \begin{cases} 100 & \text{if median minutes per day receiving physiotherapy } >40 \text{ mins} \\ 90 & \text{if median minutes per day receiving physiotherapy } >32 \text{ and } \leq 40 \text{ mins} \\ 80 & \text{if median minutes per day receiving physiotherapy } >28 \text{ and } \leq 32 \text{ mins} \\ 70 & \text{if median minutes per day receiving physiotherapy } >24 \text{ and } \leq 28 \text{ mins} \\ 60 & \text{if median minutes per day receiving physiotherapy } >20 \text{ and } \leq 24 \text{ mins} \\ 50 & \text{if median minutes per day receiving physiotherapy } >16 \text{ and } \leq 20 \text{ mins} \\ 40 & \text{if median minutes per day receiving physiotherapy } >12 \text{ and } \leq 16 \text{ mins} \\ 30 & \text{if median minutes per day receiving physiotherapy } >8 \text{ and } \leq 12 \text{ mins} \\ 20 & \text{if median minutes per day receiving physiotherapy } >4 \text{ and } \leq 8 \text{ mins} \\ 10 & \text{if median minutes per day receiving physiotherapy } >0 \text{ and } \leq 4 \text{ mins} \\ 0 & \text{if median minutes per day receiving physiotherapy } =0 \text{ mins} \end{cases}$$

Physiotherapy measure 3. Median % days on which physiotherapy is received

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Physiotherapy measure 4. % compliance against therapy target for physiotherapy

Raw measure ranges from 0% up. Note that hospitals can overachieve on this measure, so scores can go above 100%.

Standardised measure = Raw measure.

Domain 7. Speech and language therapy

Speech and language therapy measure 1. % patients reported as requiring speech and language therapy

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Speech and language therapy measure 2. Median minutes per day receiving speech and language therapy

Raw measure is elapsed time, and theoretically can take any positive value up to 1440 minutes. There are 1440 minutes in one day.

$$\text{Standardised measure} = \begin{cases} 100 & \text{if median minutes per day receiving SLT} > 40 \text{ mins} \\ 90 & \text{if median minutes per day receiving SLT} > 32 \text{ and} \leq 40 \text{ mins} \\ 80 & \text{if median minutes per day receiving SLT} > 28 \text{ and} \leq 32 \text{ mins} \\ 70 & \text{if median minutes per day receiving SLT} > 24 \text{ and} \leq 28 \text{ mins} \\ 60 & \text{if median minutes per day receiving SLT} > 20 \text{ and} \leq 24 \text{ mins} \\ 50 & \text{if median minutes per day receiving SLT} > 16 \text{ and} \leq 20 \text{ mins} \\ 40 & \text{if median minutes per day receiving SLT} > 12 \text{ and} \leq 16 \text{ mins} \\ 30 & \text{if median minutes per day receiving SLT} > 8 \text{ and} \leq 12 \text{ mins} \\ 20 & \text{if median minutes per day receiving SLT} > 4 \text{ and} \leq 8 \text{ mins} \\ 10 & \text{if median minutes per day receiving SLT} > 0 \text{ and} \leq 4 \text{ mins} \\ 0 & \text{if median minutes per day receiving SLT} = 0 \text{ mins} \end{cases}$$

Where SLT means Speech and Language Therapy.

Speech and language therapy measure 3. Median % days on which speech and language therapy is received

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Speech and language therapy measure 4. % compliance against therapy target for speech and language therapy

Raw measure ranges from 0% up. Note that hospitals can overachieve on this measure, so scores can go above 100%.

Standardised measure = Raw measure.

Domain 8. Multi-disciplinary team (MDT) working

MDT working measure 1. % applicable patients assessed by occupational therapist within 72 hours

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

MDT working measure 2. Median time until assessed by occupational therapist

Raw measure is time-to-event, and theoretically can take any positive value.

$$
\begin{aligned}
&\text{Standardised measure}\\
&= \begin{cases}
100 & \text{if median time until assessed by occupational therapist} <6 \text{ hours}\\
90 & \text{if median time until assessed by occupational therapist} <12 \text{ and } \geq 6 \text{ hrs}\\
80 & \text{if median time until assessed by occupational therapist} <18 \text{ and } \geq 12 \text{ hrs}\\
70 & \text{if median time until assessed by occupational therapist} <24 \text{ and } \geq 18 \text{ hrs}\\
60 & \text{if median time until assessed by occupational therapist} <30 \text{ and } \geq 24 \text{ hrs}\\
50 & \text{if median time until assessed by occupational therapist} <36 \text{ and } \geq 30 \text{ hrs}\\
40 & \text{if median time until assessed by occupational therapist} <42 \text{ and } \geq 36 \text{ hrs}\\
30 & \text{if median time until assessed by occupational therapist} <48 \text{ and } \geq 42 \text{ hrs}\\
20 & \text{if median time until assessed by occupational therapist} <54 \text{ and } \geq 48 \text{ hrs}\\
10 & \text{if median time until assessed by occupational therapist} <60 \text{ and } \geq 54 \text{ hrs}\\
0 & \text{if median time until assessed by occupational therapist} \geq 60 \text{ hrs}
\end{cases}
\end{aligned}
$$

MDT working measure 3. % applicable patients assessed by a physiotherapist within 72 hours

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

MDT working measure 4. Median time until assessed by a physiotherapist

Raw measure is time-to-event, and theoretically can take any positive value.

$$
\begin{aligned}
&\text{Standardised measure}\\
&= \begin{cases}
100 & \text{if median time until assessed by physiotherapist} <6 \text{ hours}\\
90 & \text{if median time until assessed by physiotherapist} <12 \text{ and } \geq 6 \text{ hrs}\\
80 & \text{if median time until assessed by physiotherapist} <18 \text{ and } \geq 12 \text{ hrs}\\
70 & \text{if median time until assessed by physiotherapist} <24 \text{ and } \geq 18 \text{ hrs}\\
60 & \text{if median time until assessed by physiotherapist} <30 \text{ and } \geq 24 \text{ hrs}\\
50 & \text{if median time until assessed by physiotherapist} <36 \text{ and } \geq 30 \text{ hrs}\\
40 & \text{if median time until assessed by physiotherapist} <42 \text{ and } \geq 36 \text{ hrs}\\
30 & \text{if median time until assessed by physiotherapist} <48 \text{ and } \geq 42 \text{ hrs}\\
20 & \text{if median time until assessed by physiotherapist} <54 \text{ and } \geq 48 \text{ hrs}\\
10 & \text{if median time until assessed by physiotherapist} <60 \text{ and } \geq 54 \text{ hrs}\\
0 & \text{if median time until assessed by physiotherapist} \geq 60 \text{ hrs}
\end{cases}
\end{aligned}
$$

MDT working measure 5. % applicable patients assessed by a speech therapist within 72 hours

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

MDT working measure 6. Median time until assessed by a speech therapist

Raw measure is time-to-event, and theoretically can take any positive value.

Standardised measure

$$= \begin{cases} 100 & \text{if median time until assessed by speech therapist} < 6 \text{ hours} \\ 90 & \text{if median time until assessed by speech therapist} < 12 \text{ and} \geq 6 \text{ hrs} \\ 80 & \text{if median time until assessed by speech therapist} < 18 \text{ and} \geq 12 \text{ hrs} \\ 70 & \text{if median time until assessed by speech therapist} < 24 \text{ and} \geq 18 \text{ hrs} \\ 60 & \text{if median time until assessed by speech therapist} < 30 \text{ and} \geq 24 \text{ hrs} \\ 50 & \text{if median time until assessed by speech therapist} < 36 \text{ and} \geq 30 \text{ hrs} \\ 40 & \text{if median time until assessed by speech therapist} < 42 \text{ and} \geq 36 \text{ hrs} \\ 30 & \text{if median time until assessed by speech therapist} < 48 \text{ and} \geq 42 \text{ hrs} \\ 20 & \text{if median time until assessed by speech therapist} < 54 \text{ and} \geq 48 \text{ hrs} \\ 10 & \text{if median time until assessed by speech therapist} < 60 \text{ and} \geq 54 \text{ hrs} \\ 0 & \text{if median time until assessed by speech therapist} \geq 60 \text{ hrs} \end{cases}$$

MDT working measure 7. % applicable patients with rehab goals agreed within 5 days

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

MDT working measure 8. % applicable patients assessed by all relevant specialists in a timely manner

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Domain 9. Standards by discharge

Standards by discharge measure 1. % applicable patients screened for nutrition and seen by dietitian by discharge

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Standards by discharge measure 2. % applicable patients with a continence plan drawn up within 3 weeks

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Standards by discharge measure 3. % applicable patients who have mood and cognition screening by discharge

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Domain 10. Discharge processes

Discharge processes measure 1. % applicable patients receiving a joint health and social care plan on discharge

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Discharge processes measure 2. % patients treated by a stroke-skilled Early Supported Discharge team

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Discharge processes measure 3. % applicable patients in atrial fibrillation discharged on anticoagulants

Raw measure ranges from 0% to 100%.

$$\text{Standardised measure } = \begin{cases} \dfrac{5 \times (\%\text{ raw measure})}{2} & \text{if raw measure } < 40\% \\ 100 & \text{otherwise} \end{cases}$$

Discharge processes measure 4. % patients discharged alive who are given a named person to contact

Raw measure ranges from 0% to 100%.

Standardised measure = Raw measure.

Appendix 2. Prompt guide for the interview study

Brief refresher on the aim of the study and interview.

- Our experience of composite indicators of healthcare quality – things like “star ratings” and quality grades and so on based on summary measures combining several indicators – is that they are often not reported in a transparent way. It is hard to understand what has been done and why in the design, development and reporting of indicators.
- We are interested in identifying the key choices when developing composite indicators.
- We want to use this as a starting point in producing reporting guidelines, but also think it is worth understanding how different professions think about this type of measurement.
- I want to get your views on the decisions involved in developing composite indicators.

Is that OK?

Have you read the information sheet? Completed the consent form?

Are you happy for me to record this interview?

1. Please can you give me a **very brief overview of your expertise** and the reasons you have been working with or researching performance measures in healthcare?
2. It is probably helpful to have a specific example in mind. If I asked you to develop a composite indicator of the quality of hospital cardiovascular services, how would you start?
    - *Can you **expand** on that?*
    - *Tell me about **another important** decision.*

Prompts if required

a. Does it matter **who you involve** in the development of a composite indicator?
    - Who? How? Why does it (not) matter? Is some form of patient/public involvement required?

b. Does the **purpose** of the indicator affect the design?
    - Choice of measures? Weights (socio-economic impact?)?
    - Does it matter if the plan is to use the indicator to **compare** organisations?

c. How will you ensure the **acceptability** and **relevance** of the indicator?

d. What would you think about when **combining measures**?
    - Weights? Gaming? Standardisation?

e. Does published documentation need to be detailed enough that someone can use it to **reproduce** the numbers in the indicator? How would you ensure they can?
    - Code? Detailed methods? Underlying datasets?

f. Are there decisions to be made around the **statistical properties** of the indicator?

  - Missing data? Reporting periods? Case-mix? Standardisation?
  - How do we identify “good” organisational performance? What do we do if we can’t?

g. Are there choices related to the final **presentation** of the indicator?
  - Uncertainty? ‘Star ratings’ vs scores?

Round up

3. Are there any extra decisions or issues involved if we are trying to measure the overall quality of care provided by a hospital, as opposed to the quality of a specific service?

4. Are there **other issues** that have come to mind during our discussion?
   - *Skip if they volunteer this spontaneously*

Closing up

5. One final question. **Who else should I speak to** about the issues we have discussed with you today?

OK, I think we’ve discussed everything I had written down. Do you have anything you’d like to add?

Thank you for your time.

Appendix 3. Approaches to combining measures into composite indicators

Perhaps the most common approach to combining individual measures into a composite indicator in use in healthcare is the weighted arithmetic average. Weights are assigned to different domains of quality and then scores on these domains are combined using the weighted arithmetic mean. While this is a common approach, it is not the only approach that could be used. Alternatives to a weighted arithmetic average approach include: geometric means and other multiplicative approaches to combining domain scores [1,175]; multivariate statistical approaches [189–191]; and various multi-criteria decision analysis approaches [1].

Geometric means (the n^{th} root of the product of the n domain scores, or equivalently exponential of the arithmetic mean of the logarithm of the domain scores) are occasionally used so that extremely poor performance on a single domain is far more consequential than performance that is slightly below par on a single domain [1,175]. This type of approach, and related issues such as the extent to which good performance in one aspect of quality is allowed to average out poor performance in another area, was not raised in my interview study.

Multivariate statistical approaches allow composite summaries of performance that go beyond a single number of star rating, potentially avoiding the need to aggregate measures at all. For example, Austin, Lee and Leckie demonstrate the use of multivariate Bayesian models to profile hospitals based on how ‘extreme’ their performance is (based on Mahalonobis distance, effectively a multivariate extension of the Z-score), how likely they are to be underperforming, and how likely they are to be overperforming [189]. Multivariate approaches including factor analysis and principal components analysis were raised by interview participants, but not in the context of producing the final summary score.

Multi-criteria decision analysis includes a broad family of approaches developed primarily by operational researchers [234], and both simple averages and the various multivariate statistical approaches are special cases of multi-criteria decision analysis. More general forms of multi-criteria decision analysis, such as the analytic hierarchy process [235], aim to account for the inconsistency of human value judgments. Applications of general decision analysis to healthcare composite indicators appear to be limited to a handful of academic papers exploring the use of data envelopment analysis or efficient frontier analysis [80,172].

Inconsistency in multi-criteria decision analysis refers to the way that human value judgments may not be commutative. That is, a human may say: A is twice as important as B; B is twice as important as C; and A is three times as important as C. This value judgment is inconsistent, as if only the pairwise comparisons of A and B and of B and C are considered, then A would be expected to be judged as four times as important as C. The analytic hierarchy process is one example of a more general multi-criteria decision analysis technique [235]. In essence, the analytic hierarchy process calculates the best consistent set of preferences by calculating the principal eigenvector of the matrix of pairwise comparisons of the importance of the different measures.

Appendix 4. Academic papers published during PhD

Based on PhD work

Barclay M, Dixon-Woods M and Lyratzopoulos G.

The problem with composite indicators.

BMJ Quality and Safety 2018. doi:10.1136/bmjqs-2018-007798

Relevant to PhD (with regard to statistical reliability, case-mix, missing data and organisational variation)

Barclay M, Abel GA, Elliss-Brookes L, Greenberg D and Lyratzopoulos G.

The influence of patient case-mix on public health area statistics for cancer stage at diagnosis: a cross-sectional study.

Eur J Public Health 2019. doi:10.1093/eurpub/ckz024

Barclay M, Lyratzopoulos G, Greenberg D, Abel GA.

Missing data and chance variation in public reporting of cancer stage at diagnosis: Cross-sectional analysis of population-based data in England.

Cancer Epidemiol 2018;52:28-42. doi:10.1016/j.canep.2017.11.005

Other first author

Barclay M, Lyratzopoulos G, Walter FM, Jefferies S, Peake MD and Rintoul RC.

Risk of second and higher order smoking-related primary cancer following lung cancer: a population-based cohort study.

Thorax 2019. doi:10.1136/thoraxjnl-2018-212456

Barclay M, Abel GA, Greenberg DC, Rous B, Lyratzopoulos G.

The Socio-demographic variation in stage at diagnosis of breast, bladder, colon, endometrial, lung, melanoma, prostate, rectal, renal and ovarian cancer in England and its population impact.

Br J Cancer 2021. In press (accepted December 2020)

Other

Bradley SH and **Barclay M**.

“Liquid biopsy” for cancer screening: Careful evaluation must consider harms as well as potential benefits.

BMJ 2021. doi: https://doi.org/10.1136/bmj.m4933

Walter FM, Pannebakker MM, **Barclay M**, Mills K, Saunders CL, Murchie P, Corrie P, Hall P, Burrows N and Emery JD.

Effect of a Skin Self-monitoring Smartphone Application on Time to Physician Consultation Among Patients With High Risk of Melanoma: A Phase 2 Randomized Clinical Trial.

JAMA Network Open 2020;3(2):e200001. doi:10.1001/jamanetworkopen.2020.0001

Okuyama A, **Barclay M**, Chen C, Higashi T.

Impact of loss-to-follow-up on cancer survival estimates for small populations: a simulation study using Hospital-Based Cancer Registries in Japan.

BMJ Open 2020;10:e033510. doi:10.1136/bmjopen-2019-033510

Brodbelt AR, **Barclay M**, Greenberg D, Williams M, Jenkinson MD and Karabatsou K.

The outcome of patients with surgically treated meningioma in England: 1999–2013. A cancer registry data analysis.

Br J Neurosurg 2019: 1-7, doi:10.1080/02688697.2019.1661965

Hsu RCJ, **Barclay M**, Loughran MA, Lyratzopoulos G, Gnanapragasam VJ and Armitage JN.

Impact of hospital nephrectomy volume on intermediate- to long-term survival in renal cell carcinoma.

BJU Int 2019. doi:10.1111/bju.14848

Newbould J, Ball S, Abel G, **Barclay M**, Brown T, Corbett J, Doble B, Elliott M, Exley J, Knack A, Martin A, Pitchforth E, Saunders C, Wilson ECF, Winpenny E, Yang M and Roland M.

A ‘telephone first’ approach to demand management in English general practice: a multimethod evaluation.

Health Services and Delivery Research 2019:7(17). doi:10.3310/hsdr07170

Hsu RCJ, **Barclay M**, Loughran MA, Lyratzopoulos G, Gnanapragasam VJ and Armitage JN.

Time trends in service provision and survival outcomes for patients with renal cancer treated by nephrectomy in England 2000-2010.

BJU Int 2018. doi:10.1111/bju.14217

Herbert A, Koo MM, **Barclay M**, Greenberg DC, Abel GA, Levell NJ and Lyratzopoulos G.

Stage-specific incidence trends of melanoma in an English region, 1996-2015: longitudinal analyses of population-based data.

Melanoma Res 2018. doi:10.1097/cmr.0000000000000489

Petrova M, **Barclay M**, Barclay SS, Barclay SIG.

Between "the best way to deliver patient care" and "chaos and low clinical value": General Practitioners' and Practice Managers' views on data sharing.

Int J Med Inform. 2017;104:74-83. doi:10.1016/j.ijmedinf.2017.05.00

Clarke G, Fistein E, Holland T, **Barclay M**, Thiemann P and Barclay SIG.

Preferences for care towards the end of life when decision-making capacity may be impaired: A large scale cross-sectional survey of public attitudes in Great Britain and the United States.

PLOS One. 2017;12(4):e0172104. doi:10.1371/journal.pone.0172104

Spathis A, Hatcher H, Booth S, Gibson F, Stone P, Abbas L, **Barclay M**, Brimicombe J, Thiemann P, McCabe MG, Campsey R, Hooker L, Moss W, Robson J and Barclay S.

Cancer-Related Fatigue in Adolescents and Young Adults After Cancer Treatment: Persistent and Poorly Managed.

J Adolesc Young Adult Oncol 2017;6(3):489-93. doi:10.1089/jayao.2017.0037

Appendix 5. Presentations given during PhD

Based on PhD work

HSRUK Conference 2019, July 2019

- Problems with composite indicators of healthcare quality and safety
    - “Highly commended” – runner-up in oral presentation popularity contest

Department of Public Health and Primary Care PhD Presentations, June 2018

- Problems with composite indicators

Other

CRUK Cambridge Centre Early Diagnosis symposium, January 2019

- Statistical properties of the ‘early stage at cancer diagnosis’ indicator of the performance of commissioning organisations

PHE Cancer Services, Data and Outcomes Conference, June 2018

- The influence of patient case-mix on public health area statistics for cancer stage at diagnosis

Appendix 6. Conferences attended

Health Services Research UK (oral presentation), Manchester, July 2019

CRUK Cambridge Early Diagnosis Symposium (oral presentation), Cambridge, January 2019

National Cancer Research Institute Conference (poster), Glasgow, November 2018

Cancer Services Data and Outcomes Conference (oral presentation), Manchester, June 2018

Cancer Data and Outcomes Conference (poster), Manchester, June 2017

Appendix 7. Formal professional development undertaken during PhD

Qualitative and Mixed Methods Approaches in Primary Care module, Dr Jenni Burt, Department of Public Health and Primary Care University of Cambridge (part of the 'MPhil in Primary Care Research'), February 2019

NVivo: An Introduction for Qualitative Research, University Information Services University of Cambridge, January 2019

Doing qualitative interviews, Social Sciences Research Methods Centre University of Cambridge, October 2018

The Secrets of Effective Facilitation & Moderation, David Rose (LACS Training), October 2018

Data visualisation 'hack day', Public Health England, May 2018
(use of R and Shiny to produce interactive data visualisations)

Writing a journal article and getting it published, University College London, June 2017

Using simulation studies to evaluate statistical methods, Dr Tim Morris, University College London, May 2017