Using information on variations to improve health system performance: from measurement to management

Laura Schang

A thesis submitted to the Department of Management of
the London School of Economics for the degree of
Doctor of Philosophy, May 2015

The London School of Economics and Political Science

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of 58, 731 words.

Abstract

Although information on variations in healthcare utilisation is increasingly available, its constructive use to improve health systems is often hindered by the lack of a clear standard to evaluate what is “good“ and “poor“ performance. This thesis investigates how regulators and managers of the system might address this lack of a standard. The thesis distinguishes between the purpose (to manage ambiguity in the absence of a standard or to determine a meaningful standard) and the approach used to achieve either purpose (socio-political or technical-evidential). The resulting four types of strategies are examined by drawing on concepts and methods from public health and epidemiology, health economics, operations research and public administration and empirical evidence from England and Scotland.

To manage ambiguity in the absence of a standard using a socio-political approach, the thesis finds that one must overcome a series of barriers including awareness, acceptance, perceived applicability and capacity of potential users. Clinical and managerial leadership appear to be enabling factors for the use of information on variations for strategic problem framing and stakeholder engagement.

To manage ambiguity in the absence of a standard using a technical-evidential approach, the use of ranking intervals and dominance relations obtained from ratio-based efficiency analysis can help to avoid the forced assignment of a single, potentially controversial ranking to each organisation under scrutiny.

To determine a standard using a technical-evidential approach, estimating capacity to benefit in populations provides a theoretically sound and feasible benchmark to assess the appropriateness of service utilisation against population needs. However, uncertainty about criteria of capacity to benefit and lack of epidemiological data remain practical challenges.

To determine a standard using a socio-political approach, an experimentalist governance logic focused on learning and dialogue between central government and local organisations can complement a hierarchist logic focused on accountability when both the ideal ends and the means for attainment are ambiguous.

As a whole, the thesis reinforces the insight that both improved technical tools and social and political processes are required to make information on variations useful to decision-makers.

Acknowledgements

First and foremost I am extremely grateful to my PhD supervisor Professor Alec Morton for his tremendous support, invaluable advice and for giving me focus over these years, always with patience and insight. His incredibly fast feedback has made London and Glasgow seem almost next door. Following our discussions, I was always motivated to go on and had a direction to do so.

I will forever be indebted to Professor Gwyn Bevan for his generous advice and support which encouraged me to pursue this area of research on variations in healthcare in the first place, and for giving me the opportunity to work within the SyMPOSE team.

For their collaboration on Chapters 2, 3, 4 and 6 and for many helpful conversations, I would like to thank my co-authors Professor Alec Morton, Professor Gwyn Bevan, Dr. Mara Airoldi, Dr. Natalie Bohm, Phil DaSilva, Chiara De Poli, Yrjänä Hynninen, Professor Monica Lakhanpaul, Professor Ahti Salo and Professor Anne Schilder.

I would like to thank my PhD thesis examiners Professor Andrew Street and Dr. Irene Papanicolas for their helpful and constructive comments and for a thought-provoking and inspiring PhD thesis defense.

I would also like to express my gratitude to Dr. Sarah Thomson with whom it was a real pleasure and privilege to collaborate on several projects at the LSE. Her expertise, support and clarity of expression have influenced my thinking about health systems a lot.

I would like to thank Professor David Goodman for his support and interest in my work at the Wennberg International Collaborative conferences, and for our collaboration together with Devin Parker to extend the model of population capacity to benefit from ventilation tubes to the United States.

Sir Muir Gray has been decisive in making it possible for me to evaluate the use of the NHS Atlas among Primary Care Trusts in England, which sparked the key questions asked in this PhD thesis. I would like to express my sincere thanks for his vision, kindness and support.

Dr. Jan-Kees Helderman provided a key stimulus to develop the analysis of performance targets in Scotland. I am very grateful to him for his insightful and

constructive feedback about experimentalist governance that raised many intriguing questions about performance management in healthcare.

Thank you to our Doctoral Programme Director Professor Chrisanthi Avgerou and my colleagues in the PhD General Management community Samer Abdelnour, Rebecca Campbell, Zhuoqiong Chen, Enrico Rossi, Catherine Hawley, Laura Zimmermann, Michael Hayle, Nofie Iman, Adeline Pelletier and Nadia Millington for creating such a friendly, supportive and productive academic environment.

A special thank you to Eva Barrenberg, Jessica Kunert, Nora Esser, Jenny Brosius, Corinna Klingler, Lisa Trigg, David Herr, Elena Nicod and Alessandra Ferrario. My friends have made these years so much more enjoyable and were always a wonderful source of support.

My brother Christopher Schang, my aunt Dr. Barbara Fischer, my cousin Matthias Fischer and my grandmother Waltraud Schultz have offered me much support, serenity and energy over these years.

I owe my deepest gratitude to my parents Bettina and Dr. Thomas Schang whose encouragement, wisdom and steady support has given me strength to endure. Thank you for everything.

Statement of conjoint work

Chapters 2, 3, 4 and 6 are based on conjoint work with Professor Alec Morton (AM), Professor Ahti Salo (AS), Professor Anne Schilder (ASc), Chiara De Poli (CDP), Professor Gwyn Bevan (GB), Dr. Mara Airoldi (MA), Professor Monica Lakhanpaul (ML), Dr. Natalie Bohm (NB), Phil DaSilva (PD), Yrjänä Hynninen (YH). I contributed 90% of the work to Chapter 2, 70% of the work to Chapter 3, 80% of the work to Chapter 4, and 90% of the work to Chapter 6. My contributions to the conjoint work are summarised in Table 0-1. I am sole author of Chapters 1 and 7 and of the empirical Chapter 5, meaning I had the original idea for the study, designed the review protocol, collected and analysed the data and drafted the Chapters. AM commented on a previous version of each of these Chapters.

Table 0-1 Contributions to conjoint work

Chapter	The PhD author	The co-authors
2	• Had the original idea for the study • Developed the conceptual framework • Collected and analysed the data • Drafted the paper and led the submission and revision process for journal publication	• Supervised data collection and commented on the draft paper: AM, GB, PD
3	• Contributed to the planning of the study • Reviewed the literature and developed the theoretical sections • Collected the data • Contributed to the interpretation of the results • Drafted the paper	• Had the original idea for the study: AM • Analysed the data using REA and contributed to section 3.3.1 describing the REA approach: YH • Contributed to the planning of the study and to the interpretation of the results and commented on the draft paper: AM, YH, AS
4	• Contributed to the planning of the study • Developed the model and conducted the data analysis • Developed selection criteria for the systematic review and co-screened studies for inclusion in the model • Led the structured elicitation process at the expert workshop • Drafted the paper and led the submission and revision process for journal publication	• Had the original idea for the study: AM, ASc, GB • Contributed to the planning of the study: AM, ASc, CDP, GB, NB, MA • Co-screened studies for inclusion in the model: CDP • Co-facilitated the expert workshop: CDP, MA • Critically reviewed the draft paper: AM, ASc, CDP, GB, MA, ML, NB
6	• Had the original idea for the study • Collected and analysed the data • Drafted the paper	• Supervised data collection and critically reviewed the draft paper: AM

Statement of inclusion of previous work

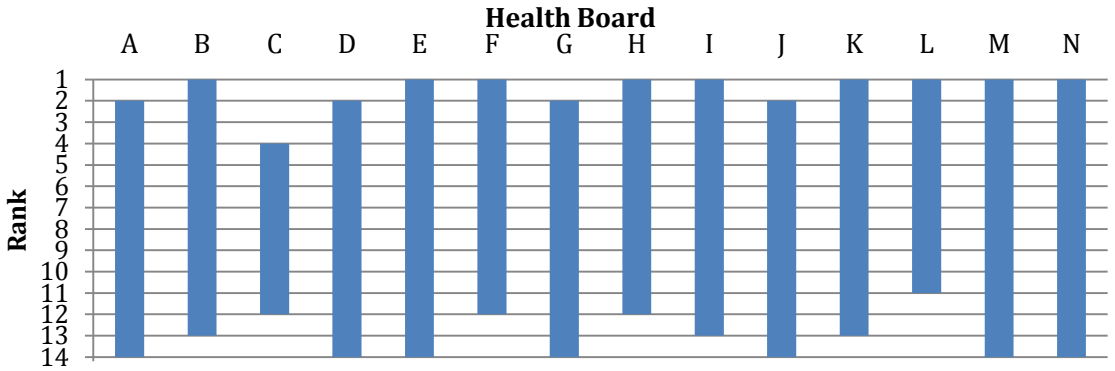
Chapter 2 draws on results from my MSc dissertation I undertook for the MSc International Health Policy in the Department of Social Policy at the London School of Economics and Political Science in 2010/11. The Chapter also draws on results from a project commissioned by NHS Right Care in 2011/12, for which I developed a series of case studies of how Primary Care Trusts were using the NHS Atlas of Variation in Healthcare (Schang and Morton, 2012).

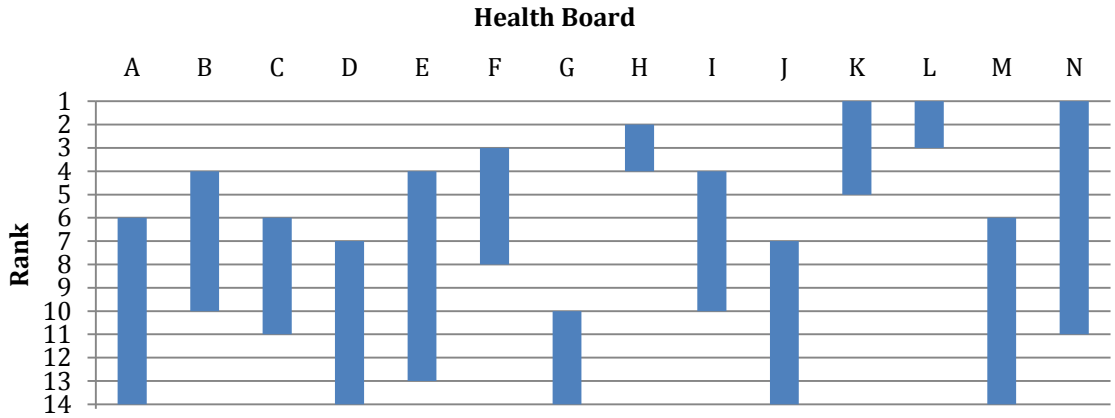
Chapter 2 of this thesis is a substantial development of these previous studies in the following ways: (i) it develops a conceptual framework of prerequisites for using information on variations; and (ii) it extends the scope of the empirical data basis from the initial 17 interviews included in the MSc dissertation to 45 interviews included in this PhD thesis. These additional interviews helped to substantiate the findings in relation to a larger sample and to clarify some important questions that were raised in my MSc dissertation, in particular with regard to enabling factors that appeared to foster the use of evidence from the NHS Atlas of Variation to inform local decision making.

Table of contents

List of tables

List of figures





List of abbreviations

A&E	Accident and Emergency
AOBD	Acute occupied bed days
CUF	Comparative utilisation ratio
dB (HL)	Decibels (hearing level)
DEA	Data envelopment analysis
HAI	Healthcare-associated infection
HEAT	Health improvement, Efficiency, Access and Treatment appropriateness
IDS	Immediate discharge service
ISD	Information Services Division
JIT	Joint Improvement Team
LDP	Local delivery plan
MRSA	Methicillin-resistant staphylococcus aureus
MSSA	Methicillin-sensitive staphylococcus aureus
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NZ score	New Zealand clinical priority score for hip and knee replacement
OECD	Organisation for Economic Co-operation and Development
OED	Oxford English Dictionary
ONS	Office for National Statistics
OME	Otitis media with effusion
PAF	Performance Assessment Framework
PCB	Population capacity to benefit
PCT	Primary Care Trust
QuEST	Quality, Efficiency & Support Team
RCT	Randomised controlled trial
REA	Ratio-based efficiency analysis
RTT	Referral to treatment
SAB	Staphyloccocus aureus bacteraemia
SBC	Shifting the balance of care for older people
STROBE	Strengthening the reporting of observational studies in epidemiology
UK	United Kingdom

US	United States of America
VT	Ventilation tube
WHO	World Health Organization
WOMAC	Western Ontario and McMaster Universities Arthritis Index

1 INTRODUCTION

Information on geographic variations in the utilisation of health services is increasingly available to policy-makers and managers. These variations are often interpreted as a marker of health system performance: as a signal of potential deficits in the appropriateness, equity and efficiency of service provision. However, prior research has tended to concentrate on the identification and measurement of variations rather than their management. The constructive use of this information is often hindered by the lack of a clear standard to evaluate what is "good" and "poor" performance. In the absence of such a standard, not only is it impossible to assess whether performance objectives have been achieved but the production and public reporting of information on variations also risks encouraging misinterpretation and causing harm. It is thus an opportune moment to investigate how regulators and managers in charge of planning, auditing and improving health services might address this ambiguity about the standard for evaluation.

The thesis includes five studies which aim to shed light on this problem from different perspectives. Two of the studies have been published in peer reviewed journals (Schang et al., 2014a, Schang et al., 2014b). The next section reviews the rationale for measuring variations in the use of health services. Subsequent sections set out the research problem, the aim and framework of the thesis, and summarise the contribution of the thesis.

1.1 Why measure variations in the use of health services?

Over the past decades, analysis of geographic variations in healthcare utilisation, expenditure and outcomes has developed into a growing field of research. A recent systematic review (Corallo et al., 2014) identified 1,114 studies of medical practice variation in OECD countries published between 1990 and 2011. National Atlases of Variation in countries including England, the United States, Germany, Spain, the

Netherlands, Canada, Australia and New Zealand have documented considerable variations in rates of hospital admission and surgical procedures between small areas (NHS Right Care, 2012a). At a cross-national scale, projects by the Organisation for Economic Co-operation and Development (OECD (Organisation for Economic Co-operation and Development), 2014) and by the European Collaborative for Healthcare Optimization (Bernal-Delgado et al., 2015) have shown that these variations exist both between and within countries.

Growing attention to evidence of variations in healthcare has evolved within a context of increasing interest in health system performance assessment more generally. Health system performance can be understood as the extent to which a health system meets its objectives (Hurst and Jee-Hughes, 2001, WHO, 2000). Measuring performance has therefore at its core an evaluative function: to ascertain the extent to which objectives such as appropriateness, equity and efficiency in the provision of services have been achieved. The importance of measuring health system performance is now widely recognised (Smith and Papanicolas, 2012, Smith et al., 2009). Advances in the quality of data reporting have resulted in unprecedented access to information on the operational performance of health systems. This information has different potential users and uses (Table 1-1). Taken together, these potential uses make performance measurement an essential part of health system governance and a key building block for the continuous improvement of health services (Berwick, 1996, Smith et al., 2009).

Table 1-1 Potential users and uses of information on performance

Potential users	Potential uses
Government and regulatory authorities	• Ensuring accountability for the effectiveness and efficiency with which resources are deployed. • Fostering improvement through appropriate regulations and incentives to purchasers and providers.
Organisations in charge of planning and purchasing health services	• Informing the planning of future service requirements. • Informing the contracting and management of healthcare providers.
Healthcare providers	• Targeting internal quality improvement efforts.
Service users	• Choosing healthcare providers.

Source: adapted from Smith et al. (2009) and Van der Wees et al. (2014).

As Goodman (2009: 749) points out, unwarranted variation in healthcare can be understood as "the variation in medical resources, utilization, and outcome that is due to differences in health system performance". Rising interest in evidence of geographic variations in healthcare use, as a potential "marker" of health system performance, is underpinned by two main policy concerns: equity concerns and concerns about appropriateness and efficiency.

Equity concerns stem from the belief that variations (inequalities) indicate inequities (unfair inequalities) in the use of health services. Variations appear to challenge the principle of horizontal equity (Oliver and Mossialos, 2004) which stipulates equal opportunity of access for equal need. This is a key concern for health systems such as the National Health Service (NHS) systems in England, Scotland, Wales and Northern Ireland which were founded on the principle of ensuring access to care based on need, regardless of ability to pay (Boyle, 2011). Evidence of variations suggests however that the type and intensity of care patients receive depends also on their place of residence, and associated factors such as supply structures and patterns of medical practice in a particular region (Busato et al., 2010, Couchoud et al., 2012). In media reports and public communication about variations, concerns about the apparent "postcode lottery" in local resource allocation decisions (The Guardian, 2011, NHS Right Care, 2011, Russell et al., 2013) imply worries that the quality of care patients receive depends on chance rather than medical need.

Appropriateness and efficiency concerns have come to the fore in light of increasing fiscal pressures on healthcare budgets. The English NHS, for instance, was required to generate efficiency savings of about 4% of total annual resources every year between 2011 and 2015 in order to meet rising demand for health services (Department of Health, 2010d). The magnitude of observed variations is often interpreted as a signal of widespread overuse and misuse of unnecessary or even harmful care (Maynard, 2013, Ham, 2013). This argument appeals to decision-makers as demographic changes and developments in medical technology challenge the financial sustainability of health systems (Busse et al., 2007). The economic crises that have affected several European nations add to these concerns (Thomson et al., 2014). Although evidence of variations does not necessarily imply evidence of inefficiency and inefficiency can exist in a system without there being any regional variations

(Göpffarth et al., 2015), tackling variations in medical practice is often cited as an opportunity to release resources that can be reinvested into care of higher value (Maynard, 2012, Rettenmaier and Wang, 2012, Huesch et al., 2013, Hollingworth et al., 2015).

Information on variations in healthcare is thus of interest due to its potential role as a signal that resources are not spent to best effect. As the presence of variations appears to contradict fundamental health policy objectives related to equity, appropriateness and efficiency, the identification of those variations that are unwarranted is increasingly recognised as a key policy challenge (OECD (Organisation for Economic Co-operation and Development), 2014).

1.2 Ambiguity about the standard for evaluation

A fundamental problem in analyses of geographic variations in healthcare, and performance comparisons more generally, lies however in defining a meaningful standard for evaluation. Following the Oxford English dictionary, the term "standard" is understood here as a "definite level of excellence, attainment, wealth, or the like, or a definite degree of any quality, viewed as a prescribed object of endeavour or as the measure of what is adequate for some purpose" (OED, 2015). In short, a standard refers to a stipulated normative level of quality or attainment that is used for comparative evaluations. In this thesis, the term standard is used interchangeably with the terms benchmark, yardstick, norm and reference point. This variation in terminology reflects the surprising fact that the notion of what constitutes an appropriate standard has not received the focused attention it deserves, even though, by definition, there is no way to assess whether system performance is "good enough" except with reference to a standard (Donabedian, 1981).

In order to evaluate the appropriateness of variations in healthcare, it is essential to establish a standard of what is meant by "good" and "poor" performance. For health outcome indicators and some indicators of the care process, ideal levels of attainment are obvious. For instance, no patient should have to die from a healthcare-associated infection (the ideal rate is zero). If the objective is to ensure equal access to cost-

effective care for equal need, then one might say that every person with diabetes should receive key interventions such as regular foot checks and eye examinations (the ideal rate is 100 per cent).

However, for most indicators that are concerned with the utilisation of health services, the standard for evaluation is essentially ambiguous. In the case of geographic variations in rates of hip replacement, for instance, it is not clear what a given utilisation rate means in terms of the appropriateness of care provided. As Robert Evans has pointed out already in 1990, potential users of research on variations are therefore confronted with a puzzle: "Are the regions, or institutions, or practitioners with high rates over-providing, or are the low ones under-providing, or does the 'best' rate lie somewhere in the middle (or beyond either end)?" (Evans, 1990: 127). More than two decades later, a recent systematic review (Mercuri and Gafni, 2011) identified a fundamental lack of theoretically sound, empirically measurable frameworks to evaluate unwarranted variations in healthcare. As a result, resolving Evans' puzzle continues to present a key hurdle to using information on variations to inform decision-making on health services (Appleby et al., 2011, Tanenbaum, 2012, OECD (Organisation for Economic Co-operation and Development), 2014, Hollingworth et al., 2015).

Addressing ambiguity about the standard for evaluation is important for several reasons. First, in the absence of a clear standard, policy-makers and managers forego potential insights about health system performance. Information on rates of hospital admissions and surgical procedures does, in itself, not allow for normative inferences about the value these services confer on people. It is hence not possible to evaluate the performance of the health system based on this information if, as stated in section 1.1, performance is defined as the degree to which objectives such as ensuring the provision of services in relation to medical need have been met.

Second, reporting information on variations without a clear standard risks encouraging misinterpretation and causing harm. It has long been argued that publishing performance information in the public sector may not only have beneficial effects but also several unintended consequences (Smith, 1995, Casalino, 1999). In the context of variations in healthcare, there is a tendency in public communication

about variations to portray regions with comparatively high utilisation rates as being afflicted by high levels of inappropriate care and regions with comparatively low rates as being characterised by high levels of unmet need (Tanenbaum, 2012).

However, existing evidence does not indicate a systematic relationship between high rates of utilisation and high levels of inappropriateness (Keyhani et al., 2012). Focusing management attention (solely) on the regions with high rates of utilisation so as to reduce presumed "overuse" seems therefore premature. Making inferences about "good" and "bad" based on information that is essentially descriptive in nature (such as the number of hip replacement operations in Liverpool compared to London) is what Rein (1976: 75) calls a "normative leap" that may lead to false conclusions about the performance of the systems studied. Consequent attempts to reduce variations across the board may eliminate also those variations that exist for legitimate reasons, such as responsiveness to population needs and patient preferences (Folland and Stano, 1990, Lilford, 2009).

Finally, in a context where resources are inevitably limited, one may question the utility of collecting and reporting data that has no clear managerial implications and is not directly linked to health system objectives. While there is no shortage of performance indicators – the 2012 US Agency for Healthcare Research and Quality (AHRQ) National Healthcare Quality and Disparities Reports, for instance, contained more than 250 quality indicators – little effort has been dedicated to prioritising the reporting of metrics in terms of their likely contribution to population health (Meltzer and Chung, 2014). This risks turning measurement into an end in itself rather than into a means to foster accountability and improvement (Spiegelhalter, 1999, Goddard et al., 2000).

In order to be able to evaluate variations in healthcare and draw implications for policy and management, achieving clarity about the standard for evaluation is therefore essential. The next section sets out a framework of how this problem might be approached and reviews limitations of prior research against this framework.

1.3 Aim and framework of the thesis

The overarching aim of this thesis is to investigate strategies through which policy-makers and managers in charge of planning, auditing and improving health services might address ambiguity about the standard for evaluation. It is proposed here that strategies to address this ambiguity can be classified according to two dimensions: their purpose and the approach by which this purpose is achieved (Table 1-2).

The purpose of addressing ambiguity about the standard for evaluation can be (i) to establish a meaningful standard; or (ii) to manage ambiguity in the absence of a standard. The approach by which either purpose is achieved can be (i) socio-political (using particular models of governance or management strategies); or (ii) technical-evidential (using particular metrics or methods for analysis). The resulting four categories are explained below in relation to prior research.

Table 1-2 Taxonomy of strategies to address ambiguity about the standard for evaluation

		Approach	
		Socio-political	**Technical-evidential**
Purpose	**Manage ambiguity in the absence of a standard**	Use models of governance or processes for decision making that recognise ambiguity about the standard.	Use methods for analysis that recognise ambiguity about the standard.
	Establish meaningful standards	Determine standards through suitable models of governance or ways of decision making.	Determine standards through suitable metrics or methods for analysis.

1.3.1 Managing ambiguity with a socio-political approach

Much of the work that public sector organisations do takes place in challenging organisational and political environments. Typically, governments, healthcare purchasers and providers are faced with multiple and conflicting demands such as the requirement to operate within a budget constraint while seeking to improve the quality of services delivered. In such a context, information that lacks clear managerial implications in terms of cost reduction, quality improvement or both is

easily overlooked. Prior research on the public reporting of comparative clinical performance data in Scotland (Mannion and Goddard, 2001, 2003) showed that, in the absence of supportive organisational contexts and appropriate incentives, the mere disclosure of performance information had little effect on the behaviour of NHS Trusts.

Attempts to motivate public sector organisations to engage with evidence of variations in healthcare utilisation thus requires an understanding of the barriers organisations face in practice and the strategies they might adopt to overcome these. As research on guideline implementation (Glasziou and Haynes, 2005) shows, the path from evidence to its use in decision making requires potential users to be aware of the data, accept its validity, perceive it as being applicable to their situation and have the capacity to use the data. However, little research seems to have explored these issues in the context of variations in healthcare utilisation. While some work exists on strategies adopted by US hospitals to tackle unwarranted clinical practice variations (Gauld et al., 2011), there is a lack of studies asking how and to what extent managers in charge of planning and purchasing services might make sense of information on variations in healthcare.

This category is therefore concerned with building understanding of the practical barriers managers of the health system face and of the strategies they use – and might use – to make sense of potentially confusing information on variations in healthcare.

1.3.2 Managing ambiguity with a technical-evidential approach

A different response to ambiguity about the standard for evaluation is to target not (only) the intended users of this information and the environments in which they operate, as under a socio-political approach, but to strengthen the ways in which metrics are reported. If standards for evaluation are absent or controversial, then methods of analysis could be deployed that recognise this ambiguity.

A typical example in the context of healthcare is the ranking of organisations based on a composite measure of performance. Since the provision of health services is typically required to meet multiple objectives (e.g. treatment appropriateness, equity

and responsiveness), composite measures are intended to provide a unified assessment of organisational performance across different domains (Smith, 2002). However, rankings on the basis of composite measures tend to rely on controversial assumptions about the relative weights of component indicators (Cherchye et al., 2007, Aron et al., 2007, Ferguson et al., 2002, Goldstein and Spiegelhalter, 1996, Kang and Hong, 2011, Goddard and Jacobs, 2009, Jacobs et al., 2005). Different techniques for determining weights – from simpler trade-off methods including ranking from most to least desired indicator and voting techniques to more elaborate multiattribute approaches such as conjoint analysis and the analytic hierarchy process – often produce different results and each method has distinct advantages and disadvantages in terms of feasibility, consistency and validity (Dolan, 1997, OECD, 2008, Appleby and Mulligan, 2000). Although it is possible to establish a single set of weights, doing so may disguise the underlying value judgements and hence limit the credibility and transparency of performance assessments based on composite scores (Hauck and Street, 2006).

Moreover, although many performance indicators are constructed as ratios (e.g. number of healthcare-associated infections/ 100,000 acute occupied bed days), it is often unclear which variables should be employed as denominators. In productivity and efficiency analysis, denominators are the inputs (labour, capital, intermediate inputs such as drugs and clinical supplies) used to produce particular quality measures (Jacobs et al., 2006b, Bojke et al., 2013). In comparisons of healthcare quality, the denominator of a ratio should be the population at risk of experiencing an event (Romano et al., 2010, Schlaud et al., 1998). Operationalising this principle is however far from straightforward (Marlow, 1995). Despite its centrality in healthcare quality assessments, little effort has been dedicated to identifying ways how to handle potential bias introduced by denominators that either overestimate or underestimate the "true" population at risk (Guillen et al., 2011). As a result, one cannot be confident that observed variations in healthcare are due to differences in system performance (the numerator) or due to misspecified populations at risk (the denominator).

A key research question to be asked in this category is therefore how the reporting of information on variations in performance might be strengthened so as to recognise the ambiguity about key modelling assumptions.

1.3.3 Establishing standards with a socio-political approach

It can be argued that policy-makers should tackle the ambiguity about the standard for evaluation "head-on" by investing more effort into establishing a meaningful standard. This is in fact a core recommendation made by international organisations such as the OECD to governments who seek to develop performance metrics (Hurst and Jee-Hughes, 2001, Hurst, 2002). After all, one of the legitimate mandates of democratically elected governments or relevant regulatory authorities is to define priorities and set standards for the health service.

In healthcare, setting targets for local organisations is a form of measuring and managing performance that by definition requires the specification of levels of "good" performance. Based on the "targets and terror" model of governance adopted by the Labour government in England between 2001 and 2005, target setting is typically conceived as a hierarchical process where central government imposes strict targets on local organisations with rewards for achievement and sanctions for failure (Bevan and Hood, 2006).

Such a hierarchist model of governance, however, requires "dials" (Carter, 1989): accurate measures of performance which unambiguously represent desired policy ends (Bevan and Hood, 2006) and whose means of attainment are known and available to the organisations under scrutiny (Jacobs et al., 2006a). Many performance indicators in social policy, however, are mere "tin openers": measures which "do not give answers but prompt interrogation and inquiry, and by themselves provide an incomplete and inaccurate picture" (Carter, 1989: 134). This holds in particular for "wicked" problems where goals are contested and means for change are ambiguous (Rittel and Webber, 1973). Some have therefore argued that setting delusively exact targets for wicked problems such as health inequalities obscures complex causal networks and necessary value judgements in determining desired levels of achievement (Blackman et al., 2009).

A key research question to be asked in this category therefore concerns the appropriate model of governance to establish standards and good practice in a context of ambiguity over ideal goals and means for attainment.

1.3.4 Establishing standards with a technical-evidential approach

As Smith and Street (2005) point out, articulation of the health system's objectives is an inherently political task. Value judgements about the objectives of a system, and corresponding standards for evaluation, cannot be derived from technical analysis (Popper, 1948). Nevertheless, research and analysis certainly have a role in the development of standards.

Most OECD countries have expressed overarching objectives for their health systems through policy documents or legislation. At an abstract level, objectives such as ensuring equal access to care for equal need may be widely accepted. In the United Kingdom, provision of services in relation to need (rather than ability to pay or non-medical factors) is in fact a foundational principle of the NHS (Boyle, 2011). Translating this principle into measurable, interpretable indicators of performance is, however, a complex endeavour.

Studies of variation in healthcare provision have sought to account for need using either of two approaches: clinical audits of care provided (see e.g. Chassin et al., 1987, Leape et al., 1990, Keyhani et al., 2008b) or standardisation of rates for variables associated with need (such as age, sex, deprivation) (see e.g. Majeed et al., 2002, Curtis et al., 2009, NHS Right Care, 2010). The former approach can only detect "overuse", defined as "ineffective care that is more likely to harm than help the patient" (Institute of Medicine, 2001: 47). It cannot identify "underuse", defined as "the failure to provide services from which the patient would likely benefit" (Institute of Medicine, 2001: 17). The latter approach is essential to enable fair comparisons between regions whose performance may differ due to factors that are outside the region's control (Nicholl et al., 2013). However, standardised rates cannot provide a benchmark of population need for healthcare, understood as the extent to which utilisation of a specific service exceeds or falls short of a level of care that is expected to be beneficial for a defined population.

A key research question to be asked in this category therefore concerns the appropriate methods to operationalise health system objectives such as "provision of services in relation to need for healthcare" in such a way that resulting estimates can

be employed as benchmarks to identify both "overuse" and "underuse" of specific interventions.

1.3.5 Use of the framework in this thesis

The rationale for the framework described above is to organise thinking about the problem of how to address ambiguity about the standard for evaluating variations in healthcare utilisation. The resulting four categories of strategies are not meant to be mutually exclusive. They might plausibly complement each other. Taken together, they offer different perspectives on the research problem and provide the overarching frame for the five studies included in this thesis.

1.4 Contribution of the thesis

To investigate how policy-makers and managers might address ambiguity about the standard for evaluation, this thesis presents five studies (Chapters 2 to 6). These studies examine different strategies to manage ambiguity about the normative standard. The contribution of each study is highlighted below and summarised in relation to the framework in Table 1-3.

Chapter 2 contributes to the health policy literature about practical barriers faced by potential users of information on variations in healthcare. To our knowledge, ours was the first study that sought to examine to what extent and how healthcare payers have used information on small area variation in rates of expenditure, activity and outcome in order to improve resource allocation. The study set out a model to frame the process of moving from information on variations to its use in decision-making on health services (Figure 2-1, Chapter 2). Taking the NHS Atlas of Variation in Healthcare in England as a case study, we examined barriers to and types of use of this information. Data collection involved a survey among Primary Care Trust (PCT) Chief Executives and a telephone follow-up to reach non-respondents (total response: 53 of 151 of PCTs, 35%). 45 senior to mid-level staff were interviewed to probe themes emerging from the survey.

The results showed that just under half of the respondents (25 of 53 PCTs) reported not using the Atlas, either because they had not been aware of it, lacked staff capacity to analyse it, or did not perceive it as applicable to local decision-making. Among the 28 users, the Atlas prompted further analysis of the reasons for variations. It was also used as a visual aid to simplify communication about comparative performance with clinicians who perceived maps as a more accessible tool for problem framing than complex statistical tables. However, only 18 of the 28 PCTs who had reviewed the Atlas also reported concrete actions taken for healthcare planning, contracting or service design. Factors that appeared to enable local managers to move beyond the data towards decisions about resource allocation and behaviour change included agreeing on responsibilities for action and the ability to define and identify those variations that were unwarranted.

These findings demonstrate that information on variations can serve as a "tin opener" (Carter, 1989) to motivate further analysis and inform strategic planning even in the absence of an a priori standard of "good" and "bad" performance. To achieve this, however, what is additionally required is leadership on variations and the provision of appropriate tools to understand which variations are unwarranted.

Chapter 3 contributes to the technical literature on healthcare performance assessment when there are multiple objectives (e.g. Hauck and Street, 2006, Castelli et al., 2015). Specifically, the study explores healthcare applications of a robust approach to ranking organisations based on a composite indicator of performance in a context of ambiguity about choices of weight sets and choices of population denominator. To that end, the study adopts a novel ratio-based efficiency analysis (REA) technique developed by operations researchers (Salo and Punkka, 2011). Previously, REA has been used to assess the efficiency of higher education institutions based on multiple inputs and quality measures (Salo and Punkka, 2011).

A key advantage of REA is its ability to use the full set of feasible weights and to take into account multiple denominator variables that represent different plausible definitions of the "population at risk". This avoids the need to settle on a single, potentially controversial set of weights and on a single, possibly biased denominator population. The results (displayed as ranking intervals and dominance relations)

allow one to identify organisations which cannot be ranked, say, worse than 4th or better than 7th. Using data from the Scottish HEAT target system, the study demonstrates the applicability of REA to comparative performance assessment in healthcare.

The study is important because it shows that assigning a single performance ranking to an organisation may not only be questionable from a policy and management perspective, but also unnecessary from a technical perspective. The use of ranking intervals and dominance relations provides one possible way to show the impact of different modelling assumptions on the results. Because REA is able to make explicit and visualise the uncertainty in rankings, the study argues that REA provides a method that may help increase the transparency and credibility of performance rankings and thus usefully complement existing tools of healthcare evaluators.

Chapter 4 makes an empirical contribution to the identification of "overuse" and "underuse" of ventilation tube (VT) surgery for otitis media with effusion (OME) against an explicit normative standard. Ventilation tubes are an insightful case study because they represent a classic case of high variation at a small area level (NHS Right Care, 2012b) and because there is a long-standing belief, since the 1980s, that these variations represent widespread overuse (Black, 1985c). This belief is supported by clinical audits in the US (Keyhani et al., 2008a) and the UK (Daniel et al., 2013) which, using different criteria of appropriateness, found that only one in three ventilation tubes was provided in line with these criteria. However, audits cannot identify the scale of underuse: i.e. patients who would benefit but are not treated.

To identify both overuse and underuse, Chapter 4 develops an epidemiological model based on: definitions of children with OME expected to benefit from VTs according to guidance from the National Institute for Health and Care Excellence (NICE); epidemiological and clinical information from a systematic review; and expert judgment. The study finds that the expected population capacity to benefit from VTs for OME based on NICE guidance exceeds, by far, the number of VTs actually provided in the NHS. About 32,200 children in England would be expected to benefit from VTs for OME per year (between 20,411 and 45,231 with 90% certainty). The observed number of VTs for OME-associated diagnoses in the NHS in 2010 however was

16,824. Hence, there appears to be substantial net "underuse" of VTs for OME if NICE criteria were applied.

These findings are important because they challenge a common policy among healthcare payers in England to improve treatment appropriateness by means of restricting access to VTs and other procedures that are deemed to be of "low clinical value" (Audit Commission, 2011). Our findings demonstrate the potential co-existence of overuse and underuse at a population level and therefore call for a more nuanced policy response.

Chapter 5 examines the feasibility and utility of a specific methodology to assess the appropriateness of variations in the use of specific interventions against a measure of population need for these interventions. Grounded in a health economic view of need for healthcare in relation to the capacity to benefit from healthcare (Culyer and Wagstaff, 1993, Culyer, 1995, Stevens and Gillam, 1998, Mooney and Houston, 2004), the study defines the concept of population capacity to benefit PCB_i^k as the number of people in some population or region *k* with a specified condition-intervention pairing *i* which represents the capacity to benefit from some intervention given defined characteristics of the health state. The study suggests that estimates of PCB may serve as a benchmark to identify potential gaps between need for and utilisation of defined interventions. Because these estimates represent a level of care that is expected to be beneficial for a specific population, they overcome a key limitation of the conventional method to account for need by standardising observed rates of utilisation for variables associated with need.

To synthesise the existing state of knowledge, the study reviews empirical applications of the PCB concept. The review identified 22 studies published between January 1990 and 2015 which applied the PCB approach to nine clinical areas in total, including amongst others hip and knee replacement for people with osteoarthritis and radiotherapy for different types of cancer. These findings show how the theoretical principle of "population need" for a specific intervention can be operationalised in practice using established methods from Health Technology Assessment (HTA) and epidemiology. The study is important because it shows the

feasibility and utility of attempting to measure capacity to benefit in populations. It also highlights persisting challenges with regard to the availability of credible criteria of capacity to benefit and accurate and comprehensive data on the incidence of these criteria in a population of interest.

Chapter 6 contributes to the public administration literature on models of governance for setting performance targets. The study asks how and to what extent a more learning-oriented logic of experimentalist governance (Sabel and Zeitlin, 2012) might complement hierarchist governance by targets focused primarily on accountability. The study postulates (i) that it is possible to disentangle and examine empirically the co-existence of hierarchist and experimentalist elements in the same performance management regime; and (ii) that the relative emphasis on experimentalist as opposed to hierarchist logics differs between policy issues depending on the degree of perceived ambiguity over ends and means.

Using a comparative embedded case study design (Yin, 2003), the study compares, within the Scottish HEAT target system, the development of HEAT targets for two policy issues which represent opposite ends on a spectrum of ambiguity over goals and means. Where ends and means were contested (the case of shifting the balance of care for older people; a typical "tin opener"), we find a stronger focus on experimentalist ideas in the form of locally agreed targets and a focus on local innovation. Where both ends and means seemed obvious (the case of healthcare-associated infections; an apparent "dial"), hierarchist elements dominated initially. However, management style drifted towards the experimentalist realm when rising rates of community-acquired infections decreased clarity about effective interventions.

To summarise, the thesis approaches the problem of ambiguity about the standard for evaluation from an interdisciplinary perspective. It draws on concepts and methods from different scientific traditions, including public health and epidemiology, health economics, operations research and public administration. As a whole, the thesis adds to our knowledge about using information on variations in healthcare. It reinforces the insight that both improved technical tools and social and political processes are required to make this information useful to decision-makers.

Table 1-3 Contribution of the five studies in relation to strategies to address ambiguity about the standard for evaluation

		Approach	
		Socio-political	**Technical-evidential**
Purpose	**Manage ambiguity in the absence of a standard**	**Chapter 2** • Provides a model to frame the process of moving from the measurement of variations in healthcare to their management. • Investigates barriers along this process faced by healthcare payers in England and examines strategies to make sense of the information locally.	**Chapter 3** • Explores healthcare applications of a robust approach to ranking using ratio-based efficiency analysis (REA; Salo and Punkka, 2011) based on ranking intervals and dominance relations that recognise ambiguity about choices of weights and choices of population denominator.
	Establish meaningful standards	**Chapter 6** • Examines how an experimentalist governance logic focused on learning and dialogue between central government and local organisations (Sabel and Zeitlin, 2012) might complement a more hierarchist philosophy focused on accountability when setting performance targets. • Contributes an empirically-based characterisation of the co-existence and potential complementarity of these logics in the Scottish HEAT target system.	**Chapter 4** • Develops an epidemiological model to investigate overuse and underuse in ventilation tube surgery for children with otitis media with effusion in England. • Shows that underuse and overuse may co-exist and that a more nuanced policy is required to increase appropriateness in the provision of ventilation tubes. **Chapter 5** • Defines the concept of population capacity to benefit (PCB) as a potential benchmark for population need for defined interventions. • Critically reviews the feasibility and utility of measuring PCB, its generalizability across conditions and persisting challenges.

2 FROM DATA TO DECISIONS? EXPLORING HOW HEALTHCARE PAYERS RESPOND TO THE NHS ATLAS OF VARIATION IN HEALTHCARE IN ENGLAND

Published as:

SCHANG, L.,[a] MORTON, A.,[a] DASILVA, P.[b] & BEVAN, G.[a] 2014. From data to decisions? Exploring how healthcare payers respond to the NHS Atlas of Variation in Healthcare in England. *Health Policy,* 114(1), 79-87.

Author information (for the time when the research was conducted):

[a] Department of Management, London School of Economics and Political Science, United Kingdom

[b] NHS QIPP Right Care Programme and NHS Derbyshire, United Kingdom

Correspondence to:

Laura Schang
Department of Management
London School of Economics and Political Science
Houghton Street | London | United Kingdom
Email: L.K.Schang@lse.ac.uk

Key words: resource allocation; small-area analysis; unwarranted variations; regional health planning; organisational decision making; quality indicators.

Abstract

Purpose: Although information on geographic variations in healthcare is now more widely available, relatively little is known about how healthcare payers use this information to improve resource allocation. We explore to what extent and how Primary Care Trusts (PCTs) in England have used the NHS Atlas of Variation in Healthcare, which has highlighted small area variation in rates of expenditure, activity and outcome.

Methods: Data collection involved an email survey among PCT Chief Executives and a telephone follow-up to reach non-respondents (total response: 53 of 151 of PCTs, 35%). 45 senior to mid-level staff were interviewed to probe themes emerging from the survey. The data were analysed using a matrix-based Framework approach.

Findings: Just under half of the respondents (25 of 53 PCTs) reported not using the Atlas, either because they had not been aware of it, lacked staff capacity to analyse it, or did not perceive it as applicable to local decision-making. Among the 28 users, the Atlas served as a prompt to understand variations and as a visual tool to facilitate communication with clinicians. Achieving clarity on which variations are unwarranted and agreeing on responsibilities for action appeared to be important factors in moving beyond initial information gathering towards decisions about resource allocation and behaviour change.

Conclusions: Many payers were unable to use information on small area variations in expenditure, activity and outcome. To change this what is additionally required are appropriate tools to understand causes of unexplained variation, in particular unwarranted variation, and enable remedial actions to be prioritised in terms of their contribution to population health.

2.1 Introduction

Over the past 40 years, medical variation research has largely focused on the identification and measurement rather than the management of variations in healthcare. Studies in particular from North America and increasingly also from other countries show that medical practice varies across regions, and that the magnitude of these variations cannot solely be explained by differences in demographic and illness profiles of regional populations (Wennberg and Gittelsohn, 1973, McPherson et al., 1982, Paul-Shaheen et al., 1987, Wennberg, 2010). Evidence of substantial variations in medical practice thus challenges the core societal objective of many health systems to provide equal access to safe and effective health care for equal need (Evans, 1990, McGlynn, 1998). But while healthcare payers now have unprecedented access to data about variations in health service utilisation and performance, there is little research on how payers might actually use this data to improve resource allocation and outcomes. Studies so far have focused on shared decision making (O'Connor A et al., 2004, Elwyn et al., 2010) and behaviour change interventions at a hospital level (Wright et al., 2006, Parente et al., 2008, Gauld et al., 2011).

However, the ways in which regional variations data might inform resource allocation at a population level by those responsible for the management of the system have not been explored. In this article we ask how a healthcare payer in charge of planning and purchasing health services for a geographical population might move from data awareness to decisions to improve quality and value in healthcare. Realising this basic quest may not be straightforward, as Glasziou and Haynes (2005) point out in the context of guideline implementation, because the path from research to improved outcomes poses a series of hurdles to clinical and managerial decision-makers. Prior to acting on the research findings, they need to be aware of and accept the data, perceive the data as applicable to their situation, and be able to use the data. These barriers seem pertinent to research use in general (Nutley et al., 2007). Data on medical practice

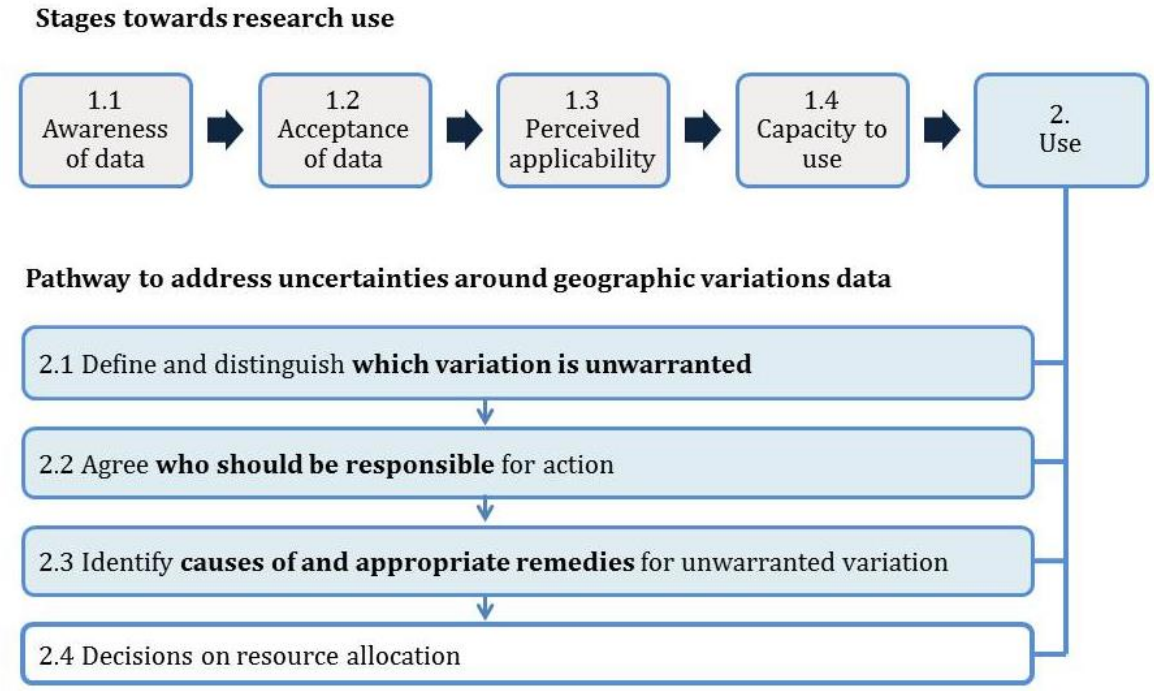
variations create the additional conundrum that, as opposed to a guideline, they rarely tell the user what to do.

There appear to be two general pathways for taking action on medical practice variation. The two principal aims of performance indicator systems stated in the literature relate to external control and accountability, and internally focused improvement and formative learning (Solberg et al., 1997, Freeman, 2002, Davies, 2005). Similarly, Carter (1989) distinguishes between "dials" that show achievement against targets, and "tin openers" that simply indicate potential problems and then lead to in-depth analysis and action. For both types of indicators, action would require agreement on who is responsible for leading investigation and change, and how to identify and remedy the causes underlying those identified variations. A key feature of classic variations research, as presented in Atlases of Variation (NHS Right Care, 2010, The Dartmouth Institute, 2012, Nolting et al., 2011, Bernal-Delgado et al., 2014), is however the essential ambiguity over the meaning of observed variations. Generally this data does not allow for direct inferences from relative rates of activity to good or bad performance of the entities under investigation. As optimal performance is not identified, this data thus differs from benchmarking where all organisations are compared with the "best" performer (Bogan and English, 1994). In this case, geographic variations data is likely to serve as a "tin opener" rather than as a "dial". As Evans (1990) pointed out, dealing with the uncertainty over how to address practice variations would thus first require defining and operationalising which part of the observed variations, if any, is unwarranted.

Figure 2-1 suggests a model to frame the process of translating evidence of geographic variations into decisions to shape resource allocation and planning. This model comprises two main stages. The first stage is informed by the literature on guideline implementation (Glasziou and Haynes, 2005) and research use (Nutley et al., 2007) and consists of a series of prerequisites for staff in a healthcare purchasing organisation to be in a position to use such evidence: that they are aware of its existence, trust the information it provides, can see its relevance to them and are capable of using this information. The second stage is structured around the pathway for using the

information (Evans, 1990): identifying unwarranted variation, agreeing who will be responsible for action, identifying causes and appropriate remedies, and making decisions on resource allocation.

Figure 2-1 A framework for moving from data on geographic variations to resource allocation decisions



Sources: adapted from Glasziou and Haynes (2005) and Evans (1990).

This model frames the questions our research sought to answer. As a case study we used the NHS Atlas of Variation in Healthcare, which in its first edition from November 2010 highlighted variation in expenditure, activity and outcomes across a wide range of clinical areas at the level of Primary Care Trusts (PCTs), the local payers in England (NHS Right Care, 2010). Our aim was to examine: (1) the extent to which PCTs met the prerequisites for using the NHS Atlas; and (2) how they were using the NHS Atlas in local decision making. We emphasise that most of this study was done before the publication of the second edition of the Atlas. We would expect awareness and capacity to use information on variations to increase over time and see this study as helping with both.

2.2 Materials and methods

2.2.1 Setting

At the time of study (July 2011–March 2012), the planning and delivery of health services in the National Health Service (NHS) in England was entrusted to 151 PCTs. They received a fixed financial allocation for their local populations (median size 284,000, ranging from under 100,000 to over one million people; Office for National Statistics (2011)) with reference to a national resource allocation formula that aimed to estimate an equitable distribution of funds against needs across the country (Boyle, 2011). Within allocated resources, PCTs were responsible for: improving health and reducing health inequalities, securing access to comprehensive, effective and efficient services, and appropriately responding to the healthcare needs of their populations. They were responsible for commissioning health services across all service sectors (public health, primary care services including dentistry, pharmacy and optometry, community health services, social care, mental health, elective and acute hospital care) and were required to engage in (Department of Health, 2006):

1) Strategic planning: assessing needs, reviewing service provision, deciding priorities.
2) Procuring services: designing services, shaping the structure of supply, managing demand for services.
3) Monitoring and evaluation: supporting patient choice, managing performance, seeking public and patient views.

The English NHS at the time of study was under expenditure constraints and required to generate efficiency savings of about 4% of total annual resources every year between 2011 and 2015, in order to meet rising demand for health services (Department of Health, 2010d). The proposed organisational reform outlined in the government White Paper Equity and Excellence: Liberating the NHS of July 2010 (Department of Health, 2010a) entailed the abolition of PCTs in April 2013, to be succeeded by general

practitioner-led Clinical Commissioning Groups. Thus, although information on variations has potential to help managers understand and focus on areas for efficiency savings in their local health economy, to be invested in areas of higher value, PCTs were likely to be distracted by their looming abolition.

2.2.2 The NHS Atlas of Variation in Healthcare

Since Glover's seminal study on variation in tonsillectomy rates among British school children in 1938 (Glover, 1938), research has repeatedly documented regional variation in medical practice in England (e.g. McPherson et al., 1982, Price et al., 1992, Congdon and Best, 2000, Majeed et al., 2002, Appleby et al., 2011). Our focus was specifically on the NHS Atlas of Variation in Healthcare, because this Atlas for the first time highlighted variation in expenditure, activity and outcome across a large range of clinical areas at PCT level and was thus likely to be particularly relevant within a commissioning context. Inspired by the US Dartmouth Atlas, the NHS Atlas was developed by the Department of Health's national Quality, Innovation, Productivity and Prevention (QIPP) programme, a large scale transformational programme intended to address these four major challenges confronting the NHS (Department of Health, 2010d), through the Right Care workstream. The first NHS Atlas, published in November 2010 (NHS Right Care, 2010), consists of 34 maps of variation (2011 Atlas: 71 maps; NHS Right Care (2011)). These maps represent the relative position of PCTs in quintiles across selected indicators, standardised for age and sex. The topics were selected in consultation with the National Clinical Directors as being of importance to their clinical specialty; for instance in terms of volume, cost, patient outcomes, or recent trends in delivery patterns.

The NHS Atlas was primarily targeted at those who manage and allocate resources for healthcare; commissioners and clinicians. Its objective was to provide information in ways that would stimulate local investigation into unwarranted variation in the NHS, its underlying causes, and remedial action. Given the complexity and variety of the different kinds of variations reported in the NHS Atlas, there was neither ranking nor evaluation of the performance of NHS organisations; nor were there any links with (external)

financial incentives. This differs from NHS star ratings (2000–2005), and the Annual Health Check (2006–2009) which gave annual summative aggregate scores of performance (Bevan, 2011); and more recent care quality targets that clearly define successful achievement (Department of Health, 2010c). The NHS Atlas carefully avoids rating PCTs as "good" or "bad" performers based on high, middle or low indicator values. Targets or "optimal" rates of activity are not defined. However, in the wide-ranging media echo to the NHS Atlas, several think tanks, academics, charities and politicians interpreted the magnitude of regional variations as indications of unwarranted variation, and urged PCTs and the government to take action (Jeffreys, 2010, The Guardian, 2011, Mays, 2011).

2.2.3 Study design

The first part of data collection involved an email survey with open-ended questions among the Chief Executives of all 151 PCTs. Given the low response (18 of 151 of PCTs, 12%), non-respondents were followed-up by telephone (total response: 53 of 151 of PCTs, 35%). The data were collected in two waves (Wave I: July to August 2011; Wave II: October 2011 to March 2012). The survey was designed to gain an indicative overview of whether the Atlas was used, why or why not, in what form and by whom, and to identify potential interviewees (see Appendix 2-A for the questions asked). The second part of the research involved interviews based on a semi-structured protocol in order to probe themes emerging from the survey (see Appendix 2-B for the questions asked).

Interviewees were chosen if they had used the Atlas or, if nobody in the organisation had used it, based on their job roles relevant to using such data. Both users and nonusers of the Atlas were interviewed as representatives of their organisations. If they were unsure whether others had used the NHS Atlas they asked other colleagues if they had. If at least one person reported using the NHS Atlas, the PCT was recorded as a "user". A working definition of "use" of the Atlas was that PCT staff reported some form of engagement with the material. Before the interviews, permission for tape-recording was obtained. In total, 45 interviews with senior to mid-level executives involved in public health,

commissioning and knowledge management from 29 PCTs were undertaken face-to-face or via telephone between October 2011 and March 2012. The interviews were transcribed verbatim and, guided by the conceptual framework, reviewed iteratively with the survey results to identify and confirm emergent themes. Themes were analysed using the Framework approach (Ritchie and Spencer, 1994), a matrix based method to construct and organise an index of central themes and subthemes, and thereby facilitate a synthesis of the findings by theme and by respondent. The recruitment of interviewees was stopped when a stage of saturation was reached; that is when no new themes emerged after several further interviews (Robson, 2002).

2.3 Results

2.3.1 Characteristics of respondents

Table 2-1 shows the characteristics of the responding PCTs for Waves I and II, overall, and in comparison to all 151 PCTs. Responding PCTs were significantly larger in terms of median population size than the totality of PCTs. There was no significant difference between responders and all PCTs in terms of the Index of Multiple Deprivation and rurality. Overall, the survey achieved representation of PCTs from all five major geographic Strategic Health Authority (SHA) regions in which PCTs are situated. However, PCTs from London were underrepresented while PCTs from Midlands and East were overrepresented.

Table 2-1 Response rates and sample characteristics (2010 data)

	Wave I (July 2011–Aug 2012)	Wave II (Oct 2011–March 2012)	TOTAL SAMPLE	TOTAL PCT POPULATION
Number of PCTs	17	36	53	151
Median population size (range)	362,345 (100,843 - 1,114,366)	389,456 (179,344 - 1,296,814)	378,907* (100,843 - 1,296,814)	284,000 (91,304 - 1,296,814)
Index of Multiple Deprivation (mean score; range)	20.53 (11.34 - 41.13)	22.75 (8.81 - 41.01)	21.81 (8.81 - 41.13)	23.64 (8.81 - 45.31)
Rural PCT	3 (17%)	13 (36%)	16 (30%)	50 (33%)
Predominantly rural[1]	*1 (6%)*	*10 (28%)*	*11 (21%)*	*24 (16%)*
Significant rural[2]	*2 (11%)*	*3 (8%)*	*5 (9%)*	*26 (17%)*
Geographic location				
London	*3*	*2*	*5 (9%)**	*31 (21%)*
North of England	*4*	*6*	*10 (19%)*	*36 (24%)*
Yorkshire and the Humber	*1*	*4*	*5 (9%)*	*15 (10%)*
Midlands and East	*7*	*13*	*20 (38%)**	*39 (26%)*
South of England	*2*	*11*	*13 (25%)*	*30 (20%)*

* Significant difference in means at 95% confidence level based on one sample mean comparison t-test (population size) and, respectively, the one sample test of proportions (rurality and geographic location).

2.3.2 Prerequisites for using the NHS Atlas

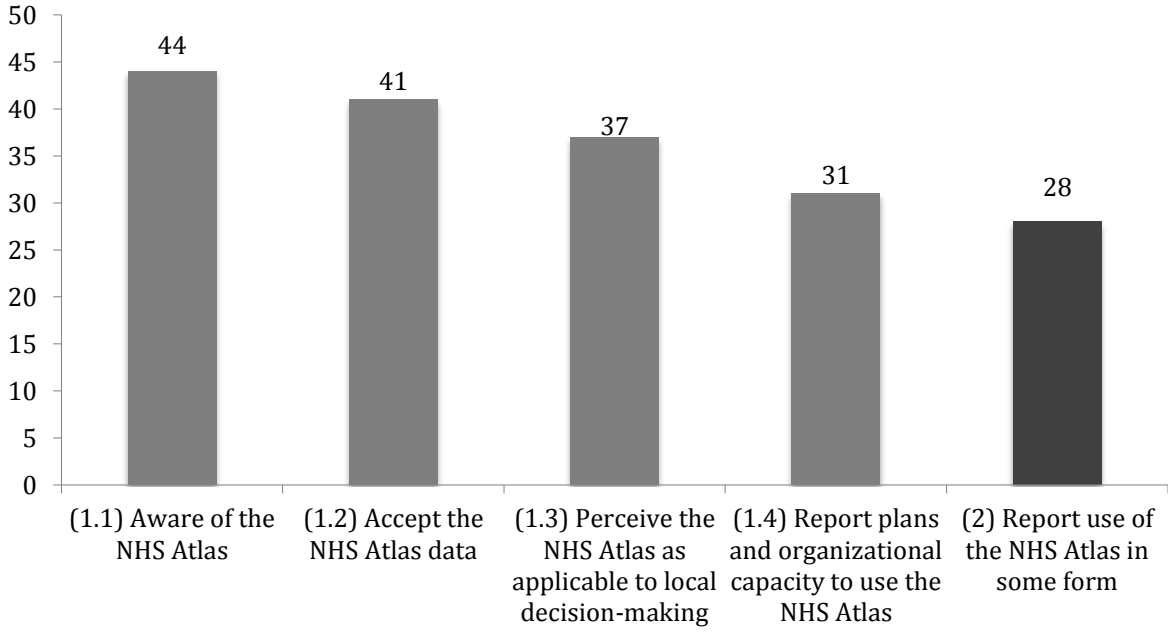
PCTs can be classified into four groups of "non-users" (groups 1.1–1.4), according to the account they gave for not using the NHS Atlas, and "users" (group 2). As the survey results (Figure 2-2) suggest, the number of PCTs appears to decline along these stages from awareness to actual use. Emerging themes from the qualitative analysis (Table 2-1) point to possible underlying reasons, as reported by PCT staff. Most PCTs were aware of

[1] Definition: more than or equal to 50% of their population in rural settlements and larger market towns.
[2] Definition: more than or equal to 26% but less than 50% of their population in rural settlements and larger market towns.

the NHS Atlas (44 of 53 PCTs, group 1.1). Those who had not been aware of the Atlas, despite it being distributed to all PCTs and the relatively large media echo following its publication, referred to being distracted by the structural reorganisation which reduced their attention to information about healthcare delivery.

Group 1.2 was aware of and accepted the NHS Atlas data as generally valid and reliable, although several respondents cautioned about taking the data at face value. In contrast, staff in three PCTs perceived these regional comparisons not as credible due to differences in local management processes, for example in coding patterns, and some noted their preference to work with local data. All PCT respondents recognised unwarranted practice variations as a challenge. This challenge was frequently linked to the NHS-wide economic constraints and the need to meet rising demand with fewer resources. However, only 37 PCTs (group 1.3) perceived the Atlas as applicable to their local situation. The main reasons for perceived limited applicability were the difficulty of (i) inferring from observed variations what ought to be done along care pathways and (ii) discerning the relationship between relative rates of activity and absolute scale of impact on population health outcomes and total service expenditure. Six PCTs who viewed the NHS Atlas as applicable to local decision making noted organisational constraints to use. In particular, annual priorities for action had already been agreed prior to publication of the Atlas and PCTs lacked staff capacity to tackle new issues. Among 31 PCTs (group 1.4) who reported the capacity for using the Atlas, three PCTs had only recently been able to make this capacity available. These PCTs were planning to use the second NHS Atlas published in December 2011. Overall, at the time of study, just over half of the respondents (28 of 53 PCTs, group 2) had thus translated the perceived need to tackle regional variations into actual use of the NHS Atlas.

Figure 2-2 Survey responses to the NHS Atlas



n= 53 Primary Care Trusts.

2.3.3 Using the NHS Atlas in local decision making

Among the users (group 2; 28 of 53 PCTs), a first basic response to the NHS Atlas was to review all maps in order to gain an overview over the PCT's relative position across a range of indicators. PCT staff seemed predominantly concerned to understand where they were "outliers"; indicators on which the PCT was in the highest or lowest quintile of rate of expenditure, activity or outcome relative to the national average. Qualitative themes on uses of the NHS Atlas in local decision making, and factors complicating and enabling its use, are illustrated in Table 2-1 and explained in more detail below.

The initial interpretation of outlier positions tended to be indicative rather than prescriptive. As respondents noted, the outliers shown in the NHS Atlas helped them to identify areas to focus on in their local health economy. Several interviewees referred to

the concept of triangulation inasmuch as a view on variation complemented various other national and local sources of data (e.g. workforce, financial, activity and outcome data insofar as it was available). In their entirety, these multiple pieces of evidence could then help to frame strategic challenges for the PCT. As public health staff in twelve PCTs pointed out, the NHS Atlas supported learning about strategic problems both internally and externally with clinicians. While the Atlas sometimes confirmed existing local suspicions rather than providing new information to PCT staff, map-based visualisations did help to communicate this understanding to clinicians who were not familiar with the statistical data, thus placing it on the management agenda. Messages from the NHS Atlas were then locally disseminated through newsletters, the Annual Public Health report, integration into evidence-into-practice packages or presentations to clinicians.

Beyond the description and illustration of variations, the evaluation of what were perceived as unwarranted variations appeared to be painstaking. As interviewees explained, they attempted to draw as much as possible on existing outcomes research and cost-effectiveness guidance. Further indications of unwarranted variations related to perverse incentives induced by payment systems, and hospital admissions perceived to be avoidable with timely diagnosis and treatment in primary care. For most PCTs, a position in the highest or lowest quintile served as but one indication of unwarranted variation that was further explored with other data sources. In turn, however, many PCTs associated a position in the medium quintile with a lower priority for any action. In some PCTs, this was because a position around the national average was, implicitly, equated with an appropriate rate of activity. These PCTs appeared to take the NHS Atlas at face value rather than as a prompt for further investigation. In other PCTs, in contrast, respondents conceded that limited staff capacity prevented them from exploring all possible sources of unwarranted variation. These respondents pointed out that, although a position in the medium quintile might not be optimal, they had decided to start exploring areas where they were outliers, relative to peers, because these areas might provide larger opportunities to reveal wasteful spending or underinvestment. While PCT respondents confirmed the difficulty of defining and identifying unwarranted variation, they also pointed out that this challenge had to be considered within the wider problem

of where they should start in improving resource allocation by investing limited funds more wisely in order to improve outcomes.

Agreements on responsibilities for action appeared to be decisive in using variations data for local decision making. For the few target-like indicators in the NHS Atlas, where existing clinical guidance would stipulate preferably high values, six PCTs emphasised the importance of involving clinicians at an early stage, as they would ultimately allocate healthcare resources. In two PCTs, for example, maps of variation showing less than 30% of patients with diabetes had received nine key care processes, as opposed to over 70% in the best PCT, helped to convince general practitioners that not only performance was unacceptably poor, in relative and absolute terms, but also that improvements were possible. PCT staff perceived the NHS Atlas as a "catalyst which motivated clinicians to take action sooner than they might have done otherwise" (Director of Commissioning, PCT22).

Among the 28 PCTs where staff had reviewed the Atlas, 18 engaged in further in-depth analysis of possible causes underlying variation. An essential factor appeared to be leadership; both in terms of support from the executive management and local champions from the PCT and clinicians who took the analyses forward. The development of structures to use data on variations also appeared to be important. Some PCTs noted the increasing role of Priority Forums to engage multiple stakeholders in order to improve value, in terms of the relationship between expenditure and health outcomes, in resource allocation. At an operational level, these PCTs had also established regular meetings with providers from primary and secondary care, in order to agree local objectives for action and foster continuous monitoring and feedback against these objectives at hospital or practice levels. In contrast, in PCTs which did not report further action on the observed variations, interviewees also frequently noted a lack of Executive and Board level support, public health and analytical capacity to address the observed variations.

Table 2-2 exemplifies some of the different logics for moving from information on variations to in-depth analysis and decisions about resource allocation. An approach to understanding variations in high-level aggregate indicators, such as total spending on a disease area as in PCT A, was to break down the data into the underlying procedures and settings of care. The objective was to identify the specific drivers of expenditure in a local health economy. Understanding variations in activity involved the exploration of specific hypotheses regarding commissioning policies and supplier behaviour, as in PCTs B and C. Depending on the particular causes identified as underlying variations in practice, PCTs decided whether changes in planning, contracting or service design would be necessary.

Table 2-2 Qualitative responses to the NHS Atlas

Theme	Sub-theme	Example/illustration
1.1. Awareness of the data	Distraction due to organisational reforms	"The development of CCGs [Clinical Commissioning Groups, successors of PCTs as from April 2013] left little room for anything else, such as improving services . . . we were mainly concerned with getting the new structures going" (Chief Operating Officer, PCT4)
1.2. Acceptance of the data	Local management processes seen as too different	"If you look at geographic differences in spending patterns, there may be distortions, in the ways costs are allocated . . . for example PCT spending on cancer may differ depending on the ways hospice costs are taken into account" (Director of Public Health, PCT7)
	Preference to work with local data	"I prefer to work with raw and more detailed local data, for many reasons. . . the data in the Atlas has been transformed and aggregated, which makes it sometimes difficult to understand what is in, and what is out . . . surely you can look up some of these issues in the meta-data [a file published by Right Care detailing the data sources and calculations of Atlas data] . . . but there is also the time lag of 1-2 years in the Atlas data, which is understandable as it takes time to do an Atlas, but at local level we have moved on since then, and have more recent data in some areas" (Information Analyst, PCT14)
1.3. Perceived applicability of the data	Single indicators versus pathways of care	"The Atlas is rather narrow in its focus on single indicators . . . what does this mean for the entire pathway, from community, primary to hospital care . . . is this variation in a single indicator actually meaningful, what does it mean for the pathway?" (Public Health Analyst, PCT3)
	Other criteria besides the magnitude of variation	"Looking at variations only can be misleading if you want to improve services. There may be large scope for improvement even for those in the top quintile nationally. Then of course some areas are simply too difficult to improve. So it's not just about reducing variations but about where to start if you want to improve population health" (Director of Public Health, PCT6)
		"What I want to know is: where do we have the largest potential for efficiency savings, that don't harm patients . . . the Atlas alone can't tell me that" (Financial Director, PCT12)

Theme	Sub-theme	Example/illustration
1.4. Ability to use the data	No staff capacity to use NHS Atlas	"We had already agreed priorities for action when the Atlas was published, and had no further resources and analysts to tackle new issues" (Medical Adviser, PCT9)
2. Use of the NHS Atlas	Strategic problem framing	"Surely the Atlas alone is not enough but we use it to triangulate with other evidence. This helps us to see where we have most potential to improve, mainly financially" (Head of Performance, PCT5)
	Problem communication	"The maps often confirmed our existing local suspicions. But they helped a lot to illustrate to GPs [general practitioners] where we stand compared to other PCTs" (Public Health Analyst, PCT13)
		"We used the Atlas to visualise problems to clinicians, in an accessible format . . . this in turn served as a catalyst which motivated clinicians to take action sooner than they might have done otherwise" (Director of Commissioning, PCT1)
Challenges in using the NHS Atlas	Unclear basis for evaluating unwarranted variation	"There is not always a clear-cut definition what variation is bad... usually we take NICE [National Institute for Health and Care Excellence] guidance as a basis, if it is available for this area" (Public Health Analyst, PCT1)
		"Variation is "unwarranted" for us if we could have avoided it with better organisation of the service, or better provider payment... but my concern is that we don't always know what better payment or delivery should look like" (Commissioning staff, PCT2)
	Role of the national average as an implicit reference point	"We were in the middle for most indicators . . . so nothing alarming really" (Medical Adviser, PCT24)
		"It's difficult to know where to start . . . we also don't have the resources to do everything. So we mainly looked at areas where we were large outliers . . . if you are very different from others, it's likely that something goes wrong in your PCT. But for respiratory disease we are around the national average for most indicators in the Atlas and still I think we could improve a lot" (Public Health Analyst, PCT2)

Theme	Sub-theme	Example/illustration
Enabling factors for coordinating further analysis and action	(Internal) responsibilities for action: Management structures and clinical involvement	"We have regular performance management meetings together with local clinicians to agree service objectives, and who does what . . . and then we monitor progress towards these objectives. The Atlas fit in naturally into our existing structures" (Director of Commissioning, PCT16)
		"It's key to have some structures to get local clinicians on board, to have a team that visits the practices, talks to clinicians . . . asking them regularly about variations and why this local health economy might differ from others" (Director of Commissioning, PCT25)
	Leadership and high-level support	"The PCT Board gave great support in using the Atlas . . . they discussed the Atlas at one of the Board meetings, and appointed a person to champion work into variations" (Public Health Analyst, PCT21)

Table 2-3 Case studies

	PCT A	PCT B	PCT C
Data from the NHS Atlas	PCT A was in the highest national quintile for total spending on cancer care	PCT B was in the highest national quintile for rates of cataract surgery	PCT C was in the highest national quintile for magnetic resonance imaging [MRI] activity
Evaluating unwarranted variation and its causes	NHS Atlas data was disaggregated using data from the regional Quality Observatory: from total spending at regional level to patterns of spending across procedures and across settings of care The cancer care team identified two main drivers of unwarranted variation: 1. Multiple charging for treatment events due to four separate charges for chemotherapy 2. High levels of emergency admissions both at active treatment stage and at the end of life	Comparisons with neighbouring PCTs showed a lower clinical threshold for cataract surgery in PCT B (6/12 versus 6/9 in the worse eye) Reasoning about unwarranted variations was based on two main observations: 1. The current clinical threshold was at the lower end of the driving standard set by the Driver and Vehicle Licensing Agency (between 6/9 and 6/12) 2. A large national audit had shown that one in three eyes with a pre-operative visual acuity of 6/9 either had no benefit or a poorer outcome post-operatively. In eyes with a pre-operative visual acuity of 6/12, only one in eight did not improve	In one of the regular performance management meetings between PCT staff and hospital medical and operating managers, clinician discretion was identified as a likely driver of variation. A retrospective audit was undertaken to compare clinical guideline recommendations with actual practice. The audit showed clinicians complied with current guidance in prompting the provision of MRIs
Responsibilities for action	Monitoring by the PCT and regular performance meetings between the Director of Commissioning and local physicians	Review by the PCT's public health team as a basis for review by the PCT's Priorities Forum	Joint leadership by the PCT's commissioning team, the medical director and operating officer of the acute hospital

	PCT A	PCT B	PCT C
Analysis and decisions on actions	Cancer-care specific decisions included: 1. The revision of contracts to ensure appropriate payment 2. Commissioning of new community services including Palliative Care Co-ordination and Rapid Response Teams to decrease the burden on hospital emergency facilities	The Priorities Forum (which advises the PCT on the treatments that should be given high or low priority and comprises public health and commissioning staff, primary and secondary care representatives, a lay representative and a librarian) agreed: 1. To increase the clinical threshold for cataract surgery to the 6/12 level 2. To introduce special clauses for occupations in which small gains in binocular visual acuity can be essential to the ability to work (e.g. watchmakers, microsurgeons) to prevent inequities	Current practice and relatively high rates of MRI utilisation were considered to be appropriate

2.4 Discussion

2.4.1 General comments and impact of this study

Internationally, there is a growing policy interest in and information on geographic variations in healthcare. In a rising number of countries including Canada, England, Germany, Spain, the Netherlands, New Zealand and the United States, Atlases of Variation have either been or are being developed to raise awareness of regional differences in patterns of expenditure, activity and outcomes. But although healthcare payers have unprecedented access to variations data, how to use such information to improve decisions about the value of resource allocation remains little understood. Even the Dartmouth Atlas of Health Care (The Dartmouth Institute, 2012), the oldest Atlas of Variation published on a regular basis since 1995, has not been examined in terms of its impact on healthcare decision-making. Although data from the Atlas was

allegedly used to underpin political statements in recent healthcare reform debates (White, 2011), the routes through which this information might influence decision-making has received less attention. This lack of systematic impact analysis seems surprising given the importance afforded to information on variations.

The findings of this study suggest some general lessons for using Atlases of Variation. Detailed recommendations for policy-makers, in terms of further development of the NHS Atlas, and for managers and users of the information ("5 questions to ask yourself when looking at the NHS Atlas") are provided in Appendix 2-C and 2-D, respectively. These recommendations were developed by the PhD author on behalf of NHS Right Care. They were used to inform further production and dissemination of future Atlases (NHS Right Care, 2011). While some of these recommendations will be specific to the English NHS, most will have generic relevance also to other countries.

2.4.2 Three lessons for using information on variations

Below we emphasise three general lessons that have emerged from the findings of this study. First, publishing an Atlas of Variation may have great merit in stimulating the search for and understanding of variations, but it may not be sufficient for achieving an impact on decision-making about resource allocation. Generic hurdles to using research evidence – such as awareness, acceptance and perceived applicability of the data (Glasziou and Haynes, 2005, Nutley et al., 2007) – also appear to be relevant for geographic variations research. Once these barriers have been overcome, it appears that Atlases of Variation can serve as a "tin opener" (Carter, 1989) to inform strategic planning by healthcare payers. They may also help communicate strategic problems to clinicians.

However, additional factors appear to be necessary for moving beyond an initial stage of gathering and communicating data towards subsequent stages of the decision-making process where data are analysed and action is taken. On the one hand, decision-makers will have to achieve some clarity and consistency on the definition and operationalisation of the concept of unwarranted variation. The current paucity of corresponding scientific frameworks identified in a recent systematic review

(Mercuri and Gafni, 2011) argues this challenge. On the other hand, agreements on responsibilities for action and leadership also appear to influence the uptake of variations data. Although all 53 participants in this study emphasised addressing unwarranted practice variations as an opportunity to reduce inappropriate use of resources within increasingly tight economic constraints, only 18 of 28 PCTs who had reviewed the Atlas were also able to coordinate further analysis and action. This is a missed opportunity.

Second, who should lead in identifying and acting on variations in medical practice, and how other stakeholders should be involved, is increasingly becoming an issue as the public availability of geographic variations data continues to grow. The NHS Atlas mainly addresses commissioners and clinicians. Given the regionalised planning and purchasing structure, this perspective seems relatively straightforward for England, as the level of analysis – the Primary Care Trust – is thus consistent with the locus of responsibility for action. In countries with competitive social health insurance systems, in contrast, a regional level of analysis tends to conflict with more dispersed responsibilities for action. In Germany, for instance, no institutionalised bodies exist to exercise cross-sectorial planning and purchasing for geographically defined populations (Ettelt et al., 2012). While the NHS Atlas is mainly targeted at health service professionals, a recently published German Atlas of Variation seeks to create pressure for change by targeting citizens and the wider public (Nolting et al., 2011). Further research might examine how a given health system context shapes the uses and users of data on variation in health service performance, and the respective interactions between stakeholder groups in identifying and addressing unexplained variations.

Third, the findings also illustrate the difficult relationship between relative rates of service provision and treatment appropriateness. The purpose of an Atlas of Variation is to reveal variations, and among the respondents to this study, attention logically tended to focus on the top and bottom "outliers". The downside of stimulating action based on outliers was some indication of false assurance derived from an average position. However, research does not suggest a systematic relationship between high, average and low rates of activity and rates of

inappropriate utilisation at a regional level (Leape et al., 1990, Chassin et al., 1987, Keyhani et al., 2012). Simulation studies also suggest that considerable variations at lower provider levels of analysis may in some cases be averaged out at a higher regional level of analysis (Diehr et al., 1990). While an outlier position can be a powerful trigger for further scrutiny, healthcare payers thus need to be wary of not conceiving the national average as an implicit reference point or even target; the danger is complacency.

To prevent an overemphasis on individual outliers, future research may need to move from the measurement of single indicators towards a more systemic view of variation and its management. This may include not only the linkage of all three domains of quality of care – structure, process and health outcomes (e.g. Donabedian, 1978, 1988) – but also a "value for money" framework which relates the outcomes achieved to the resources deployed (National Audit Office, 2011). Possible starting points may be the modelling of patients' pathways across all settings of care (Porter, 2010, Porter and Teisberg, 2006) and, at a population level, the modelling of population health gain from implementing alternative interventions in relation to the required expenditure (Airoldi et al., 2014). Future research may need to focus more strongly on developing requisite models and designing them in such a way that they can easily be applied by health service professionals.

2.4.3 Limitations

This study was constrained by two main classes of limitations; those inherent to qualitative research, and those specific to this study. Interview-based research is well-suited to explore personal experiences and perceptions known only to the people involved (Patton, 1990). However, potential inaccuracies may arise due to poor recall and misrepresentation of facts, when respondents give answers they assume the interviewer wants to hear (Robson, 2002). Interviews with multiple respondents per PCT, if possible, and an emphasis on the open-ended, non-directive character of the interview questions were intended to address these challenges.

A study-specific challenge was the potential for selection bias. It remains unclear whether the non-respondents to this study lacked the capacity to participate in the research, in light of the large scale structural reorganisation of the NHS at the time of study, or whether they were not interested in the topic of variations in healthcare. Despite the wide spectrum of responses to the NHS Atlas illustrated in this study, the respondents may have been more motivated or even pioneers in engaging with geographic variations data compared to their peers. PCTs who reported using the NHS Atlas also tended to be of a larger size (responsible for median populations of about 378,907, compared with the national median size of 284,000 people; Office for National Statistics (2011)) or tended to be collaborating with a University. Presumably these PCTs thus had access to greater analytic capacity than the "average" PCT. This issue deserves greater attention in future research, considering that the new Clinical Commissioning Groups (CCGs) introduced by the Healthcare Reform Bill are will involve even smaller populations and thus possibly lead to even lower capacity for data analysis and strategic planning.

Overall, the survey achieved geographic coverage across all regions in England. The underrepresentation of PCTs in London may, according to the survey respondents, be due in part to the additional pressure perceived by PCTs situated in the capital in the context of healthcare reform. In contrast, the overrepresentation of PCTs in Midlands and East is likely due to the fact that one PCT (Lincolnshire) appeared particularly motivated to use information from the NHS Atlas and disseminated positive experiences to neighbouring PCTs. This issue merits further research so as to understand the routes through which good practice might diffuse across geographic areas.

2.5 Conclusions

Based on a case study from England, we have explored key considerations and challenges along the process of moving from data on geographic variations in medical practice towards decisions to improve the value of resource allocation. Explicit

attention to these and other factors may help governments and payers understand the pathways through which this information might inform decision-making. Our findings illustrate that an Atlas of Variation can support healthcare payers in framing, communicating and prompting the search for strategic problems, but that its mere publication may not be sufficient to influence decision-making even in an ideal context where responsibilities for planning and purchasing health services across sectors are integrated in one regional organisation. The provision of appropriate tools to help planners understand what variation is unwarranted, and to prioritise remedial actions on the basis of their contribution to population health, should be a key focus for promulgators of variations data.

2.6 Acknowledgements

This research project received financial support from the NHS QIPP Right Care Programme, responsible for developing the NHS Atlas of Variation. Phil DaSilva is Co-Director of the NHS QIPP Right Care Programme and Joint-Author of the NHS Atlas of Variation. We are very grateful to all respondents from PCTs who invested their time and provided insight and judgement in a time of organisational turmoil in the NHS. We would also like to thank two reviewers for their helpful comments. The usual disclaimers apply.

2.7 Commentary in relation to the organising matrix of strategies to address ambiguity about the standard for evaluation

This Chapter has explored how policy-makers and managers might manage ambiguity about the standard for evaluation with a socio-political approach. The Chapter has provided a model to frame the process of moving from the measurement of variations in healthcare to their management. On this basis, the Chapter has investigated barriers along this process faced by healthcare payers in England and examines strategies to make sense of the information locally.

The findings demonstrate that information on variations can serve as a "tin opener" (Carter, 1989) to motivate further analysis and inform strategic planning even in the absence of an a priori standard of "good" and "bad" performance. To achieve this, however, what is additionally required is leadership on variations and the provision of appropriate tools to understand which variations are unwarranted.

2.8 Appendix

2.8.1 Appendix 2-A. Survey questions: LSE/ Right Care study on the NHS Atlas of Variation in Healthcare

Dear Chief Executive

As you are aware there is an increasing focus on dealing with variation in the NHS and in particular eliminating unwarranted variations.

The first step to this is the understanding of variation as an issue for patients, clinicians and managers all coming at it a different way. The NHS Atlas of Variation has been successful in highlighting variation.

The Right Care workstream has commissioned LSE to examine how local commissioners use the NHS Atlas. The aim is to understand how the NHS Atlas might be applied, as a tool for change, within a local health economy, for example in terms of stimulating discussion with clinicians, promoting further analysis and action. These experiences will be published by the Right Care team to support local health organizations in identifying and addressing unwarranted variation in healthcare.

We would be delighted if your PCT Cluster and forming CCGs participated in this project. This will be an opportunity for sharing ways that your organisation uses the NHS Atlas as a tool in its decision making process relative to its peers. An output of your participation could include us doing a presentation to staff or board and/ or provide a written report for internal use.

We would be grateful if you could take the time to answer the following questions or if you could ask relevant colleagues to answer them:

1. Are you aware of the NHS Atlas of Variation in Health Care published in November 2010 as part of the QIPP Right Care Programme?

2. Has your Board considered the findings of the NHS Atlas of Variation?

3. Have you or your colleagues used the information provided by the NHS Atlas in some form?

4. If nobody in your organisation has used the Atlas, could you describe why not?

5. Do you use the NHS Atlas? If so, could you briefly state how you use it?

6. Could you briefly state in what form the publication of the NHS Atlas has stimulated new action, and/or supports existing work? For instance, does the NHS Atlas support your work related to

- Strategic planning and evaluation
- Contracting of providers
- Engagement of clinicians, patients or the wider public?

7. How do you think could the NHS Atlas be developed to be more useful for your organisation? How could the NHS Atlas be improved?

8. Would you be willing to be interviewed to share your experiences about barriers and enablers of using the NHS Atlas of Variation? If yes, please could you provide your name and email address or telephone number?

Individual interviewees could of course remain anonymous if they wish. Your participation is valued and we would highly appreciate to hear from you as soon as possible.

Please contact Laura Schang under L.K.Schang@lse.ac.uk or 07586250538. We would also be very grateful if you could notify colleagues or indicate potential interview partners for us to contact.

2.8.2 Appendix 2-B. Interview guide

Your experiences in using the NHS Atlas of Variation are invaluable in enabling learning across Primary Care Trust Clusters and forming Clinical Commissioning Groups. Our aims are to use your experiences to improve outcomes for patients and to help develop future versions of the NHS Atlas.

The Right Care work stream has commissioned LSE to develop case studies of how different stakeholders use the NHS Atlas. The aim is to understand how the NHS Atlas might be applied, as a tool for change, within a local health economy, for example in terms of stimulating discussion with clinicians, promoting further analysis and action. These case studies will be published by the Right Care team to support local health organizations in identifying and addressing unwarranted variation in healthcare.

We would like to discuss with you

- **<u>Looking backward</u> –The situation or problem:**
    - What issues did you encounter in 2011 and in previous years which made it important or helpful to use the NHS Atlas of Variation and other comparative information?
    - Were these special concerns and/or linked to a particular stage in the commissioning cycle?

- **What action was taken:**
    - How did you go about identifying what variation 'matters' in your region?
    - What tools and processes you use to explain medical practice variation, and
    - How did you define whether these causes were 'warranted' or 'unwarranted'?
    - What did you do about unwarranted variation?

- **Stakeholders:** How were/ are different stakeholders involved in the process of identifying and acting on (unwarranted) practice variations? Such as
    - Public health professionals?
    - Clinical Commissioning Group/ commissioning leaders?
    - GP?
    - Hospital clinical and managerial teams?
    - Other stakeholders?

- **What happened as a result:**

    - what expected and unexpected changes the NHS Atlas (and other sources of evidence) triggered or supported in terms of
        - awareness of and attitudes towards unwarranted variation among clinicians/ providers,
        - changes in commissioning and/or public health policies, service design and clinical practice
        - quantifiable results (e.g. service volume, referral patterns, patient outcomes, savings)

- **Enablers and barriers:**
    - What levers enabled you in using information on variations effectively?
    - What facilitating factors you experience(d) in identifying and acting on (unwarranted) practice variation and how you attempted to overcome them?

- **Looking forward:** How do you intend to use the NHS Atlas 2.0 and other information on variations in 2012 in the future?
    - For example, in what stage(s) of the commissioning cycle?
    - Might there be a review of the Atlas requested by the Board, etc?
    - How do you decide on the variations (e.g. disease areas, primary or secondary care) to be addressed first, and why?
    - Will there be a review/ policy for the entire PCT Cluster, and/or will you look at lower levels and how?

- **Might you be able to share graphs/ diagrams,** such as charts used for planning and commissioning and in working with clinicians, to illustrate your work on unwarranted variations?

Contact: To facilitate learning between PCT Clusters we would also like you to consider us sharing the name of your PCT Cluster and a **contact if possible.**

Additional interviews: Can you suggest additional people at different levels whom we might approach for sharing their experiences, such as the

1. Director of Public Health,
2. Director of Commissioning,
3. Medical Director,
4. Public Health Analysts,
5. GPs/ clinical commissioning group leaders

Thank you so much for your time!

2.8.3 Appendix 2-C. Recommendations for policy-makers: Development of the NHS Atlas

1. ***Include several units of analysis:*** The provision of data at a regional (PCT) level only will be too coarse for local decision-making and service planning. Future versions of the NHS Atlas should therefore enable the disaggregation of data from PCTs to Clinical Commissioning Groups, individual practices and hospitals,
   (i) to enable comparison e.g. of admission rates at different levels of analysis; and
   (ii) to inform the identification of PCT-wide and of provider-specific priorities.

2. ***Revisit the choice and organisation of indicators:*** The core purpose of the NHS Atlas should be clarified. This core purpose should then ideally inform the selection of indicators. While different purposes may not necessarily be mutually exclusive, depending on the core purpose, the criteria for selecting indicators may differ:
   (iii) **The NHS Atlas as a system performance tool:** Indicators should ideally be chosen in relation to the commissioning envelope of PCTs/CCGs. They could be organised according to high-level programme budget categories, and key sub-indicators underneath. Wider coverage of areas and indicators would be important to enable a more comprehensive view of performance.
   (iv) **The NHS Atlas as a tool to identify and realise opportunities for cost savings:** Indicator selection could thus focus on resource-intensive procedures as well as on procedures of high-variation across the country.
   (v) **The NHS Atlas as a tool to track patient pathways:** This would include grouping indicators in a meaningful way, e.g. in terms of patient pathways for long-term conditions would improve the relevance of the data to local decision-making.
   (vi) **The NHS Atlas as a tool to stimulate curiosity:** Indicator selection could thus focus new indicators each year to maintain salience.

3. ***Ensure a transparent and robust methodology:*** This includes giving renewed consideration to the following areas:
   a. ***More detailed meta-data:*** Specify the individual disease codes allocated to programme budget categories to facilitate analysis of apparent outliers.
   b. ***Access to data tables:*** Provide the underlying data as public use files such that possible users can check the data.

c. ***Disease-specific weighting:*** The standardised rates draw on overall age, sex and need weighted populations. However, as the patterns of need, age and sex vary across diseases and different areas of healthcare, a disease-specific weighting may be more appropriate (e.g. per 1,000 people with a cancer diagnosis). For musculo-skeletal conditions, the apparently higher levels of spending in PCTs with older age profiles might be explained by the much stronger correlation of musculo-skeletal conditions with age than this was the case with other programme budget categories.

4. ***Provide analytic tools:*** Being an outlier in terms of a given utilisation rate can affect very few cases in total with therefore little practical significance. Where rates are based on low prevalence this should be demonstrated. Furthermore, the following aspects could be strengthened:
    a. ***Relative scale of benefits:*** Demonstrate the scale of cost savings and improved health outcomes that could be achieved by reaching the top-quintile in empirical terms (compared to the 'best' PCTs) or in theoretical-normative terms (compared to what *could* be achieved).
    b. ***Correlations*****:** Enable further analysis of relationships between different indicators e.g. drug spend versus hospital spend versus community spend.
    c. ***Expanded overviews*** over possible reasons for variation should be provided.
    d. ***Trends:*** A means of tracking change will be important to identify trends over time.
    e. ***Profiles:*** Enable pulling a single file showing the overall position for a PCT (SHA, CCG) in the interactive version, complementing the separate thematic overviews. A download option for charts and improved resolution for screen capture should be provided.

5. ***Facilitate dissemination, learning and action:*** The publication of the NHS Atlas should be linked to the planning cycles of PCTs to enable incorporation into next year's commissioning and provision strategies. In addition, the following aspects should be addressed:
    a. ***Wide and ongoing dissemination:*** Conduct follow-up meetings and events to disseminate the NHS Atlas more widely and remind of its availability.
    b. ***Standard means of posting enquiries into the data*** could help to identify whether PCTs had already looked into their apparent high or outlier position, to demonstrate the reasons identified for this position.
    c. ***Case studies*** of how other organisations have made use of the NHS Atlas information could inform local learning and should be disseminated at a national scale.

2.8.4 Appendix 2-D. Recommendations for managers: "5 questions to ask yourself when looking at the NHS Atlas"

Understanding variations in healthcare can be time-consuming and complex. This complexity should not impede PCT boards to tackle unwarranted variation. Doing nothing is not an option at a time when the NHS has to make unprecedented savings while enhancing value from the budget allocated to healthcare.

PCT boards might ask themselves 5 questions when looking at the NHS Atlas (see list). This might provide a structure to guide commissioners to take action. It might be a starting point to break down the question: what does this indicator mean for what my organisation should do, tomorrow?

A word of caution is required for high and low ("outlier") values in the NHS Atlas. Outlier values may indicate areas where PCTs are falling behind, compared to peer organisations. At the same time, even the good end of the empirical spectrum might offer large scope for improvement. Commissioners should be mindful of not making average values a priority, as the national average is not necessarily the optimal rate of expenditure or activity. Despite this risk the NHS Atlas might help commissioners to ask whether high rates of a particular intervention are beneficial for patients. The resources invested in activity that exceeds the national average might yield larger health gains if they were allocated to better-value care, for example to meet unmet needs in this or in another group of patients.

1. What sort of indicator is this?

(a) **High-level aggregate indicator** e.g. mental health or cancer programme budget spending
(b) **Rate of activity** e.g. hip replacements per 1000 people
(c) **Compliance with effective care standard** e.g. proportion of eligible patients treated in stroke units
(d) **Health outcome** e.g. coronary heart disease mortality (negative), cancer survival rate (positive)

2. Where does my health economy sit on this indicator compared to peers?

(a) **For high-level aggregate indicators,** if ranking is high or low, then break down the indicator into activity included in this indicator (e.g. procedures, prescribing)
(b) **For rates of activity,** if ranking is high or low, then investigate where in your health economy this outlier activity occurs

(c) **For indicators which measure compliance with effective care standards,** if ranking is low, then investigate where in your health economy this outlier activity occurs

(d) **For negative (positive) health outcome indicators,** if ranking is high (low), then investigate where in your health economy this outlier activity occurs

(e) **Otherwise,** comparative analysis does not suggest that this indicator should give cause for concern

3. Where in my health economy does this outlier expenditure, activity or outcome (not) occur?

(a) **For expenditure and activity data,** drill down to provider level (GP practice, hospital, community provider) using e.g. NHS Comparators, Quality Observatory data

(b) **For outcome data,** look at patient sub-groups e.g. by geographic area, socio-economic status (e.g. income, education, occupation) or age group

4. What might explain the variation?

(a) **Demand factors** e.g. patient decisions, GP decisions, illness, commissioning priorities

(b) **Supply factors** e.g. service design, clinical decisions, government policy, resource availability, payment structures

(c) Determinants beyond the health system

(d) Data inaccuracy

5. How might we move towards the right treatment rate if the variation is unwarranted?

(a) **Use a systematic appraisal approach** such as Program Budgeting & Marginal Analysis or Decision Conferencing (http://www.health.org.uk/publications/commissioning-with-the-community/) to prioritise investment

(b) **Review and implement regional and national guidelines** (e.g. NICE, Royal Colleges)

(c) **Inform and engage providers** through discussions, physicians profiling, peer education, clinical practice guidelines, financial incentives

(d) **Inform and engage patients** through decision aids and shared decision-making

for treatments with large trade-offs for patient quality of life or life expectancy

(e) **Network with other commissioners** to exchange best practices e.g. via the Health Investment Network and review policies for commissioning and provision

3 HOW TO HANDLE AMBIGUITY ABOUT WEIGHTS AND CHOICES OF DENOMINATOR IN COMPOSITE MEASURES OF HEALTHCARE QUALITY? ROBUST RANKING INTERVALS AND DOMINANCE RELATIONS FOR SCOTTISH HEALTH BOARDS

Authors:

Laura Schang[a], Yrjänä Hynninen[b], Alec Morton[c], Ahti Salo[b]

Author information:

[a] Department of Management, London School of Economics and Political Science, London, United Kingdom

[b] Department of Mathematics and Systems Analysis, Systems Analysis Laboratory, Aalto University School of Science, Aalto, Finland

[c] Department of Management Science, Strathclyde Business School, University of Strathclyde, Glasgow, United Kingdom

Correspondence to:

Laura Schang
Department of Management
London School of Economics and Political Science
Houghton Street | London | United Kingdom
Email: L.K.Schang@lse.ac.uk

Key words: performance comparison; ranking process; composite indicator; weight.

Abstract

Composite indicators of healthcare quality typically embed contentious assumptions. This includes, in particular, the choice of weights of constituent indicators to obtain a single number. Moreover, although many comparative measures are constructed as ratios, the choice of denominator is often ambiguous. The conventional approach is to determine a single set of weights and to choose a single denominator, although this involves considerable methodological challenges. This study examines an alternative approach to handle ambiguity about weights and choices of denominator in composite indicators which considers all feasible weights and can incorporate multiple denominators. We illustrate this approach with an application to comparative quality assessments of Scottish Health Boards. The results (displayed as ranking intervals and dominance relations) allow one to identify Boards which cannot be ranked, say, worse than 4th or better than 7th. Such rankings give policy-makers a sense of the uncertainty around ranks, and the extent to which action is warranted.

3.1 Introduction

The increasing complexity of health systems and the multidimensionality of health system performance have reinforced calls for the production of composite measures of performance (WHO, 2000, Healthcare Commission, 2005, CMS, 2009, Carinci et al., 2015). Summarizing the information contained in diverse indicators in a single index and ranking organisations or countries on that basis has the potential to present the "big picture", by highlighting in a unified way to what extent the objectives of health systems related to health outcomes, treatment appropriateness, access and other dimensions have been met (WHO, 2000, Smith, 2005). Rather than having to identify a trend across a range of separate indicators, a single number may be easier to interpret and thus offer a rounded evaluation of performance. As such, summary measures may seem an attractive approach to strengthen accountability, facilitate communication with the public and focus improvement efforts on poorly performing organisations (Goddard and Jacobs, 2009, Smith, 2002).

However, there are several arguments against the use of composite indicators. Fundamentally, aggregate measures may disguise the sources of poor performance and thus obscure the best focus for remedial action (Smith, 2002). Rankings based on composite measures are typically also highly sensitive to methodological choices, in particular to the choice of weights attached to constituent indicators (see e.g. Jacobs et al., 2005, Reeves et al., 2007, Gravelle et al., 2003, OECD, 2008). In their analysis of hospital performance based on star ratings in the English NHS, Jacobs et al. (2005) show, for instance, how subtle changes in the weighting system lead some hospitals to jump almost half of the league table. However, the techniques by which weights are determined are unlikely to be straightforward. In addition, although many comparative quality measures are constructed as ratios, it is not necessarily obvious which indicators should be employed as denominators (Schlaud et al., 1998). In the context of low-birthweight survival rates, Guillen et al. (2011) illustrate how the choice of population denominator results in considerable variation depending on whether survival is reported relative to all births; live births; or neonatal intensive care unit admissions.

These concerns are critical especially when rankings have serious consequences for the rankees. For example, six of the Chief Executives of the twelve lowest ranked hospitals in England's star rating system (the so-called "dirty dozen") lost their jobs as a result (Bevan and Hamblin, 2009). It has been argued that France and Spain's apparently high ranking in the WHO's 2000 assessment of health systems substantially diminished pressure for reform in these countries (Navarro, 2000). In Medicare's Premier Hospital Quality Incentive Demonstration, a pay-for-performance scheme based on a composite quality score, hospitals below the ninth decile faced a 2% deduction in their Medicare payment (CMS, 2009). With such high stakes, understanding whether ranks are robust to alternative assumptions seems critical.

We here examine a methodological strategy to handle ambiguity about weights and choices of denominator in composite indicators of performance. We make two main contributions. First, we demonstrate the use of an approach to ranking organisations based on ranking intervals and dominance relations which accounts for the full set of feasible weights. This avoids the need to settle on a single, potentially controversial set of weights as it is required for instance in data envelopment analysis (DEA), in which weights are chosen such that each organisation appears in its best possible light (Cherchye et al., 2007). Feasible weights are less restrictive and thus potentially better able to increase transparency and to acknowledge lack of information about the "correct" set of weights. The ranking intervals obtained with this approach can be said to be robust in the sense that they reflect the full range of rankings that the entities under comparison may attain when weights are selected from their respective feasible weight sets. Second, we highlight the problem of choice of denominator in ratio-based measures of performance and how it might be tackled through the use of ranking intervals. We provide an empirical application to healthcare quality comparisons of Scottish Health Boards.

3.2 Challenges in developing composite indicators of healthcare quality

A composite indicator is commonly expressed as an additive model based on a weighted sum of a set of performance indicators:

$$C_k = \sum_{j=1}^{J} w_j \, x_{jk} \qquad (1)$$

where J is the number of indicators, w_i is the weight attached to indicator j, and x_i the score on indicator j for organisation k. Composite measures of this form require choices about (i) the set of indicators included; (ii) the methods used to transform the constituent indicators (in order to achieve a common unit of measurement); (iii) the weights applied; (iv) any specific aggregation rules used; and (v) potential adjustments for environmental or other uncontrollable influences on performance. In addition (vi), although many healthcare quality indicators that are used to construct a composite indicator are reported as ratios, the choice of denominator is not always straightforward.

The focus of this study is on problems (iii) and (vi), how to handle ambiguity about the choice of weights and the choice of denominator. Below we set out the conceptual background and problems with conventional strategies to handle these problems. In the empirical application, we explain and justify the approaches taken to problems (i), (ii), (iv) and (v).

3.2.1 Valuation of multiple healthcare quality measures

Healthcare performance measures are heterogeneous and multidimensional. However, without a functioning market, there is no price mechanism for comparison. To aggregate different indicators into a summary measure of performance, weights are required which – analogous to prices – should represent the opportunity cost of achieving improvements on each individual measure by capturing the relative value attached to an extra unit of it (Smith, 2002).

In practice, arriving at explicit trade-offs between different healthcare quality measures – and thus exact specifications of weights – is highly contentious. First, it is often unclear *whose* preferences should be elicited. Weights used often reflect a single set of preferences, although the evidence suggests substantial heterogeneity in preferences between and within groups of policy-makers, patients and the public (Smith, 2002, Decancq and Lugo, 2012). Making precise judgments about the relative value of sub-indicators to the composite is typically both politically controversial and cognitively demanding, thus triggering reluctance among respondents to agree on a set of weights.

Second, there is no consensus on a single best method *how* to elicit weights. Different techniques for valuing health(care) outcomes – from simpler trade-off methods including ranking from most to least desired indicator and voting techniques to more elaborate multi-attribute approaches such as conjoint analysis and the analytic hierarchy process – tend to produce different results and each method has distinct advantages and disadvantages in terms of feasibility, consistency and validity (Dolan, 1997, OECD, 2008, Appleby and Mulligan, 2000).

To circumvent perceived difficulties with normative approaches to set weights, data-driven weighting systems are frequently used. For example, data envelopment analysis (DEA) – one of the most widespread methods to compare organisations with multiple outputs and inputs (Hollingsworth and Street, 2006) – uses empirically derived, flexible weights, following a "benefit of the doubt" approach. It is however questionable whether data-driven weights reflect meaningful trade-offs between performance domains (Decancq and Lugo, 2012). There is no logical reason why an organisation necessarily values most some performance domain because it performs relatively well on it: data-driven approaches thus are confronted with the impossibility to derive values from facts (Popper, 1948).

The conventional recommendation to address ambiguity about weights, and the best method to elicit weights, is to conduct extensive sensitivity analysis on the chosen weights (Jacobs et al., 2005). However, traditional sensitivity analysis is problematic insofar as the choice of ranges of weights typically depends on the analyst. This form

of sensitivity analysis thus corresponds to a "blind search" which is not explicitly oriented towards changes in ranks and the maximum and minimum plausible ranks an organisation can attain.

3.2.2 Choice of denominators

Healthcare quality measures are often reported as ratio measures where a specific quality measure is divided by some measure of population. Not all comparative assessments of healthcare quality require necessarily a denominator. So-called "never events", events which are deemed to be entirely preventable, are reported as absolute numbers without reference to a denominator (NHS England, 2015). However, typically a ratio-based measure is used in order to make entities of different sizes comparable and to establish a common currency unit in which performance is assessed as "good" or "poor" relative to other organisations.

To construct ratio-based quality measures, the denominator should represent the best available proxy for the population at risk (Romano et al., 2010, Schlaud et al., 1998). However, the population at risk of experiencing a specific event is not always obvious. Consider two health authorities A and B with the same number of healthcare associated infections (HAIs) but a lower number of bed days in authority A. On a simple ratio measure of HAIs/ 1,000 bed days, authority B would seem to score better, but this conclusion would be warranted only if there were no groups at risk of HAIs other than hospitalised populations. However, if the numerator also included community-acquired infections, then a narrowly defined denominator such as hospital bed days would underestimate the actual number of exposed individuals (in particular, it ignores populations in non-acute hospital settings exposed to HAI e.g. in geriatric wards, nursing homes). A comprehensive denominator, such as total population, in contrast, would overestimate the population at risk by including individuals facing no or a negligible risk of experiencing the event (Marlow, 1995).

In addition, the use of bed days as the denominator may be problematic insofar as it might penalize Boards which succeed in reducing length of stay (another frequent health policy objective). Such Boards would then appear to have poorer performance

on HAIs. Yet, the use of total population as the denominator does not account for Boards with a high number of hospitalised populations. These populations may plausibly have a higher risk of acquiring an HAI than general populations, since they are typically sicker and thus more susceptible to infection.

To address ambiguity about the choice of denominator, it is clearly essential to define the unit of analysis and, on this basis, the correct denominator. For system-level comparisons, population might be appropriate; for hospital comparisons, total admissions or bed days. Ideally, one would therefore specify a numerator that is unambiguously linked to one single denominator (McKibben et al., 2005); for example, by excluding community-acquired infections that are present on admission to hospital from the numerator. In practice, it is however often difficult to distinguish between HAIs that were present on admission and those acquired during a hospital stay (Naessens and Huschka, 2004, Zhan et al., 2007).

Since there will always be some uncertainty about the correct population at risk, it makes sense to consider different denominators since that enables a more complete perspective on the outcome of interest (Guillen et al., 2011). To do this, one could produce multiple ratios between all reasonable numerator and denominator combinations. However, the manual comparison of multiple performance ratios quickly becomes unwieldy. In a situation with, say, four numerators and three denominators, one would obtain 12 performance ratios for each entity under scrutiny.

3.3 Methods

3.3.1 Ranking intervals and dominance relations for all feasible weights

We here examine the use of an alternative approach to handle ambiguity about choices of weights and choices of denominator. Rather than specifying explicit weights (that are subsequently subjected to sensitivity analysis), this approach consists in developing ranking intervals and dominance relations which consider the

full set of feasible weights. The approach is also able to handle different choices of denominator variables.

We adopt a ratio-based efficiency analysis (REA) technique (Salo and Punkka, 2011). Suppose there are *K* Decision-Making Units (DMUs – the entities to be evaluated) that have *N* different measures for the numerator of a ratio and *M* measures for the denominator of a ratio. The values of the *n*th numerator and the *m*th denominator of the *k*th DMU are $y_{nk} \geq 0$ and $x_{mk} \geq 0$, respectively. Thus, the possible performance ratios of the DMU *k* are y_{nk}/x_{mk}, where $n = 1, \dots, N$ and $m = 1, \dots, M$.

REA enables the aggregation of different numerators and denominators in a summary measure of performance. The relative importance of the *n*th numerator and the *m*th denominator is captured by nonnegative weights u_n and v_m, respectively. The aggregated performance ratio of DMU *k* is defined as

$$E_k(u, v) = \frac{\sum_n u_n y_{nk}}{\sum_m v_m x_{mk}}. \quad (2)$$

To examine the pairwise relations between DMUs, REA uses the concept of dominance: DMU k dominates DMU l if the performance ratio of DMU k is at least as high as that of DMU l for all feasible weights and there exist some weights for which its performance ratio is strictly higher. If a dominance relation exists between two DMUs, one can be confident that for any set of assumption, one DMU outperforms the other. The dominance relation between DMUs k and l is determined by the pairwise performance ratio

$$D_{k,l}(u, v) = \frac{E_k(u,v)}{E_l(u,v)}. \quad (3)$$

The maximum and the minimum of $D_{k,l}(u, v)$ over all feasible weights provide upper and lower interval bounds on how well DMU k performs relative to DMU l. Thus, if the minimum of $D_{k,l}$ is greater than one, DMU k dominates DMU l. The dominance structure is computed with linear programming.

The ranking interval indicates the best and worst performance rankings a DMU k can attain relative to other DMUs over all feasible weights. The best ranking is

determined by the minimum number of other DMUs with a strictly higher performance ratio. For instance, the best ranking as third for a given DMU means that, no matter how the weights are selected, there are at least two other DMUs with a strictly higher performance ratio. If for some feasible weights the performance ratio of a DMU is higher than or equal to the ratio of any other DMU, then its best ranking will be one. The worst ranking is computed similarly.

3.3.1 Method strengths and limitations

There are several innovative characteristics, and distinct advantages, to this approach. First, the aggregation of numerators and the denominators is achieved without fixing the weights of constituent indicators. By comparing the relative magnitude of the performance ratios between DMUs with all feasible weights, one can produce robust information about the performance of DMUs in the sense that the resulting intervals reflect the full range of rankings that DMUs may attain for feasible weight sets.

Second, REA calculates pairwise comparisons between DMUs rather than comparing each DMU to an efficient frontier as in DEA or stochastic frontier analysis. This makes REA results more robust than frontier-based results, since the introduction or removal of an outlier DMU can substantially change the location of the efficiency frontier (Banker et al., 1986). Pairwise dominance relations obtained from REA, in contrast, cannot change if a new DMU is added and the ranking intervals can shift by no more than one ranking at the end points.

Third, because the REA uses no efficient frontier, there is no minimum number of DMUs needed to conduct performance comparisons. For DEA, Banker *et al*. (1986) have proposed the simple rule of thumb that the number of DMUs should be at least three times the number of variables. This is problematic since the number of indicators typically far outstrips the number of organisations. REA, in contrast, is based on pairwise comparisons only. It thus requires a minimum of only two DMUs and there is no upper limit to the number of indicators.

It is important to point out that, where the choice of denominator is relatively unambiguous, ratio-based analysis is not necessary. One can calculate individual performance rates for the respective indicators and aggregate them as a weighted sum as in equation (1). This is akin to evaluating the numerator of the performance ratio (2).

In this study, we use a ratio-based analysis in order to illustrate robustness to different choices of denominator. However, it is important to recall that ratio-based measures have possible limitations. In particular, the use of a ratio function assumes constant returns to scale in the sense that it does not account for structural differences (such as a higher share of fixed costs) between organisations. This assumption implies that, in evaluating organisational performance, one does for instance not allow an organisation a comparatively higher number of healthcare-associated infections (in terms of the performance ratio, e.g. per 100,000 population) only because it relatively small in size. However, in the context we examine here – Scottish Health Boards, as outlined below – this assumption seems justified since these Boards are allocated resources in line with a formula which seeks to compensate for structural differences so as to ensure a level playing field across organisations.

Ratio measures may be preferred when there is primarily a concern with evaluation (examining which organisations obtain higher or lower performance ratios) rather than explanation (examining why organisations achieve particular performance outcomes, as in regression analysis). Alternatively, one could identify empirically the population at risk by means of a regression model akin to a production function, where a specific quality measure is analysed as a function of different possible populations at risk. Variables with positive coefficients have an influence on the quality measure in question and could thus be interpreted as populations at risk. The scope of this paper, however, is limited to highlighting the problem of choice of denominator and to examining the implications for comparative evaluations and organizational rankings.

3.3.2 System context and data

Selection of indicators. We here illustrate the robust ranking interval approach with an application to the comparative quality of Scottish Health Boards. In Scotland, responsibility for the allocation of resources is decentralized to 14 territorial Boards. The ultimate objectives of these Boards are to protect and improve the health of their populations through planning for and delivering health services (Scottish Government, 2014). To construct a composite indicator of the quality of care provided by Scottish Health Boards, we confined ourselves to indicators used in the HEAT target system. This existing performance management system is used by the Scottish Government to assess the performance of Health Boards. All indicators used here (Table 3-1) come from the official performance measurement system, but are not meant to represent an exhaustive set of health system objectives. To address the two problems examined in this study, we use two data sets:

- **Data for part I:** To examine robustness to choices of weights and dominance relations, we analyse six indicators from the HEAT target system which are intended to measure Health Boards' relative degree of achievement in ensuring appropriate and accessible treatment. This analysis is based on an additive model which is akin to analyzing the numerator of the performance ratio in equation (2).

- **Data for part II:** For most of the six quality indicators, the correct denominator variable is quite straightforward. However, as discussed in section 3.2.2, for healthcare-associated infections, the choice of denominator may be ambiguous. To examine robustness to alternative choices of denominator (here, the population at risk of experiencing an infection), we relate the number of healthcare-associated infections to hospitalised and general populations. This analysis relies on the more complex ratio-based model in equation (2).

Data transformation. To avoid mixing different units of measurement and to achieve scale invariance, data were normalized to the [0;1] range by dividing each value by the maximum value for a given indicator.

Environmental adjustment. The 14 Health Boards differ in terms of various environmental factors that are beyond the control of Boards but that might influence observed performance on the chosen indicators. Such factors include, in particular, demographic, epidemiological and regional structures. However, in Scotland, such factors should be fully compensated for within the funding mechanism. Health Boards are allocated resources based on a formula that explicitly takes account of variations in healthcare needs that arise as a result of age and sex composition, morbidity, life circumstances and other factors; and excess costs of delivering services in some (especially rural) regions that are deemed unavoidable (ISD Scotland, 2010b). Thus, Boards with older and sicker populations have already been compensated to take account of the greater healthcare needs of their populations so that they can ensure the same level of quality. We acknowledge that the risk adjustment provided by this formula is not perfect. However, following this line of argument, it is not unreasonable to assume that Boards are comparable with respect to the performance indicators analysed here.

Table 3-1 Variables and descriptive statistics

	Definition	Mean	SD	Min	Max
Data for part I: robustness to choices of weights and dominance relations					
18WRTT[a]	Number of patient journeys from referral to treatment over 18 weeks (among patients seen) per 100,000 RTT patient journeys from referral to treatment (among patients seen)	9,858	8,791	1,851	30,603
4-hour A&E waiting[a]	Number of recorded A&E waits lasting over 4 hours per 100,000 A&E attendances	9,412	10,096	859	31,731
Emergency admissions[a]	Number of emergency admissions among +75 years per 100,000 population	13,419	9,274	4,107	37,256
MRSA/MSSA[a]	Number of MRSA/MSSA infections per 100,000 population	137	107	24	413
C.difficile[a]	Number of Clostridium difficile infections per 100,000 population	164	116	42	399
Delayed discharges[a]	Number of bed days lost due to delayed discharges per 100,000 occupied bed days	131	99	13	373
Data for part II: robustness to choices of denominator					
Quality indicator (numerator variable)					
C.difficile[a]	Number of Clostridium difficile infections	164	116	8	399
Population indicators (denominator variables)					
Total population[b]	Resident population (mid-year estimates)	475,232	318,214	113,880	1,214,587
AOBD[a]	Number of acute occupied bed days	113,244	98,182	20,723	365,951

Sources: [a]HEAT target system; [b]National Records of Scotland. All data are for 2012/13.

Table 3-2 Comparative performance of Boards on the constituent six quality indicators, based on rates per 100,000 shown in Table 3-1

		18WRTT	4-hour A&E waiting	Emergency admissions	MRSA/MSSA	C.difficile	Delayed discharges
A	**Ayrshire & Arran**	8,691	8,312	3,646	23	49	3
B	**Borders**	6,204	3,267	3,612	21	44	2
C	**Dumfries & Galloway**	6,170	5,987	3,130	27	36	7
D	**Fife**	6,899	4,559	2,725	35	26	11
E	**Forth Valley**	15,123	8,238	2,513	26	14	9
F	**Grampian**	9,343	3,812	2,239	25	24	10
G	**Greater Glasgow & Clyde**	8,523	6,956	3,061	34	33	5
H	**Highland**	5,817	2,199	2,825	17	24	9
I	**Lanarkshire**	5,551	8,667	2,671	24	35	5
J	**Lothian**	12,293	9,172	2,495	30	42	10
K	**Orkney**	2,649	1,663	2,661	9	84	5
L	**Shetland**	2,209	730	2,555	13	34	9
M	**Tayside**	8,701	1,119	2,964	36	50	5
N	**Western Isles**	4,876	1,666	3,320	4	123	18
	Max	15,123	9,172	3,646	36	123	18
	Min	2,209	730	2,239	4	14	2
	SD	3,475	3,090	424	10	28	4
	Mean	7,361	4,739	2,887	23	44	8

3.3.3 Weight restrictions on quality measures

An advantage of REA is its ability to address incomplete information about weight specifications by using the full set of feasible weights. This can be an attractive option when one assumes complete ignorance about the relative value of averting particular events. However, while an elicitation of cardinal preferences over "how much" worse a, say, MRSA infection is compared to, say, an emergency admission may not feasible (e.g. due to high cognitive demands) or desirable (e.g. due to biases introduced by specific elicitation methods), it may be possible to obtain statements about which events are worse than others. Introducing plausible weight restrictions based on ordinal preferences can be useful because this recognises people's ability to provide limited preference information about the relative badness of particular events without imposing implausibly exact weights. Restrictions on weights can reasonably

be used to prevent inconsistencies with accepted views on the relative importance of measures analysed (Allen et al., 1997, Pedraja-Chaparro et al., 1997).

For illustrative purposes, the research team arrived at a set of ordinal weights through pairwise comparisons of any two quality measures, along the lines "*If you could avoid either an emergency admission to hospital or an MRSA infection, which event would you rather avoid"*. Corresponding to their relative badness, events were ranked as follows (from worst=1 to least bad=6):

1. an MRSA/MSSA infection;
2. an emergency admission[3];
3. a clostridium difficile infection;
4. having to wait longer than 18 weeks from referral to treatment;
5. having to wait more than 4 hours in A&E[4];
6. a delayed discharge.

Another challenge in flexible weighting systems is that the final composite score may be heavily influenced by an indicator that is considered of marginal importance in the wider health system context (Goddard and Jacobs, 2009). We here made the (illustrative but reasonable) assumption that avoiding a particular event can at most have half of the overall value attached to avoiding an event of each of the six quality measures. This resulted in the following proportional weight restrictions: avoiding an event of the worst healthcare quality measure cannot be more than ten times as valuable as avoiding an event of the least bad quality measure (since with six indicators, a ratio of 1/10 means that one quality measure can have at most half of the weight mass).

To examine robustness to different choices of denominator (part II), no weight restrictions were used. In efficiency analysis, denominator weights have a clear interpretation, since they indicate the substitutability between different types of

[3] We assumed an avoidable admission e.g. for acute exacerbation of COPD that could have been prevented with timely primary care.

[4] We assumed a condition where patients are in mild to moderate discomfort.

inputs (labor, capital, other intermediate inputs). In quality comparisons, denominators represent different populations at risk. However, denominator weights lack a clear interpretation as in efficiency analysis since it is hard to think about trade-offs between different populations at risk. To simplify the methodology, the analyses in part II therefore do not use any weight restrictions. In part I, there are no denominator variables since the analysis relies on rates as model variables (see Table 3-1).

3.4 Results

3.4.1 Robustness to choices of weights: Unrestricted and restricted ranking intervals for feasible weight sets

Figures 3-1 to 3-3 illustrate the use of ranking intervals to handle ambiguity about choices of weights assigned to performance on the different indicators. They also show what the most critical assumptions are with respect to weights. The ranking intervals show the possible rankings that Boards can attain. The wide ranking intervals for measures of quality across all feasible weights (Figure 3-1) suggest considerable sensitivity to feasible choices of weights. If one adds ordinal weight restrictions (Figures 3-2) and ordinal and proportional weight restrictions (Figure 3-3), then variations in performance appear to be manifested more clearly.

The width of the ranking interval reflects the impact of different assumptions about changes in weights. A small interval suggests that a Board's performance is robust to alternative modelling assumptions. For example, Board *L* (Figure 3-2) is ranked 3rd or higher no matter which assumptions are used. The interval bounds show the impact of modelling assumptions on relative ranks. Thus, one can be confident that Board *F,* for example, cannot be ranked worse than 7th and not better than 3nd.

Figure 3-1 Performance rankings without weight restrictions

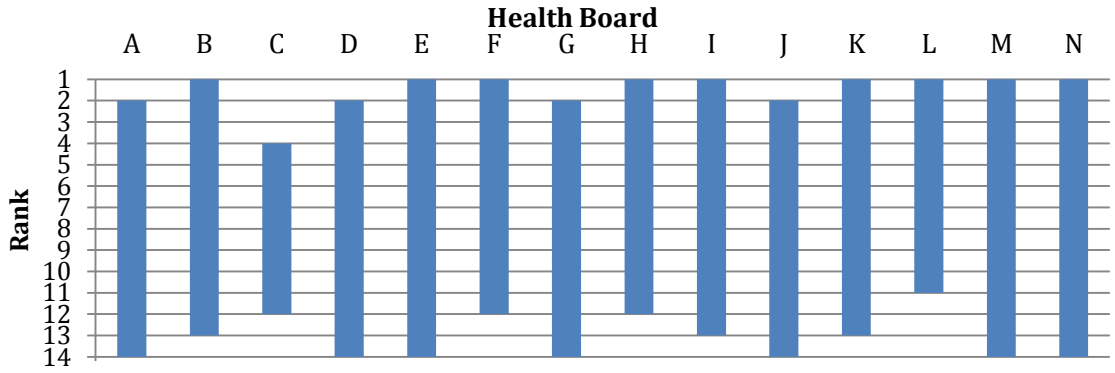


Figure 3-2 Performance rankings with ordinal weight restrictions

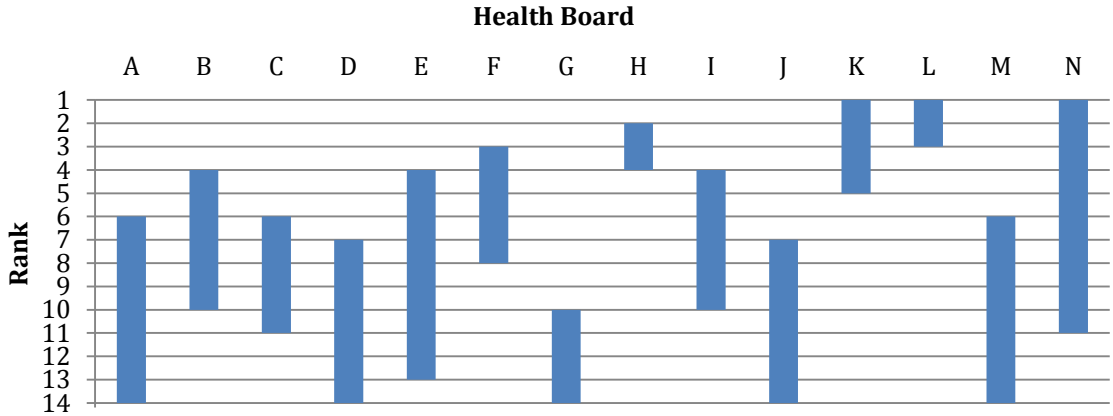
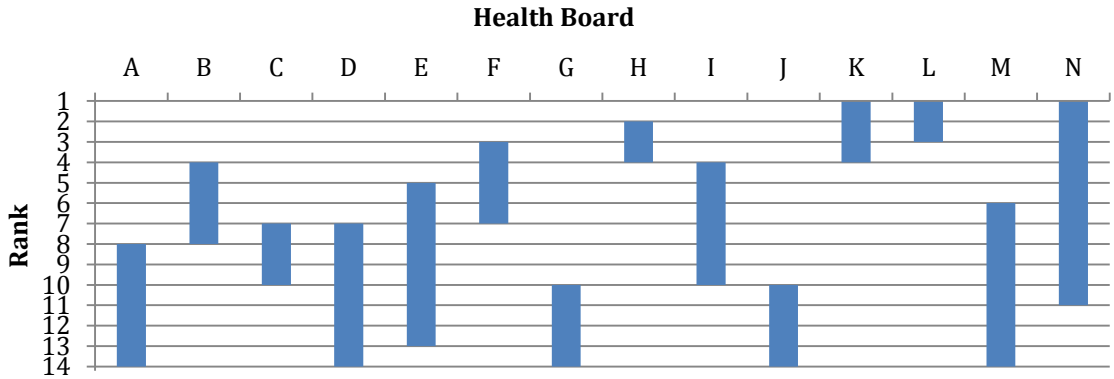


Figure 3-3 Performance rankings with ordinal and proportional weight restrictions



3.4.2 Dominance relations and comparative scope for improvement

Based on pairwise comparisons, the REA results can be displayed in a unified way as a dominance relation (Figure 3-4): insofar as Boards are more superordinate or "higher up", their relative performance is more robust to changes in the weights attached to the constituent indicators of performance. This graph suggests that the three island Boards NHS Orkney (K), Shetland (L) and Western Isles (N) are top performers since they are not dominated by any other Board. In turn, the Boards NHS Ayrshire and Arran (A), Fife (D), Greater Glasgow and Clyde (G), Lothian (J) and Tayside (M) are dominated by the other Boards.

There are two main reasons for this differentiation status. First, a Board's performance on the indicators that are used to construct the composite measure play a role (Table 3-2). For instance, although NHS Western Isles scores worst, by far, on rates of clostridium difficile infections, all three island Boards perform comparatively better than the rest of Scotland on MRSA/MSSA infections, 4-hour A&E waiting times and 18WRTT. Second, the ordinal weight restrictions used influence the dominance relations. In this example, performance on MRSA/ MSSA infections is weighted more highly than performance on emergency admissions, which in turn receives a higher weight than performance on c.difficile, and so on. Indeed, inspection of the underlying data (Table 3-2) suggests that the five Boards who appear at the bottom of the dominance graph perform comparatively worse on MRSA/ MSSA infections and emergency admissions. Nevertheless, their poor overall differentiation status is the result of poor performance on several (up to four different) indicators and thus not exclusively the result of the weighting scheme.

Table 3-3 shows the radial improvements needed for some Board X (depicted by row) that is dominated by some Board Y (depicted by column) to improve its performance (i.e. decrease its rates, since these are all "lower is better" indicators) so as not to be dominated. This means, for instance, that Board A needs to reduce its rates on all the indicators by 8 % so as not to be dominated by Board B. If a cell is empty, then Board X is not dominated by Board Y. Looking horizontally, one can see, for instance, the improvements that would be needed for the five worst performing Boards J, G, D, M, A

to become non-dominated by the better-performing Boards. Looking vertically, one can identify the distance that differentiates each Board from the national leaders, Boards K, L and N.

Figure 3-4 Dominance graph for Scottish Health Boards, based on ordinal and proportional weight restrictions

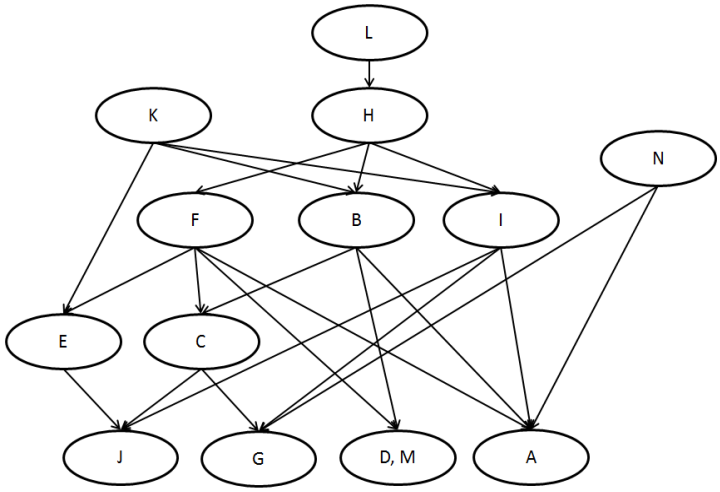


Table 3-3 Comparative scope for improvement needed to reach another target or reference Board in Scotland

Dominated Board		Target or Reference Board												
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**
Ayrshire & Arran	**A**		8 %				2 %		25 %	2 %		22 %	36 %		2 %
Borders	**B**								9 %			14 %	27 %		
Dumfries & Galloway	**C**		<1 %				7 %		21 %			15 %	31 %		
Fife	**D**		3 %				11 %		24 %			17 %	32 %		
Forth Valley	**E**						7 %		12 %			3 %	21 %		
Grampian	**F**								6 %				15 %		
Greater Glasgow & Clyde	**G**		9 %	8 %			16 %		29 %	11 %		22 %	36 %		2 %
Highland	**H**												10 %		
Lanarkshire	**I**								12 %			6 %	23 %		
Lothian	**J**		4 %	2 %		6 %	18 %		23 %	11 %		18 %	33 %		
Orkney	**K**														
Shetland	**L**														
Tayside	**M**		8 %				4 %		20 %			25 %	36 %		
Western Isles	**N**														

3.4.3 Ratio-based analysis: Robustness to choice of denominator

The correlation between rates of clostridium difficile infections per population compared to bed days as the denominator is low (r = 0.15537). This suggests that the choice of denominator will make a difference to the relative ranks Boards may attain. Table 3-4 examines robustness to different choices of denominator. Although most Boards seem to perform either relatively well (Forth Valley, Grampian, Greater Glasgow and Clyde) or relatively poorly (Tayside, Ayrshire and Arran) on either of the two denominators, for some Boards different assumptions about appropriate denominators lead to notable rank reversals. The three island Boards appear to perform distinctly better when infections are measured relative to bed days while NHS Borders, Lanarkshire, Fife and Highland jump up the ranking for a population denominator.

Table 3-4 Performance on healthcare associated infections (HAIs; includes C.difficile) relative to different choices of denominator

Board	Per 100,000 bed days		Per 100,000 Total population		Ranking interval for bed days and population
	Number of HAIs	Rank	Number of HAIs	Rank difference compared to bed days	
Shetland	**55**	**1**	**34**	***-5***	**1-6**
Forth Valley	78	2	14	+1	1-2
Grampian	105	3	24	+1	2-3
Greater Glasgow & Clyde	109	4	33	-1	3-4
Orkney	**114**	**5**	**84**	***-8***	**5-13**
Highland	**124**	**6**	**24**	**+3**	**3-6**
Western Isles	**140**	**7**	**123**	***-7***	**7-14**
Fife	**155**	**8**	**26**	**+4**	**4-8**
Dumfries & Galloway	161	9	36	+1	8-9
Lanarkshire	**162**	**10**	**35**	**+3**	**7-10**
Lothian	177	11	42	+2	9-11
Tayside	195	12	50	0	12
Ayrshire & Arran	211	13	49	+2	11-13
Borders	**241**	**14**	**44**	**+4**	**10-14**

3.5 Discussion

We have focused on two pervasive sources of ambiguity, in the sense of a lack of knowledge about the best modelling choices, which make the use of composite measures for robust performance comparisons in healthcare difficult: How should different indicators be weighted to obtain an aggregate measure of performance? What is the "correct" denominator in ratio-based performance indicators? As Jacobs et al. (2005) note, two possible implications to respond to the uncertainty inherent in composite indicators would be to dismiss composite indicators altogether and instead estimate relative performance separately for each objective (an example of this is Hauck and Street's (2006) multivariate multilevel approach that requires no aggregation and weighting of multiple objectives at all); or to invest considerable resources into more sophisticated modelling, such as by means of elaborate preference elicitation or by seeking to estimate meaningful weights from existing health service information.

In a context where information is inevitably incomplete but policy-makers might still be interested in an overall measure of health system performance (OECD, 2008), we have demonstrated how the REA technology offers a third way that openly provides indications of the uncertainty inherent in the valuation of objectives and choices of denominators. The approach is essentially based on agnosticism: When there are multiple reasonable denominators which each highlight aspects of performance – such as that an organisation can deliver high-quality in terms of few quality measures relative to hospitalised and/or general populations – then analysts need not restrict themselves to a single denominator. Our results reinforce the insight that healthcare quality may be best thought of as a collection of possible rates depending on how the population denominator is specified rather than as a single "right" rate (Guillen et al., 2011). Ranking intervals based on multiple denominators thus may enable a more complete account of performance.

Similarly, if we know that healthcare quality measures are heterogeneous but are ignorant of the best method to weight them, then methods to construct a composite indicator of performance need to capture that lack of knowledge. Sensitivity analysis

on weights is not a new idea; several prior attempts – especially in the multidimensional well-being literature – include explicit use of ranges of weights (e.g. Zhou et al., 2010); computation of multiple weighting schemes (e.g. Osberg and Sharpe, 2002); and global sensitivity analysis (e.g. Saltelli et al., 2008).

The REA approach adds to this work in two ways. First, consideration of incomplete information is built into the very structure of the model. Ranking intervals give policy-makers a sense of the uncertainty around ranks, indicating the extent to which action is warranted. Our results show that, when one assumes complete ignorance about the relative weights assigned to different indicators, then it is essentially impossible to differentiate the performance of Scottish Health Boards (Figure 3-1). In other words, one cannot say which Boards perform comparatively better or worse. Regulatory action solely based on such rankings would thus clearly be premature.

However, once some reasonable ordinal and proportional weight restrictions are applied, organizational performance appears much clarified. Clearly, the choice of weight restrictions may differ between groups of people: different individuals may come up with different orderings or proportionate weights concerning the relative badness (or goodness) of particular events. However, if some restrictions can be established (e.g. based on existing consensus or medical evidence of disease severity), then they may provide useful insights. When an organisation consistently appears at the bottom (Board G) or at the top (Board L; in Figure 3-2) whichever set of weights is used, this may provide a stronger rationale for policy intervention. It supports the notion that settling on a unique set of weights may not always be necessary to inform judgments in many situations (Foster and Sen, 1997).

Second, ranking intervals and dominance relations appear to offer relatively intuitive ways to synthesise key messages contained in disparate indicators. This may help to communicate in a unified way the results of comparative assessments to policy-makers, possibly addressing the limitations of frontier-based approaches such as DEA and stochastic frontier analysis whose complexity has tended to limit their practical influence outside academic circles (Hussey et al., 2009, Hollingsworth and Street, 2006). Visualisation of uncertainty also mitigates the loss of transparency due to

opaque methodological choices made about the valuation of objectives (Hauck and Street, 2006). Whether REA can live up to these expectations in practice remains to be seen, but some promising experiences from public sector education institutions (Salo and Punkka, 2011) suggest that REA may usefully complement existing methods of healthcare evaluators.

REA-type analyses are likely to be particularly useful under conditions where: (i) there are concerns about rank reversals due to sensitivity to outliers and the introduction or removal of DMUs (since pairwise comparisons make REA results relatively robust to these biases); (ii) the audience are policy-makers and managers rather than academics (since results such as being "30% below the efficient frontier" may not be easily accessible to non-technical audiences and REA requires no concept of an efficient frontier); and (iii) there are relatively few DMUs (since no large number of DMUs is needed to construct an efficient frontier and "peer groups" as in DEA).

3.6 Implications for policy and research

The agnosticism implied in the REA approach may come at a price of incomplete orderings (in the form of wide ranking intervals). This will depend on the extent to which weight restrictions are used; and on the correlation between indicators. These factors are linked, because strongly correlated indicators will make rankings less sensitive to different sets of weights (Foster et al., 2012). The appropriate degree of correlation will depend on the purpose of the analysis. For policy-makers and managers, wide ranking intervals simply reinforce the need to be cautious in using comparative assessments based on composite indicators for definitive judgments rather than as signals to motivate further analysis.

Dominance relations that are based on pairwise comparisons between Boards provide comparative performance assessments one can be confident about. Since dominance relations indicate that some DMU k performs at least as well as some other DMU l for all feasible weights and there exist some weights for which it

performs strictly better, this information could, for instance, be used for setting performance targets across all the indicators included in the analysis. Since improvements on some indicators may require less effort than others, indicator-specific improvements would also be informative. However, this would require a different approach. A recent study by Gouveia *et al.* (2015) employs slack-variables which define the variable-specific distance to the efficient frontier. This helps to estimate the improvements required for a DMU to reach the best comparative performance level. However, this approach does not indicate the improvements needed to reach some specific, non-efficient DMU as it is possible with the approach used in our study. This is particularly relevant from a managerial and policy perspective and an important strength of our approach, since the top performing organisation may not always be the most meaningful (and practically feasible) benchmark for another organisation that performs considerably worse.

Finally, it is essential to re-emphasize the importance of the other methodological choices (listed in section 3.2) that must be made when constructing a composite measure of performance. This concerns, in particular, the initial selection of indicators and any adjustment for environmental constraints on performance. If important indicators are not included in the analysis, then any performance evaluation will be meaningless. To mitigate dangers of omitting important variables, or including irrelevant variables (Smith, 1997), the set of performance metrics will need to reflect a country's definition of valued quality measures of the health service (Dowd et al., 2014).

With regard to uncontrollable influences on performance, in Scotland the funding formula mechanism is designed to enable all NHS Boards to produce equal levels of performance. Since this formula takes account of differences in population and other geographic characteristics (e.g. rurality), it can be argued that prior adjustment has already been carried out via the funding system (Jacobs et al., 2006b). However, the degree to which this argument holds depends on the context of analysis as well as the degree to which the formula accurately and comprehensively compensates for uncontrollable determinants of performance. While for Scottish Health Boards the funding formula argument may hold, it may not hold for hospital-level analyses

within these Boards since patients visiting these hospitals are likely to differ in relation to a number of additional case-mix variables.

Furthermore, as Smith (2003) notes, formula funding is fraught with challenges and imperfections, such as that performance criteria have proved hard to include in the formula. This means that poor quality of care which increases levels of morbidity might be 'rewarded' with higher levels of funding. In Scotland, the issue is further complicated by a policy called 'differential growth' where actual allocations to Boards do not entirely follow the formula. Instead, annual real-terms growth for Boards who are above parity (i.e. above their target share estimated by the formula) is lower than for Boards who are below parity until the formula-based funding distribution is achieved (ISD Scotland, 2010a). Finally, although several techniques exist to adjust for environmental variables (reviewed by Jacobs et al., 2006b: 115-17 in the context of DEA-type analyses), there is no universally accepted "best" method to tackle this problem in a satisfactory way. As a result, the link between resource allocation and performance measurement remains complex and an important avenue for future research.

3.1 Acknowledgements

The research was supported by the New Professors Fund at the University of Strathclyde. We thank Yan Feng for her discussion of an earlier draft of this paper at the Health Economists' Study Group, Glasgow 2014, and the participants in that conference. The views expressed here and any mistakes are those of the authors.

3.2 Commentary in relation to the organising matrix of strategies to address ambiguity about the standard for evaluation

This Chapter has explored how policy-makers and managers might manage ambiguity about the standard for evaluation with a technical-evidential approach. The Chapter has explored healthcare applications of a robust approach to ranking based on ranking intervals and dominance relations that recognise ambiguity about choices of weights and choices of population denominator.

A key problem remains the choice of healthcare quality measures. The indicators against which the performance of Scottish Health Boards is assessed, and which we have used here, focus on adverse events patients experience, thus embracing the imperative that healthcare should do no *harm*. However, measuring health system performance also requires some concept of the *benefit*: the good that is produced by a health system. Since health outcomes are a function of activities of the health system and exogeneous (e.g. lifestyle, demographic and socio-economic) factors, measures of health system performance should ideally indicate the "value-added" (Goldstein and Spiegelhalter, 1996): the health gain or incremental healthcare outcome conferred to patients that is attributable to the workings of the health system. In practice, operationalising this concept remains however difficult due to limited information on the counterfactual i.e. the health status in the absence of the health intervention (Jacobs et al., 2006b).

Using *activities* rather than health outcomes may offer insights into healthcare performance if a clear link between activities and health gain exists (Jacobs et al., 2006b). However, activity-based analyses often suffer from limited information about treatment appropriateness not only at a patient level (e.g. whether cataract surgery was clinically indicated) but also at a population level (e.g. whether all patients with capacity to benefit from cataract surgery also had access to the procedure). Chapters 4 and 5 explore this issue further by examining a technical-evidential methodology to establish a meaningful standard in the form of population capacity to benefit.

4 USING AN EPIDEMIOLOGICAL MODEL TO INVESTIGATE THE GAP BETWEEN NEED AND UTILISATION: THE CASE OF VENTILATION TUBES FOR OTITIS MEDIA WITH EFFUSION IN ENGLAND

Published as:

SCHANG, L.,[a] DE POLI, C.,[a] AIROLDI, M.,[a] MORTON, A.,[b] BOHM, N.,[c] LAKHANPAUL, M.,[d] SCHILDER, A.[c] & BEVAN, G.[a] 2014. Using an epidemiological model to investigate unwarranted variation: the case of ventilation tubes for otitis media with effusion in England. *Journal of Health Services Research & Policy,* 19(4), 236-44.

Author information (for the time when the research was conducted):

[a] Department of Management, London School of Economics and Political Science, London, UK

[b] Department of Management Science, Strathclyde Business School, University of Strathclyde, Glasgow, UK

[c] Ear Institute, University College London, England, UK

[d] General and Adolescent Paediatrics Unit, UCL Institute of Child Health, London, UK

Correspondence to:

Laura Schang
Department of Management
London School of Economics and Political Science
Houghton Street | London | United Kingdom
Email: L.K.Schang@lse.ac.uk

Key words: epidemiological models; otitis media with effusion; unwarranted variations.

Abstract

Objectives: To investigate the gap between need for and utilisation of ventilation tube (VT) insertions for otitis media with effusion (OME) in children in England. This procedure is known to be "overused" from audits of care provided, as only one in three VT insertions conform to the appropriateness criteria by the National Institute for Health and Care Excellence (NICE); but audits cannot identify the scale of "underuse": i.e. patients who would benefit but are not treated.

Methods: To explore both "underuse" and "overuse" of VTs for OME we developed an epidemiological model based on: definitions of children with OME expected to benefit from VTs according to NICE guidance; epidemiological and clinical information from a systematic review; and expert judgement. A range of estimates was derived using Monte Carlo simulation and compared with the number of VTs actually provided in the NHS in 2010.

Results: About 32,200 children in England would be expected to benefit from VTs for OME per year (between 20,411 and 45,231 with 90% certainty). The observed number of VTs for OME-associated diagnoses however was 16,824.

Conclusions: The expected population capacity to benefit from VTs for OME based on NICE guidance appeared to exceed, by far, the number of VTs actually provided in the NHS. So, while there is known overuse, there also may be substantial underuse of VTs for OME if NICE criteria were applied. Future investigations of unwarranted variation should therefore not only focus on patients who are treated, but consider potential to benefit at the population level.

4.1 Introduction

Systems of healthcare in countries that are under severe fiscal pressures (Thomson et al., 2014) seek to do more for less: to increase the benefits from healthcare and reduce its costs. There is evidence of large and persistent variations in medical practice across small areas, which have been documented in various countries (Corallo et al., 2014). This evidence is generally seen as an indication of "overuse": i.e. where reductions in rates of treatment could release resources with gains in health (Ham, 2013). In England, commissioners are allocated budgets for their populations and have to develop policies for services for which they are and are not prepared to pay. One such policy seeks to reduce unwarranted variation by restricting access to procedures listed as being of "low clinical value" (Audit Commission, 2011). However, due to the lack of an objective reference point against which to evaluate overuse (defined as "ineffective care that is more likely to harm than help the patient"; Institute of Medicine (2001: 47) or underuse (defined as "the failure to provide services from which the patient would likely benefit"; Institute of Medicine (2001: 17)), information on variations remains essentially ambiguous (Evans, 1990). The purpose of this article is to investigate unwarranted variations by modelling the scale of underuse or overuse of ventilation tubes (VTs; grommets) for children with otitis media with effusion (OME) in England.

VT insertions are a classic case of high geographic variation. Variations in England have been documented since the 1980s (Black, 1985b) and have persisted: in 2010/11 there was about eight-fold variation across 151 commissioners with a mean population of about 300,000 (NHS Right Care, 2012b). VTs have been listed by commissioners as a "low value"procedure (Audit Commission, 2011), which seeks to restrict referrals by general practitioners (GPs). Despite that, VT insertions remain one of the most frequent surgical interventions in children: with over 32,000 insertions in 2010/11, of which 23,500 were among children younger than 14 years (NHS Information Centre, 2011). Clinical audits in the United States (Keyhani et al., 2008a) and the United Kingdom (Daniel et al., 2013), using different criteria of

appropriateness, found that only one in three VT insertions were appropriate, suggesting substantial "overuse".

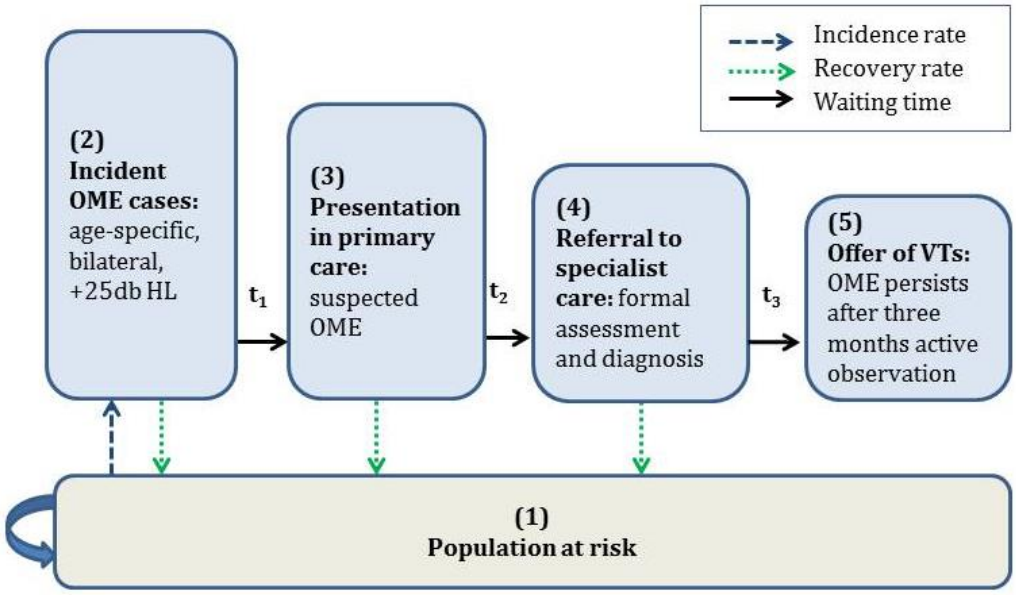
However, audits of care delivered cannot address the scale of "underuse" of VTs for OME. We therefore developed an epidemiological model to estimate the number of children with capacity to benefit from VTs for OME, if NICE guidance (NICE Guidance, 2008) were being followed, and compared this with the number of VTs actually provided in England. The study population are children aged 2 to 8 years.

4.1.1 Recommended clinical pathway

OME is defined as an effusion in the middle ear cleft, in the absence of signs of acute inflammation. It may cause conductive hearing loss which, if persistent, can affect speech and language development, educational performance and behaviour (Simpson et al., 2007). By the age of four years, about 80% of children have had episodes of OME (Zielhuis et al., 1990d). As OME is transitory for most children, the NICE clinical pathway (Figure 4-1) recommends an initial period of active observation over three months and repeat audiological testing at the end of this period.

At that stage, it is recommended that VTs are offered to children younger than 12 years who meet three core criteria: (1) bilateral OME with (2) a hearing level in the better hearing ear of 25 to 30 db hearing level or higher that (3) is documented over a period of three months. The crucial point is that NICE guidance does not define VTs as an intrinsically "low value" procedure, but recognises their value in relation to a set of evidence-based criteria. In exceptional cases, VTs may also be offered if clinicians judge the impact of OME-related hearing impairment on the child's development, well-being or social functioning to be substantial (NICE Guidance, 2008).

Figure 4-1 Conceptual model: NICE pathway of care



Explanation:

(1) The model starts with a population of children at risk of developing OME.

(2) Of these children, some will develop bilateral OME with a hearing level of +25 db.

(3) The recovery rate determines the proportion of children recovering and returning to the susceptible population. The remaining (persistent) cases present in primary care.

(4) Children who are referred to specialist care undergo formal assessment and diagnosis.

(5) Patients for whom a diagnosis of OME is confirmed after three months "watchful waiting" have a capacity to benefit from VTs for OME and should be considered for surgical intervention according to NICE guidance.

Legend:

Boxes represent mutually exclusive, collectively exhaustive states in which parts of the population of children find themselves.

Arrows represent the transition probabilities (incidence and recovery rates) and the waiting times that link the states.

4.2 Methods

Based on the NICE criteria, our epidemiological model to estimate population capacity to benefit from VTs for OME is formulated below. The modelling assumptions are summarised in Table 4-1. The parameters, their definition and estimation are given in Table 4-2.

4.2.1 Epidemiological model

1) Incidence: The number of new cases of OME in any given year, N(OME), is determined by the annual age-specific cumulative incidence (risk) I_j of OME multiplied by the susceptible population in a given age group S_j, summed over all eligible age groups *j*. The subgroup of cases with bilateral OME and a hearing level at NICE threshold level is expressed by

$$N\ (OME) = \sum_{j=0}^{12} \left(S_j * I_j * P\ (HL|Bilateral\ OME) * P(Bilateral\ OME\ |OME)\right)$$

2) Disease process: We model the probability of OME persisting at time *t* from the onset of OME as an exponential process (adapted from Zielhuis et al. (1990c)) of the form

$$P\ (OME \mid t) = \frac{1}{2^{\frac{t}{m}}}$$

3) Capacity to benefit from VTs for OME: As OME is transitory, the population with capacity to benefit will diminish as time passes since the onset of OME. Population capacity to benefit from VTs for OME is estimated as

$$PCB\ (t) = P\ (OME \mid t) * N\ (OME)$$

4.2.2 Data sources and extraction

To estimate parameter values, we carried out a systematic literature review according to PRISMA guidelines (Moher et al., 2009) (see Appendix 4-A for details of

the search strategy and data extraction, Appendix 4-B for the rationale for the study inclusion criteria).

4.2.3 Setting and population

The setting is the National Health Service (NHS) in England. The population includes children younger than 12 years covered by NICE guidance. However, as we were unable to find incidence studies that met our inclusion criteria for the age groups 0, 1, 4 and 9 to 12 years, we focused the analysis on children aged 2 to 8 years (extrapolating the incidence for 4-year olds from 3-year olds) which is the age group with the majority of VT insertions (0 to 12 years: 19,805; 2 to 8 years: 16,824 procedures with OME-associated diagnoses in 2010/11; NHS Information Centre (2011)). To estimate the susceptible population, the total population of children has been corrected for an estimate of OME prevalence (Appendix 4-C). We focused on children meeting the three NICE core criteria for VT insertion. The number of exceptional cases, which are identified through clinical judgement, was not modelled. This means that estimates from our epidemiological model are probably conservative and underestimate the number of children with capacity to benefit from VTs.

4.2.4 Model validation

All modelling assumptions were iteratively refined in consultation with the Project Steering Group. During an expert workshop in September 2012, ten participants with complementary expertise in audiology, ENT, general practice and epidemiology were invited to conduct a structured "walk-through" (Eddy et al., 2012: 846) to examine the model's overall structure and individual components. The group judged the model to be a fair representation of the NICE care pathway and of the disease process governing OME given the existing evidence base.

4.2.5 Sensitivity analysis

Data retrieved from the literature raised the issue of potential for bias in terms of internal validity (the extent to which the design of original studies ensured accurate

measurement of the parameters of interest) and external validity (the extent to which studies conducted e.g. two decades ago in a different setting were applicable to the present UK context). While we recognise the relevance of the literature-based data, we felt the different sources of uncertainty in the evidence would merit supplementing this with expert judgement. We followed a structured approach to expert elicitation (O'Hagan et al., 2006). We provided the panel of experts with the literature-based estimates, encouraged discussion and elicited fractiles of subjective probability distributions. We then used these estimates in a Monte Carlo simulation performed in @RISK 5.0 to gain an insight into the impact of the combined uncertainty in parameter estimates on the modelling results (Briggs et al., 2012).

Table 4-1 Modelling assumptions

Assumption	Comment
Exponential disease process	For the population level an exponential and rate-constant recovery process is applied based on Zielhuis et al. (1990c). The authors found a good fit (r^2 = 0.98) between the exponential model estimated with Kaplan-Meier technique and the empirical data from a prospective cohort study (n=816 children with valid measurements). For a discussion of the epidemiological models for representing the natural course of OME see Zielhuis et al. (1990a). However, this may mask the few children suffering from highly persistent OME. At the individual level, OME may also be more episodic.
Stationary population	Assumes a stable age distribution within each age group and year (based on mid-year population estimates).
t	Total waiting time t represents a parameter that reflects demand- and supply-side aspects of patient utilisation behaviour, access and referral policies and the organisation of care delivery. Is varied over a range to account for uncertainy in three distinct sub-intervals: • t_1, *time to presentation in primary care:* Seeks to account for the time lag for detecting hearing loss associated with OME. As OME is an often asymptomatic or "silent" condition, conductive hearing loss is likely to be noted by parents, teachers or carers only after some time (if at all). • t_2, *time from presentation in primary care to diagnosis in specialist care:* According to the NHS Constitution, patients have a right to be seen by a consultant within maximum 18 weeks after referral (Department of Health, 2010b). This is a political rather than clinical standard. It also refers to maximum not to optimum waiting times. National HES data confirms a median waiting time of 7.3 weeks (51 days) for grommets (NHS Information Centre, 2011) from the decision to admit to actual admission (excluding days of deferment and suspension). • t_3, *time from diagnosis to confirmation:* supposed to be 3 months according to NICE guidance.
Incidence is represented as a function of age	Age-based incidence rates are used as the association of OME with age is well-established and most reliably documented (Zielhuis et al., 1990b).
Incidence rates are at a population level and include both first and recurrent cases	About 50% of children recovering from OME experience a further episode of OME (Fiellau-Nikolajsen, 1983, Zielhuis et al., 1990c). However, due to the often asymptomatic character of OME, even robust incidence studies cannot rule out the possibility that a child has previously suffered from OME. Modelling history of OME could thus lead to an overestimation of cases. Therefore incidence rates used in the model do not differentiate between first-time and recurrent cases and are assumed to include both.

Assumption	Comment
Measurement of incident cases	The studies used to estimate the incidence of OME were based on screening intervals of 3 months (Zielhuis et al., 1990d) or 4 months (Williamson et al., 1994). This will underestimate transient cases occuring and recovering during this successive screening intervals. However, the assumption is justified insofar as OME is considered a disease occuring only after several weeks of middle ear pathology (Zielhuis et al., 1990a).
Seasonal variation in incidence is averaged out over one year.	While the incidence of OME is known to be higher in winter (Fiellau-Nikolajsen, 1983), the incidence data used in the model and the model quality measure represent an annual average.
Fixed proportion of bilateral OME.	Reflects the nature of the data that has been collected at (discrete) screening time points; although at individual level, children may switch between unilateral and bilateral states.

Table 4-2 Model parameters

Parameter	Definition	Base value used in model	References	Distribution for sensitivity analysis	Lower quartile; upper quartile [c]
S_j	Number of susceptible children in age group *j* at risk of developing OME in a given year (reference year 2010).	See the Appendix 4-B		-	-
I_j	Age-specific cumulative incidence (risk) of transiting to the OME state over a period of one year by year of age. Diagnosis based on type B tympanogram by the Jerger classification and otoscopy.				
2		0.350	Zielhuis et al. (1990d)	β (1.93;1.93;0.15;0.54)	0.280;0.420
3		0.160	Zielhuis et al. (1990d)	β (1.93;1.93;0.06;0.25)	0.128;0.192
4		0.160	[a]	β (1.93;1.93;0.06;0.25)	0.128;0.192
5		0.278	Williamson et al. (1994)	β (1.93;1.93;0.12;0.43)	0.222;0.334
6		0.151	Williamson et al. (1994)	β (1.93;1.93;0.06;0.23)	0.121;0.181
7		0.111	Williamson et al. (1994)	β (1.99;1.99;0.04;0.17)	0.088;0.133
8		0.065	Williamson et al. (1994)	β (1.93;1.93;0.03;0.11)	0.056;0.084
P (Bilateral OME | OME)	Conditional probability of bilateral OME given a diagnosis of OME.	0.4	Williamson et al. (1994)	β (303;455)	0.38;0.41
P (HL | Bilateral OME)	Conditional probability of a hearing level of +25dBgiven a diagnosis of bilateral OME.	0.35	Sabo et al. (2003)	β (11;11)	0.3;0.4
m	Median time to recovery ("half life" of OME)	3 months (three-month recovery rate of 0.5)	Thomsen and Tos (1981); Fiellau-Nikolajsen (1983); Tos (1984); Zielhuis et al. (1990c)	Used as deterministic value in the model as found to be consistent across different settings and time periods by various studies.	

Parameter	Definition	Base value used in model	References	Distribution for sensitivity analysis	Lower quartile; upper quartile [c]
t	Total waiting time t from OME onset	$t_1 + t_2 + t_3$	See Table 4-1	Varied over a range from 0 to 25 weeks	
t_1	Time from OME onset to presentation in primary care	1 month	[b]		
t_2	Time from presentation in primary care to formal diagnosis	1 month	[b]		
t_3	Time from formal diagnosis to offer of treatment (active observation or "watchful waiting")	3 months	NICE Guidance (2008)		

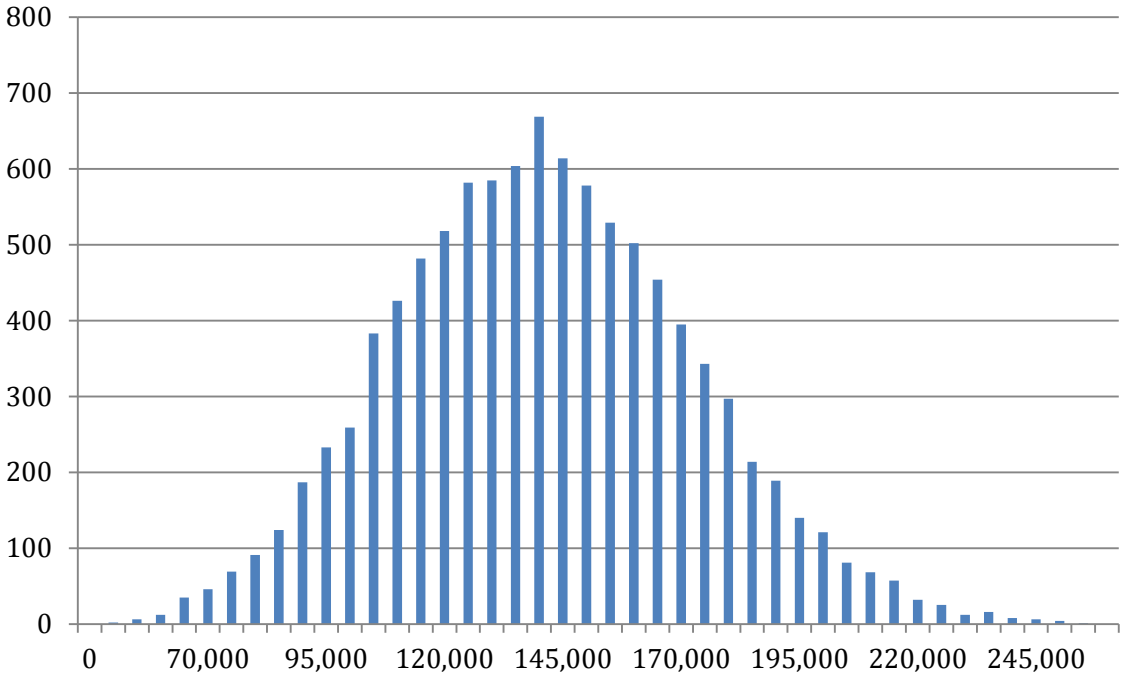
Estimates from clinical expert panel:
(a) extrapolating the incidence for 3-year olds;
(b) reflecting ideal circumstances;
(c) based on structured probability elicitation (O'Hagan et al., 2006).

4.3 Results

Figure 4-2 illustrates the combined uncertainty in the expected incidence of bilateral OME with a hearing level of +25 db. Based on 10,000 iterations of the simulation model and given the set of input distributions, the resulting distribution of the expected incidence ranges between 63,800 and 143,600 cases per year in England with 90% certainty (mean estimate: 102,083 cases). These results from the Monte Carlo simulation are used to model the expected number of children with capacity to benefit from VTs for OME as the total waiting time from the onset of OME is varied over a range.

Since OME is transitory, the expected population capacity to benefit from VTs for OME depends on the total waiting time from the onset of OME to the point where treatment is considered (Figure 4-3). NICE guidance recommends a three-month period of active observation following the first formal diagnosis. Thus, if we were to assume the first outpatient appointment took place instantaneously after the onset of OME, then the mean estimate of children for whom VTs would be clinically indicated would be approximately 51,000 (at t=3 months; between 32,400 and 71,800 with 90% certainty). There is currently no national guidance on the recommended waiting time from the onset of OME until the first outpatient appointment (waiting time intervals t_1 and t_2 in Figure 4-1). Since our model aims to provide a benchmark of expected care, rather than a reflection of actual practice, our assumptions about the length of these intervals (Table 4-2) represent clinically "ideal" circumstances based on expert group consensus. Assuming a one-month buffer period before parents become concerned about the symptoms of OME and visit a GP and another month before children have their first outpatient appointment, we would expect approximately 32,200 children to benefit from VTs for OME (at t=5 months; 90% certainty interval 20,411 to 45,231). This contrasts with an observed number of 16,824 VTs that were actually provided for OME-associated diagnosis codes in the age group of 2 to 8 years in 2010/11 in England. As can be seen in Table 4-3, even if we were to assume coding inaccuracies in VTs coded with OME-associated diagnoses, the conclusions would be unaffected.

Figure 4-2 Monte Carlo simulation of expected annual incidence of bilateral OME with a hearing level of +25 dB in England

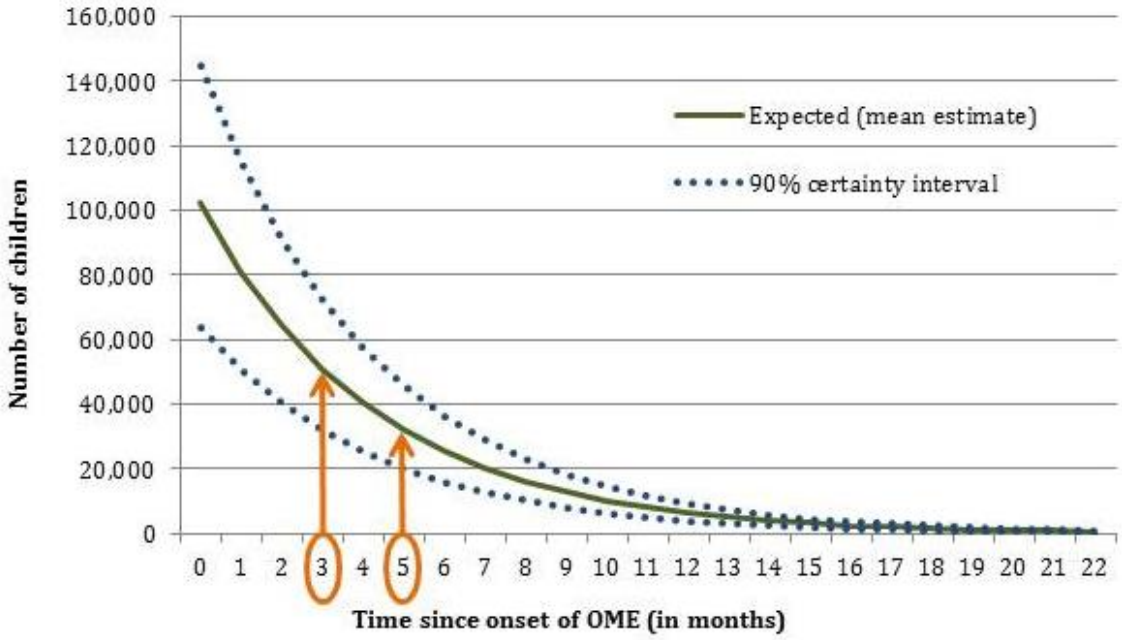


Legend:

x-axis: expected annual incidence of bilateral OME with +25 dB hearing level in England (2010).

y-axis: frequency of observing a particular quality measure value based on 10,000 iterations of the simulation model.

Figure 4-3 Expected number of children with capacity to benefit from VTs for OME depending on total waiting time in England (reference year 2010, age groups 2 to 8 years)*



*Given different starting estimates of the total annual incidence of bilateral OME with hearing level of +25dB for the age groups 2 to 8 from the Monte Carlo simulation (Figure 4-2) of approximately 102,083 cases (mean estimate); 63,800 cases (lower 5% bound); and 143,600 cases (upper 95% bound).

Table 4-3 Observed VT insertions in England, 2010/11

Observed VT insertions	Count
Total admissions	32,716
Day case	29,566
Age 0-14	23,459
Age 0-12, OME-associated diagnosis codes (2010/11)*	19,805
Age 2-8, OME-associated diagnosis codes (2010/11)*	16,824

Source: NHS Information Centre. Main procedures and interventions: 4 character. Hospital Episode Statistics for England. Inpatient statistics, 2010-11.

* Procedure code D15.1 Myringotomy with insertion of ventilation tube through tympanic membrane for DIAG1=H652: Chronic serous otitis media or H653: Chronic mucoid otitis media or H654: Other chronic nonsuppurative otitis media or H659: Nonsuppurative otitis media, unspecified. Both as primary and secondary procedure (e.g. besides adenoidectomy); including both elective and emergency admissions, in- and outpatient cases.

4.4 Discussion

This study shows that the expected capacity to benefit from VTs for OME among children in England, according to NICE guidance, exceeds the number of VTs that were actually provided in the NHS. Our model hence reveals the possibility of underuse of VTs for OME at the aggregate national level. However, the findings also need to be interpreted in the light of the roughly eight-fold variation in treatment rates across PCTs in England (NHS Right Care, 2012b), which suggests that overuse might still occur in some regions.

4.4.1 Strengths and weaknesses of the study

The model draws on evidence-based clinical guidance to obtain an indicative estimate of the scale of potential underuse or overuse of VTs in a given population. This estimate does not represent the "right (treatment) rate", which would also depend on informed patient choice. It attempts to approximate a level of treatment that the NHS would be expected to offer to patients, if NICE criteria were accepted as a valid basis for identifying patients with capacity to benefit from VTs. We recognise that NICE criteria can only be approximate predictors of benefit from VTs for hearing outcomes, especially for cases located just above or below the +25 dB hearing level threshold, with even more uncertainty over the impact of VTs on childhood development and the child's quality of life. Thus, from a normative standpoint, our model can only give an approximate estimate of how many VTs "should" be offered, which may change once better predictors of benefit become available.

The model uses best available evidence identified through a systematic review. The shortage of high-quality studies meeting our inclusion criteria did not allow for a meta-analysis, and we have demonstrated the consequent uncertainty in our parameter estimates and their combined impact on the modelling results by Monte Carlo simulation. The observed number of VTs provided covers patients treated in the NHS; unfortunately we were unable to obtain estimates of the scale of private practice in England. However, total private sector expenditure on healthcare in the UK (2011) is 17.2% (Office for National Statistics, 2012) which would not substantially affect the conclusions of our study.

4.4.2 Findings in relation to studies of utilisation

Our study using a population model complements utilisation-based studies of treatment appropriateness. A recent multi-centre study in England found that only 32.2% of VTs inserted complied with the three core NICE criteria, while 54.8% of VTs were provided on the basis of exceptional circumstances (Daniel et al., 2013). Although NICE guidelines explicitly encourage the provision of VTs also beyond the three core criteria if clinicians judge the impact of OME on the child's development and social functioning to be substantial (NICE Guidance, 2008), the apparent reframing of

"exceptions" under clinical guidance as the "rule" in clinical practice does raise questions over treatment appropriateness. This study adds to these findings by illustrating that, while there may be deviation from NICE core criteria, which could either reflect patient-oriented treatment or overuse of VTs, unmet clinical need according to these core criteria may be present simultaneously.

There are three possible lines of explanation for the divergent findings between the Daniel et al. study, which found overuse of VTs, and this study, which identifies a net underuse at the population level. These explanations are as follows:

Explanation 1: The model is biased. The model depends on a number of assumptions (see Tables 4-1 and 4-2) and limitations (see section 4.4.1). In particular, the epidemiology of OME is complex (Zielhuis *et al.,* 1990a) and the use of age-specific incidence rates in this study may not have been able to account for other predictors of incidence. However, as highlighted in Appendix 4-B, much care was taken in specifying the selection criteria for the studies. To enhance external validity of the data, for instance, data only from areas with similar climate conditions was chosen since climate is a known risk factor for OME (Rovers *et al.,* 2004). As a result of comprehensive sensitivity analysis across all model parameters at the same time, one can be confident that the model estimates are reasonable.

Explanation 2: The Daniel *et al.* (2013) study is biased. While the epidemiological model has England-wide coverage, the Daniel *et al.* (2013) study examined only five hospital centres (Derbyshire Hospitals NHS Trust, Nottingham University Hospitals NHS Trust, Royal Free Hampstead NHS Trust London, Sherwood Forrest Hospitals NHS Trust, and United Lincolnshire Hospitals NHS Trust). It is possible that these five centres were exhibiting an unusually high level of inappropriate insertions of ventilation tubes. Therefore, they may not have been representative of England as a whole. Furthermore, the use of clinical audits of patients who were treated suffers from a fundamental methodological limitation: per definition, such audits are not able to identify patients who would benefit but who did not access healthcare in the first place or failed to be referred in time by their GP.

Explanation 3: Both studies show a true, complementary aspect of the reality of healthcare delivery. There are several good reasons why the apparent co-existence of both overuse and underuse is a genuine phenomenon rather than a statistical artefact. On the one hand, there are three logical explanations for an apparent underuse of VTs as identified by our model. First, as parents, teachers and nurseries may fail to recognise hearing loss associated with OME (Rosenfeld *et al.*, 1998), it is possible that many patients do not present to primary care in the first place. Second, GPs may lack the knowledge or capacity to diagnose OME correctly. In a recent UK-based study, participating GPs correctly identified OME only in 53% of cases, which is not much higher than chance (Buchanan and Pothier, 2008). This means that, on average, a GP misses every second child with OME. At a national scale, this factor alone would provide a plausible explanation of why the expected population capacity to benefit from VTs was twice as high as the observed rate of utilisation. Third, the finding of apparent underuse is consistent with the "low value" policy among healthcare commissioners in England which has entailed restricting access to VTs (Audit Commission, 2011). This means that GPs might be inclined to avoid or delay referrals even for patients for whom NICE guidance would recommend a referral.

On the other hand, there are several possible explanations for Daniel *et al.'s* (2013) finding that, once patients have been referred to specialist care, surgeons may have a tendency to insert VTs even if the three NICE 'core criteria' have not been met. First, as Daniel *et al.* (2013) point out, it may be the case that surgeons perceive the NICE guidelines as overly restrictive. There may hence be an attempt to take into account the patient's situation as a whole and personalise care to what surgeons perceive to be in this particular patient's best interest. Second, economic incentives to increase rates of surgery may play a role. This may be particularly relevant given the increasing financial pressures and challenging economic environments faced by many NHS hospitals (Hurst and Williams, 2012). Third, parental preferences might influence the decision to operate. It has long been argued that some socio-economic groups, allegedly the relatively well-off middle class patients, might exercise their 'voice' and demand treatment even though objectively they are not in medical need (Le Grand, 2007). For these reasons, it seems plausible to suggest that the English NHS is likely to suffer from

both overuse and underuse with respect to VTs for OME. Further empirical analysis is necessary to examine these factors.

4.4.3 Policy implications

An increasingly common policy among healthcare commissioners in England is to label VTs *per se* as "overused" and "low value" (Audit Commission, 2011) and hence restrict access to the procedure. Our findings highlight the possibility of substantial "underuse" among children in England for whom VTs are deemed beneficial and thus call for a more nuanced policy response. Because there is no evidence of a systematic relationship between high rates of utilisation and high rates of inappropriateness (Keyhani et al., 2012), we need a policy that tackles overuse by clinical audit of treatment, and ensures access to effective care for children suffering from persistent bilateral OME with a degree of hearing loss that is disabling and may affect their health and development. This policy would use the ideas of epidemiologic surveillance of medical care (Caper, 1987) to enlarge the framing of clinical appropriateness from audits of services delivered to population capacity to benefit. Understanding the number of people who might be expected to benefit, given local population characteristics and clinical guidance, has relevance also for other high-volume services such as cataract surgery, joint arthroplasty or spinal procedures: it could help widen clinical concerns from individual patients towards the entire population who could (not) benefit and should hence (not) be offered a procedure. This policy would require investments in: (1) recommended intervention criteria that are more directly related to patient benefit, based on evidence from everyday practice (high-quality clinical databases rather than RCTs) on the real-world impacts of surgery on health outcomes compared to a control group; and (2) good information on disease epidemiology.

4.4.4 Implications for research and quality improvement

To explain the discrepancy between observed VT provisions and the expected number of VTs offered, a multi-faceted qualitative and quantitative approach involving commissioners, professionals and families is needed to identify barriers

along the whole pathway and then design interventions for improvement. As parents, teachers and nurseries may fail to recognise hearing loss associated with OME (Rosenfeld et al., 1998), it is possible that many patients do not present to primary care in the first place. GPs, school nurses and health visitors need the knowledge and capacity to identify patients with suspected OME and ensure timely referral and diagnosis according to NICE criteria. In a recent UK-based study, participating GPs correctly identified OME only in 53% of cases, which is not much higher than chance (Buchanan and Pothier, 2008). Since VTs feature widely as a "low value" procedure (Audit Commission, 2011), GPs might also tend to withhold referrals even for patients for whom VTs could be a clinically and cost-effective option. Delays in care and a long history of "watchful waiting" in community services may thus, in practice, exceed the two-month interval from the onset of OME to formal diagnosis which we assumed as a clinically "ideal" benchmark in our model. To overcome fragmentation, GPs, audiologists and ENT specialists need to work together to ensure early recognition and referral of children with capacity to benefit from treatment. Patients and carers deliberately choosing non-surgical treatment alternatives, such as hearing aids or medical management, may also in part explain the apparent discrepancy between "expected" and "observed". However, many patients and carers may not be given the opportunity to discuss and understand their options for treatment, resulting in uninformed use of other care. Future research might therefore also examine regional variations in patient preferences and approaches to shared decision-making (Elwyn et al., 2010) and how these add to, or interact with, differences in local commissioning criteria and socio-economic inequalities.

4.5 Conclusions

This study has examined the case of VTs for OME which, although known to be "overused" based on audits of care provided, also seem to be substantially "underused" at a population level in England based on NICE guidance. Because overuse and underuse may co-exist as sources of unwarranted variation, clinicians and managers should examine if all children who would be expected to benefit from

VTs for OME also have access to the procedure. The study is of one condition in England but raises an important general issue over using studies of medical practice variations to inform policies to reduce overuse and thus release resources to meet rising demand in times of austerity. To maximise benefits for patients within resource constraints, policies where medical practice varies ought to tackle overuse by auditing care that is provided, and underuse by assessing capacity to benefit in populations.

4.6 Acknowledgements

This work was supported by the Health Foundation [grant number 6179]. We would like to especially thank Martin Birchall for his enthusiasm and support for the project; Simon Swift and Adam Ceney for facilitating access to Hospital Episode Statistics data; members of the steering group including Bengi Beyzade, General Practitioner in Islington; Ian Colvin, General Practitioner at Elizabeth Avenue Group Practice, Islington; Kelvin Kwa, ENT Clinical Fellow at the UCL Ear Institute; James Mountford, Director of Clinical Quality at UCLPartners; and the expert panel who generously gave their time, insight and judgement to discuss the model: Martin Birchall, Professor of Laryngology at RNTNE; George Browning, Professor of Otorhinolaryngology at the MRC Institute of Hearing Research in Glasgow; Thembela Guzula, Senior Audiologist at Barts Health; Julie Hare, Consultant Speech and Language Therapist at RNTNE; Martin Marshall, Professor of Healthcare Improvement at UCL; Seema Patel, Lead Audiologist at Barts Health; and one paediatric ENT consultant who wished to remain anonymous. The conclusions reached are those of the authors. The authors alone are responsible for any mistakes.

4.1 Commentary in relation to the organising matrix of strategies to address ambiguity about the standard for evaluation

This Chapter has explored how policy-makers and managers might establish meaningful standards for evaluation with a technical-evidential approach. To this end, the Chapter has developed an epidemiological model to investigate overuse and

underuse in ventilation tube surgery for children with otitis media with effusion in England. The Chapter has shown that underuse and overuse may co-exist and that a more nuanced policy is required to increase appropriateness in the provision of ventilation tubes.

4.2 Appendix

4.2.1 Appendix 4-A. Systematic literature review: Search strategy and data extraction

A systematic literature review was carried out using the databases PubMed, DARE, Scopus, Web of Science and the Cochrane Library (timespan: all available years; restriction to studies in English language). After removing duplicates, 1302 studies were screened independently by the first and second authors based on pre-defined criteria. To be eligible, studies needed to (i) be population-based screening studies; (ii) have a prospective design; (iii) follow defined case finding and diagnostic methods; (iv) provide incidence rates by year of age; and (v) be conducted in Europe or North America. The detailed rationale for each criterion is stated in Appendix 4-B. Study selection was discussed among members of the research team, with the Project Steering Group and during a workshop with UK-based clinical and epidemiological experts. Those studies judged to be in line with the selection criteria were retained.

Database	Search criteria	Number of results
DARE	(("otitis media with effusion" OR "glue ear" OR "non suppurative otitis media" OR "serous otitis media" OR "secretory otitis media" OR "middle ear effusion" OR "purulent otitis media with effusion")) AND (("prevalence" OR "incidence" OR "epidemiology" OR "occurrence")) AND (("child*" OR "kid*" OR "infan*"))	15
Cochrane library	("otitisQUOTESPACEmediaQUOTESPACEwithQUOTESPACEeffusion" OR "glueQUOTESPACEear" OR "nonQUOTESPACEsuppurativeQUOTESPACEotitisQUOTESPACEmedia" OR "serousQUOTESPACEotitisQUOTESPACEmedia" OR "secretoryQUOTESPACEotitisQUOTESPACEmedia" OR "middleQUOTESPACEearQUOTESPACEeffusion" OR "purulentQUOTESPACEotitisQUOTESPACEmediaQUOTESPACEwithQUOTESPACEeffusion") and ("prevalence" OR "incidence" OR "epidemiology" OR "occurrence") and ("child*" OR "kid*" OR "infan*") not ("acuteQUOTESPACEotitisQUOTESPACEmedia" OR "recurrentQUOTESPACEacuteQUOTESPACEotitisQUOTESPACEmedia") not ("adult*") NOT ("animal*") NOT ("cleftQUOTESPACEpalate" OR "down'sQUOTESPACEsyndrome" OR "downQUOTESPACEsyndrome") in Cochrane Database of Systematic Reviews"	57
Web of science	Topic=(("otitis media with effusion" OR "glue ear" OR "non suppurative otitis media" OR "serous otitis media" OR "secretory otitis media" OR "middle ear effusion" OR "purulent otitis media with effusion")) AND Topic=("prevalence" OR "incidence" OR "epidemiology" OR "occurrence") AND Topic=(("child*" OR "kid*" OR "infan*")) NOT Topic=(("acute otitis media" OR "recurrent acute otitis	635

Database	Search criteria	Number of results
	media”))NOT Topic=("adult*")NOT Topic=("animal*")NOT Topic=(("cleft palate" OR "down's syndrome" OR "down syndrome")) Refined by: Languages=(ENGLISH) AND [excluding] Subject Areas=(PHYSICS OR URBAN STUDIES OR PLANT SCIENCES OR HISTORY) Timespan=All Years. Lemmatization=On	
PubMed	((((((("otitis media with effusion"[All Fields] OR "glue ear"[All Fields] OR "non suppurative otitis media"[All Fields] OR "serous otitis media"[All Fields] OR "secretory otitis media"[All Fields] OR "middle ear effusion"[All Fields] OR (("otitis media, suppurative"[MeSH Terms] OR ("otitis"[All Fields] AND "media"[All Fields] AND "suppurative"[All Fields]) OR "suppurative otitis media"[All Fields] OR ("purulent"[All Fields] AND "otitis"[All Fields] AND "media"[All Fields]) OR "purulent otitis media"[All Fields]) AND effusion[All Fields])) AND ("prevalence"[All Fields] OR "incidence"[All Fields] OR "epidemiology"[All Fields] OR "occurrence"[All Fields])) AND ("child*"[All Fields] OR "kid*"[All Fields] OR "infan*"[All Fields])) NOT ("acute otitis media"[All Fields] OR "recurrent acute otitis media"[All Fields])) NOT "adult*"[All Fields]) NOT "animal*"[All Fields]) NOT ("cleft palate"[All Fields] OR "down's syndrome"[All Fields] OR "down syndrome"[All Fields]) AND ("humans"[MeSH Terms] AND English[lang])	538
Scopus	(ALL(("otitis media with effusion" OR "glue ear" OR "non suppurative otitis media" OR "serous otitis media" OR "secretory otitis media" OR "middle ear effusion" OR "purulent otitis media with effusion")) AND ALL(("prevalence" OR "incidence" OR "epidemiology" OR "occurrence")) AND ALL(("child*" OR "kid*" OR "infan*")) AND NOT ALL(("acute otitis media" OR "recurrent acute otitis media")) AND NOT ALL(("adult*")) AND NOT ALL(("animal*")) AND NOT ALL(("cleft palate" OR "down's syndrome" OR "down syndrome"))) AND (LIMIT-TO(LANGUAGE, "English")) AND (EXCLUDE(SUBJAREA, "AGRI") OR EXCLUDE(SUBJAREA, "PHYS"))	947

4.2.2 Appendix 4-B. Study inclusion criteria

Inclusion criteria	Rationale	Exclusion criteria
(i) Population-based screening study	For valid estimates of incidence, the denominator should include all, or a representative sample of, individuals at risk. As regards hearing loss: most literature focuses on clinical populations which are likely to, on average, experience higher levels of hearing loss than children with OME in the general population. The model therefore uses epidemiological data on hearing loss from a community-based study (Sabo et al., 2003: 44).	1) *Utilisation-based studies* (i.e. with the number of people actually visiting the doctor as the denominator): A single hospital or practice cannot usually be assumed to provide care for a well-defined population that is representative of a larger group (Fiellau-Nikolajsen, 1983). 2) *Trial-based studies:* Results may be difficult to generalise to a general population setting if particular groups are over- or underrepresented. 3) *Studies with high-risk populations* e.g. pre-term babies on intensive care units, exclusive focus on children in daycare. 4) *Clinical specialist populations* (for estimating the proportion of hearing loss among all OME cases): If the denominator are children who have already been referred to ENT (Haggard, 2009, Fria et al., 1985, Ungkanont et al., 2010) this may either lead to overestimation (due to selectivity of more severe cases) or underestimation (due to bias in detection and presentation among parents and/ or gaps in referral from primary care). 5) *Self-report studies:* As regards incidence and hearing loss, parents have been shown to be inaccurate in their judgments regarding the presence of hearing loss that may accompany an episode of OME (Rosenfeld et al., 1998).
(ii) Prospective design	OME often presents asymptomatically, which complicates retrospective diagnosis of OME.	*Retrospective designs* (e.g. parent interviews or analysis of doctor consultations): These will substantially underestimate the true incidence of OME (Roland et al., 1989) and are thus not a reliable case finding design for OME.

Inclusion criteria	Rationale	Exclusion criteria
(iii) Case finding methods and diagnosis	The recommended diagnostic algorithm for OME combines impedance audiometry (tympanometry) with pneumatic otoscopy (Rovers et al., 2004). OME is diagnosed when tympanometry reveals a flat curve (relative gradient less than 0.1, type B) or middle ear pressure between -399 to -200 daPa (C2 curve), when the tympanic membrane has no or reduced mobility, or fluid or air bubbles are evident behind the ear drum (Simpson et al., 2007).	Studies that do not provide correspondingly defined case finding and diagnostic methods.
(iv) Stratified by year of age	Incidence of OME is known to vary considerably by age (Zielhuis et al., 1990b).	Studies that report only aggregate (e.g. five-year) rates since these are likely to obscure key variations in incidence across age groups.
(v) Studies conducted in Europe or North America	Incidence of OME may be influenced by climatic settings (Black, 1985a).	Studies conducted in different climatic settings than England.

4.2.3 Appendix 4-C. Estimation of susceptible population

For valid estimations of incident cases, children with prevalent OME at the beginning of the study period need to be subtracted from the total population to obtain an estimate of the susceptible population (i.e. the population at risk). This is because the denominator of the cumulative incidence is defined as the number of children at risk at beginning of the study period rather than the total population (Morgenstern et al., 1980). Point prevalences are taken from population-based studies. The estimates are lower than those reported in a review by Zielhuis et al. (1990b) which may be due to the amalgamation of point and period prevalences (time frames over which prevalence was measured were not reported) in their review.

j	**P_j**		**N_j**	**$S_j = N_j-(N_j*P_j)$**
Age group	**Point prevalence (%)**	**Reference**	**Total population**	**Susceptible population**
2	10.61	Tos (1984)	667,185	596,423
3	9.8	Fiellau-Nikolajsen (1983)	640,232	577,489
4	8.8	Fiellau-Nikolajsen (1983)	620,326	565,737
5	10	Fiellau-Nikolajsen (1983)	606,770	546,093
6	6.1	Fiellau-Nikolajsen (1983)	598,725	562,203
7	3.04	Tos (1984)	577,767	560,183
8	1.11	Tos (1984)	560,460	554,233

Source: Own calculation based on Office for National Statistics (2011) population data.

5 GEOGRAPHIC VARIATIONS IN HEALTHCARE AND THE PROBLEM OF POPULATION NEEDS: CAN ESTIMATES OF CAPACITY TO BENEFIT IN POPULATIONS PROVIDE A FEASIBLE AND USEFUL BENCHMARK? A REVIEW

Author:

Laura Schang
Department of Management
London School of Economics and Political Science
Houghton Street | London | United Kingdom
Email: L.K.Schang@lse.ac.uk

Key words: need for healthcare; population capacity to benefit; overuse; underuse; variations in healthcare; health service planning.

Abstract

Background: The conventional approach to account for population needs in studies of geographic variation in healthcare is to standardise utilisation rates for variables associated with need (e.g. age, deprivation). However, this approach provides no benchmark of the extent to which actual utilisation meets the appropriate level of care. This paper examines the concept of population capacity to benefit (PCB) based on criteria of capacity to benefit from a particular intervention and their prevalence and incidence in the population of interest.

Methods: Studies following the PCB approach were identified from a keyword search of the PubMed, Scopus, Web of Science and Cinahl databases.

Results: 22 studies from the United Kingdom, Ireland, Canada and Australia were identified which estimated population requirements for hip and knee replacement, radiotherapy, coronary revascularisation, cataract surgery, dental care, prostatectomy, stroke care and ventilation tube surgery. Criteria of capacity to benefit were obtained from consensus panels, guidance endorsed by professional associations or Health Technology Assessment institutions. Fifteen studies extrapolated epidemiological information from other contexts but only six studies assessed the consequent uncertainty through sensitivity analysis. Estimated population benchmarks varied depending on the chosen criteria and threshold values for intervention and whether patient preferences where taken into account.

Conclusion: Measuring PCB provides a theoretically sound complement to standardised rates but is unlikely to produce a single unambiguous “right rate“ of population need. Progress with evidence-based guidelines, population surveillance systems and better use of established methods to handle uncertainty create scope to overcome some of the hurdles faced by studies in the 1990s and 2000s. National agencies should consider developing population benchmarks for resource-intensive conditions and use these to support regional healthcare planning and surveillance.

5.1 Introduction

Geographic variations in rates of hospital admission, surgical procedures and resource supply have been widely documented within countries in the form of national Atlases of Variation in England, Germany, Spain, the Netherlands, the United States and many other countries (Corallo et al., 2014) and in cross-national projects by ECHO (Bernal-Delgado et al., 2015) and the OECD (2014). These variations challenge the goal of many health systems in Europe to ensure equal opportunity of access for equal need. They are often interpreted as a signal of widespread overuse of unnecessary or even harmful care and underuse of effective care that fails to meet population needs (Appleby et al., 2011, OECD (Organisation for Economic Co-operation and Development), 2014).

The conventional method to account for population needs in analyses of variations is to standardise crude utilisation rates for variables shown to be associated with population need for care, such as age, sex and area-level deprivation. This is essential to enable fair comparisons between regions whose performance may differ due to factors that are outside the region's control (Nicholl et al., 2013). For example, geographic variations in rates of hip replacement should be standardised at least by age, as regions with a higher proportion of older people would be expected to have a higher incidence of osteoarthritis and thus higher rates of need for joint replacement than regions with younger age distributions, *ceteris paribus*.

This paper is motivated by the problem that such standardised rates can however not provide a benchmark of population need for healthcare, defined as the extent to which utilisation of a specific service exceeds or falls short of a level of care that is expected to be beneficial for a defined population. The directly age-standardised rate is the rate of utilisation in a region that would be expected *if* that region had the same age structure as an (arbitrarily chosen) standard population (Breslow and Day, 1987). This hypothetical estimate does not indicate the number of people with capacity to benefit from a particular intervention to gauge "how much" scope there is

for reducing or expanding service levels commensurate with region-specific needs for care.

As Tversky and Kahneman (1991) have shown, a common psychological response is to anchor judgement on the reference points that are provided in a set of data. In the absence of a reference point for local need, research shows that health service managers are tempted to evaluate their region's performance in relation to the national average as the reference point, inferring that an above-average position suggests "overuse" while a below-average position suggests "underuse" (Schang et al., 2014b). However, the average utilisation rate, adjusted for variables associated with need, has no normative justification as a benchmark of population need for care. There is also little empirical evidence that above-average intervention rates entail a higher proportion of ineffective care. A systematic review of studies of overuse in the United States (Keyhani et al., 2012) found that high and low use areas showed similar levels of inappropriate procedures. This problematises the comparative approach (Bradshaw, 1972) to inferring the degree to which actual utilisation meets population needs from the distribution of standardised utilisation rates. The lack of a normative benchmark of population needs therefore limits the usefulness of information on variations for healthcare planning (Mercuri et al., 2013).

The aim of this paper is twofold: To define the concept of population capacity to benefit (PCB); and to examine its feasibility and usefulness to identify underuse or overuse relative to a population's need for a specified intervention based on a review of empirical applications. The next section defines the PCB concept and delineates it from alternative conceptions of "need". Subsequently, the methods and results of the literature review are presented. Finally, implications for policy are discussed.

5.2 Theoretical background

5.2.1 How to define "need for healthcare"?

Epidemiology has traditionally focused on measuring the burden and distribution of disease in populations (Ezzati et al., 2004). The basis for ascribing a "need" to a

person is thus evidence of a poor initial health state (Hasman et al., 2006). However, not every need for *health* entails a need for *healthcare* (Culyer and Wagstaff, 1993). The magnitude of health deficits (i.e. gaps between actual and desired health status) is hence not equivalent to the level of health services required to *improve* population health (Culyer and Wagstaff, 1993, Wright et al., 1998, Mooney and Houston, 2004). This may be, firstly, because no effective interventions exist to prevent, cure or care for a health problem. Secondly, effective interventions that exist may fall outside the remit of health systems. For instance, policies to tackle inequalities in life expectancy between socio-economic groups may be found largely in employment and education sectors and in the physical and social environment (McQueen et al., 2012, Mackenbach, 2012). The definition of a "need for healthcare" thus requires the ability to address a health problem within the boundaries of the health system.

In addition to comparatively defined need mentioned in the introduction, Bradshaw (1972) proposed three further types of need: felt need (wants, desires); expressed need (vocalised wants or use of services); and normative need (which is assessed in relation to a desirable standard). The first three types are problematic insofar as they risk confusing the concepts of need, preference and utilisation. The next paragraphs thus focus on normative standards of need for healthcare. Two fundamental challenges in their definition are reviewed: the choice of an underlying value basis and of the desired level of the standard. Against these considerations, a definition of need for healthcare in the context of unwarranted variation in healthcare utilisation is provided.

A natural basis to define "need for healthcare" is the capacity to benefit from healthcare (Culyer and Wagstaff, 1993, Culyer, 1995, Stevens and Gillam, 1998). This view requires the availability of effective interventions which are likely to improve clinical outcomes (e.g. physical functioning) or quality of life (e.g. less pain and anxiety). Within the paradigm of evidence-based medicine, the guiding question is thus whether the balance between benefits and risks of an intervention is expected to produce a net benefit. As Hasman et al. (2006) point out, need for some healthcare intervention has to be expressed as a condition-intervention-pairing: that is, a health state (e.g. severe osteoarthritis with moderate pain and functional impairment) for

which a particular intervention (e.g. total hip replacement) is deemed to enable gains in health.

A sole focus on effectiveness in defining need ignores however the potential for an inefficient use of limited resources (Cochrane, 1972, Acheson, 1978, Culyer, 1995, Mooney and Houston, 2004). Culyer (1995) defines need for healthcare therefore as "the minimum amount of resources required to exhaust a person's capacity to benefit" (Culyer, 1995: 728). This health economic definition has two important implications and, arguably, advantages. First, if similar benefits can be achieved with medical intervention (e.g. a visit to the doctor) and without medical intervention (by patience and watchful waiting), then one cannot assert the presence of a "healthcare need". This may be the case for many typically self-limiting conditions such as mild headaches and common colds. Second, an intervention cannot be said to be as "needed" as another intervention which is equally effective but requires less resources for its delivery (Culyer, 1995). Thus, while under a medical effectiveness paradigm one would be indifferent to the choice between a branded and a generic drug (with identical bioactive ingredients), under an economic evaluation paradigm the cheaper generic drug obviously dominates.

Notably, Culyer's definition implies a *maximum* level of capacity to benefit as the standard against which need is assessed: a "need" for services can be said to exist up to the point where capacity to benefit is exhausted. In reality, however, need for health or social care interventions may instead be defined in relation to an acceptable or "normal" functioning range (Hasman et al., 2006: p.149f.). In the United Kingdom, for instance, Department of Health (1991) guidance defined need for care management as the "requirements of individuals to enable them to achieve, maintain or restore an *acceptable* [added emphasis] level of social independence or quality of life, as defined by particular care agency or authority" (p.12f.). Following this view, a funding agency may thus recognise a person's need for physiotherapy premised on restoring the ability to participate in activities of daily living, but not necessarily a need for athletics training to prepare for the Olympics.

Clearly, not every healthcare need as defined by the prevalence of avoidable (by the health system) ill health must necessarily be funded by a public or private payer. Determining what is meant by "acceptable" or "normal" is a value judgement that may result in different conclusions in different societies. The range and intensity of health services that can be provided is inevitably constrained by the resources that are available and by how much a society or an individual is prepared to spend on healthcare rather than, say, education or pensions (Papanicolas and Smith, 2014).

Extending Culyer's definition to the population level, a population's need for healthcare can be understood as the minimum amount of resources required to exhaust a population's capacity to benefit. Population capacity to benefit PCB_i^k is defined here as the number of people in some population or region k with a specified condition-intervention pairing i which represents the capacity to benefit from some intervention given defined characteristics of the health state.

For the purpose of evaluating the appropriateness of rates of utilisation against population need for specified interventions, this definition has several advantages: (i) it is derived from the fundamental objective of a health system to produce benefit in some form (defined depending on a society's values e.g. in terms of potential gains in health or quality of life); (ii) it relates need for healthcare to the share of the distribution of ill health that is amenable to the range of available and cost-effective preventive, acute or chronic health services; and (iii) it provides a measure of need that is not contaminated by regional variations in the supply and demand for healthcare.

5.2.2 Relationship between need as capacity to benefit, healthcare supply and demand

As further discussed below and illustrated in Figure 5-1, it is worthwhile to highlight the relationship between need (as estimated by PCB), supply and demand for healthcare. PCB is not entirely independent of the supply of healthcare, in terms of the general stage of technological development. In particular, the presence of a

capacity to benefit requires the existence of a specific technology to treat the underlying condition. In turn, if new effective technologies are identified (or if existing technologies that were deemed effective are recognised to be ineffective based on new evidence), then PCB estimates will change. Thus, PCB is related to supply in the sense of a "general" existence of medical technologies to treat a condition.

However, provided that some effective medical technology exists "at large" (e.g. in the form of a national guideline as produced by NICE that is applicable across the entire country), then the number of people with a capacity to benefit from this technology does not depend on the local degree of access to or availability of this technology. Suppose, for instance, that there are no orthopaedic surgeons on the Isle of Skye to provide hip replacements. Clearly, this does not necessarily mean that no island resident might benefit from hip replacements.

The critical point for the purpose of this paper is that PCB estimates are independent of regional variations in levels of supply. This is important because, while regional levels of utilisation may depend to some extent on regional levels of supply, regional levels of medical need (as defined by the capacity to benefit) are not influenced by regional levels of supply (health professionals and capacity e.g. in terms of hospital beds).

Similarly, while levels of utilisation will certainly depend on levels of demand (in particular patient preferences for particular interventions), capacity to benefit is, in my view, a technical concept that requires no information on preferences. For instance, in order to state that some person has a capacity to benefit from hip replacement, one might require certain information on medical parameters (e.g. type and severity of osteoarthritis) as well as on quality of life (e.g. degree of pain as experienced by an individual). However, one does not require information on whether the individual "wants" a hip replacement. Such information on patient preferences is clearly essential for the decision to provide the intervention. It is irrelevant, however, for the judgment as to whether a "capacity to benefit" exists.

5.2.3 Capacity to benefit in populations: measurement and interpretation

To use estimates of PCB as a benchmark to identify unwarranted variation in healthcare, one faces three sets of problems:

1. Defining the condition-intervention pairing *i* in terms of measurable criteria of capacity to benefit from a specific intervention;
2. Estimating the number of new (incident) cases of *i* in a general population or region *k* that meet these criteria over a specified time interval (e.g. one year); and
3. Comparing the PCB estimate with actual utilisation over the same time period.

Each of these problems is characterised by measurement uncertainties. The first problem, increasingly the realm of Health Technology Assessment (HTA) agencies in many countries (Sorenson et al., 2008), requires criteria which accurately and reliably predict improvements in clinical outcomes or quality of life. The second problem requires epidemiological methods. In England, health authorities have since 1990 been mandated to assess the health and care needs of their local populations to inform priorities for improving population health (Wright et al., 1998). This has led to a field of health services research termed the epidemiology of indications (Frankel, 1991) which consists of identifying the prevalence and incidence of "cases of the condition where treatment would be indicated, tolerated, and desired by the particular sufferer, and also approved of in general as a proper use of the health budget" (Frankel, 1991: 258). Early on, however, it was pointed out that clinical uncertainty about outcomes, heterogeneity among patients in valuing these outcomes,[5] and effort in conducting high-quality epidemiological studies would likely limit the routine use of epidemiologically based needs assessment for healthcare planning and purchasing (McKee and Clarke, 1995, McKee, 1996).

[5] Even if one can predict with some certainty the probability of an outcome for an individual, each outcome may be valued differently. For example, two 65 year-old men with osteoarthritis facing a choice between hip replacement and continued pharmacological treatment who are fully informed about the benefits and risks of each potential outcome may choose differently because they place different values on each outcome.

If a credible approximation of PCB is possible, one finally requires reasonably complete and accurate data on utilisation. The needs-utilisation comparison could then be interpreted as follows:

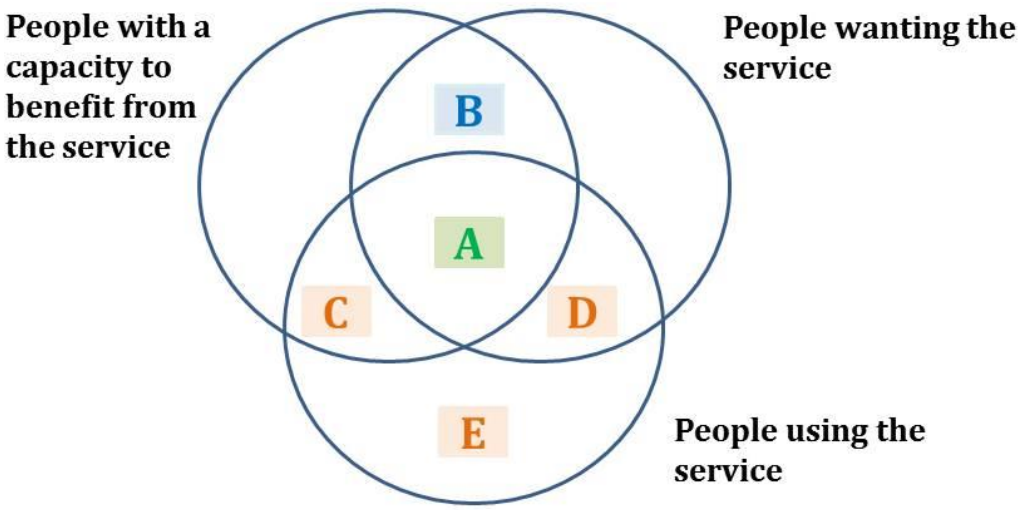
- If utilisation exceeds the PCB estimate, then this suggests overuse at the population level.
- If utilisation falls short of PCB, this suggests underuse of beneficial care.
- If utilisation and PCB are roughly equal, this suggests needs-based care.

For each of the three scenarios above, however, two additional factors are required in order to assess the degree of "overuse" or "underuse" in a system. The first factor is the appropriateness of clinical indications for care that has actually been provided. Audits of service utilisation are therefore necessary to ensure that overprovision to people for whom the intervention is not effective is not masked by simultaneous unmet need among those who would benefit (Schang et al., 2014a).

The second factor is patient preference. Unwarranted variation in healthcare utilisation – manifest in the form of either overuse or underuse – is defined as that part of variation that is not explained by population needs or patient preference (Wennberg, 2010). The PCB concept provides a normative measure of population need for defined interventions. It indicates the range and intensity of services a health system would be expected to *offer* to patients if evidence-based criteria of capacity to benefit were being applied (Schang et al., 2014a). Where individuals may place different values on the same outcome and the best choice of treatment is thus sensitive to patient preferences (Wennberg, 2010, Weinstein et al., 2007), it is not desirable to interpret this normative measure of population need as being equivalent to the "right (treatment) rate". The PCB estimate exceeds the appropriate rate of intervention by that margin of patients who, if eligible, make an informed choice for another option of treatment (including no treatment). Figure 5-1 depicts the relationships between the concepts of PCB, patient preference and utilisation and the resulting interpretations of underuse and overuse of defined interventions.

The aim of this literature review is to take stock of empirical studies that apply the PCB approach, guided by the following question: Can estimates of capacity to benefit in populations become a feasible and useful benchmark to quantify the discrepancy between actual utilisation and population need for a defined intervention?

Figure 5-1 Relationship between population capacity to benefit, patient preference and service utilisation



Legend:

A + B **= Population benchmark for appropriate utilisation ('right treatment rate'):** The intervention was clinically indicated and wanted.

A **= Actual utilisation that was appropriate (the intersection of need, preference and utilisation):** The intervention was clinically indicated, wanted and provided.

B **= Underuse:** The intervention was clinically indicated and wanted, but not provided.

C **= Overuse:** The intervention was clinically indicated and provided, but not wanted.

D **= Overuse:** The intervention was wanted and provided, but not clinically indicated.

E **= Overuse:** The intervention was provided, but neither clinically indicated nor wanted.

5.2.4 Comparison of the PCB concept with conventional needs indices and standardised utilisation measures

In the literature, various indices have been published to estimate need for health care for populations. One well-known example are the needs indices as they have been used in the National Health Service (NHS) in England in order to allocate resources to small areas (Carr-Hill et al., 1994, Sutton et al., 2002, Morris et al., 2007, PBRA Team, 2009).

The core feature that distinguishes the PCB measure from such indices is that it starts from explicit criteria of capacity to benefit. Conventional needs indices (e.g. Glover et al., 2004) are estimated by means of regression models where, firstly, a proxy measure of "need for care" in a population (e.g. admission rates, expenditure per patient in region k) is chosen. Secondly, the model seeks to explain variance across regions in the magnitude of this measure through a range of predictor variables (e.g. regional rates of long-term illness, deprivation etc.). Finally, expected rates of the chosen proxy measure of "need" for health care are produced for each region.

One of the key problems with this approach is that it does not account for whether rates of the chosen proxy measure of "need for care" in a population (e.g. admission rates, expenditure per patient in region k) was actually "needed" in terms of whether there was a capacity to benefit from care. The PCB measure, in contrast, starts from a set of criteria of capacity to benefit. These criteria are ideally derived from high-quality medical evidence about the effectiveness and safety about a particular procedure that has been shown to improve specified patient outcomes. The task is then to estimate the epidemiological prevalence of these criteria in a population.

The PCB concept differs from conventional needs indices and standardised utilisation measures also in a number of other ways. These differences are summarised in Table 5-1.

Table 5-1 Comparison of the PCB concept with conventional needs indices and standardisation of utilisation rates

	Standardisation	Population Capacity to Benefit	Needs indices
Purpose	**Performance measurement:** Adjustment for causes of regional variations that are not attributable to differences in health system performance	**Performance measurement:** Benchmark for the ***region-specific*** need for services	**Resource allocation** to small areas and/or providers responsible for populations (e.g. general practice populations)
Scope	**Single procedure or whole system**	**Single procedure** (which can be linked to a concrete capacity to benefit)	**Typically whole system:** general need for services across the totality of procedures and sectors of care
Nature of the standard	An arbitrarily chosen standard population (e.g. the national average)	The (absolute) numbers of people who would benefit from an intervention	A proxy measure of „need" for services (e.g. expenditure or rates of admission)
Guiding question(s)	Which rate of interventions can be expected if region *k* had the same [age-, morbidity-etc] distribution as the standard population?	How many people in region *k* have a 'capacity to benefit' from intervention *i* ?	Which „legitimate" predictors of „need" (e.g. age, deprivation) explain variations in the proxy measure of need, after adjusting for supply?
Model	Standardised utilisation = (utilisation in region *k*, utilisation in the standard population)	PCB_i^k = (criteria of capacity to benefit, population)	Expenditure = (need factors, supply factors, other variables)
Nature of variables included in the model	*No variables are required apart from a simple standardisation variable (usually age and/or sex)*	Variables represent criteria of capacity to benefit. These criteria are derived from existing clinical guidelines and/or HTA evidence.	Variables that explain a specified amount of the variance in the proxy measure of "need" and/or meet other criteria are retained.
Challenges	No benchmark for the *region-specific* need for services	Availability of criteria of capacity to benefit and epidemiological data Interpretation: Service use < PCB → *Suggests* underuse Service use > PCB → *Suggests* overuse Service use ≈ PCB → Assess appropriateness of clinical indications of care provided	No direct relationship to the capacity to benefit from services The chosen proxy variable of need (e.g. expenditure, admission rates) may be confounded by local supply and other factors: Disentangling the effects of needs factors

5.3 Methods

The PubMed/Medline, Scopus, Web of Science and Cinahl databases were searched using logical combinations (with the Boolean operator "AND") of the following search terms: "capacity to benefit" AND population; "healthcare needs assessment"; "needs assessment" AND "healthcare need"; "Population requirement"; "Healthcare requirement"; normative AND "treatment rate"; "needs assessment" AND healthcare AND population AND criteri*; "needs assessment" AND healthcare AND population AND indication.

The review included empirical studies published between January 1990 and 2015 that:

(i) defined criteria of capacity to benefit from a specific intervention; and
(ii) applied these criteria to estimate their prevalence or incidence in a defined general population.

Studies which assessed morbidity only or sought to identify treatment needs among patients in healthcare settings were thus excluded.

January 1990 was chosen as a cut-off date for two reasons. First, in 1990, the National Health Service Act in the United Kingdom (UK) for the first time in the UK and internationally mandated health authorities in the UK to assess the health and care needs of their populations (Wright et al., 1998). This year hence marks an important policy development likely to inspire research on this topic. Second, considering the advances in the development of HTA guidelines and epidemiological methods during the 1990s, earlier studies were deemed to be of less relevance to a present context.

The framework for data extraction focused on the basis for criteria of capacity to benefit; methods of epidemiological assessment; and any comparisons with actual utilisation. Within these categories, relevant sub-categories were developed from repeated reading and comparison of the identified PCB studies.

5.4 Results

22 studies published between 1995 and 2014 reported the results of empirical studies of population capacity to benefit (Figure 5-2). Table 5-2 shows the clinical areas covered and the countries of origin. The reporting of results (Tables 5-3 to 5-5) follows the framework for data extraction.

Figure 5-2 Literature review process

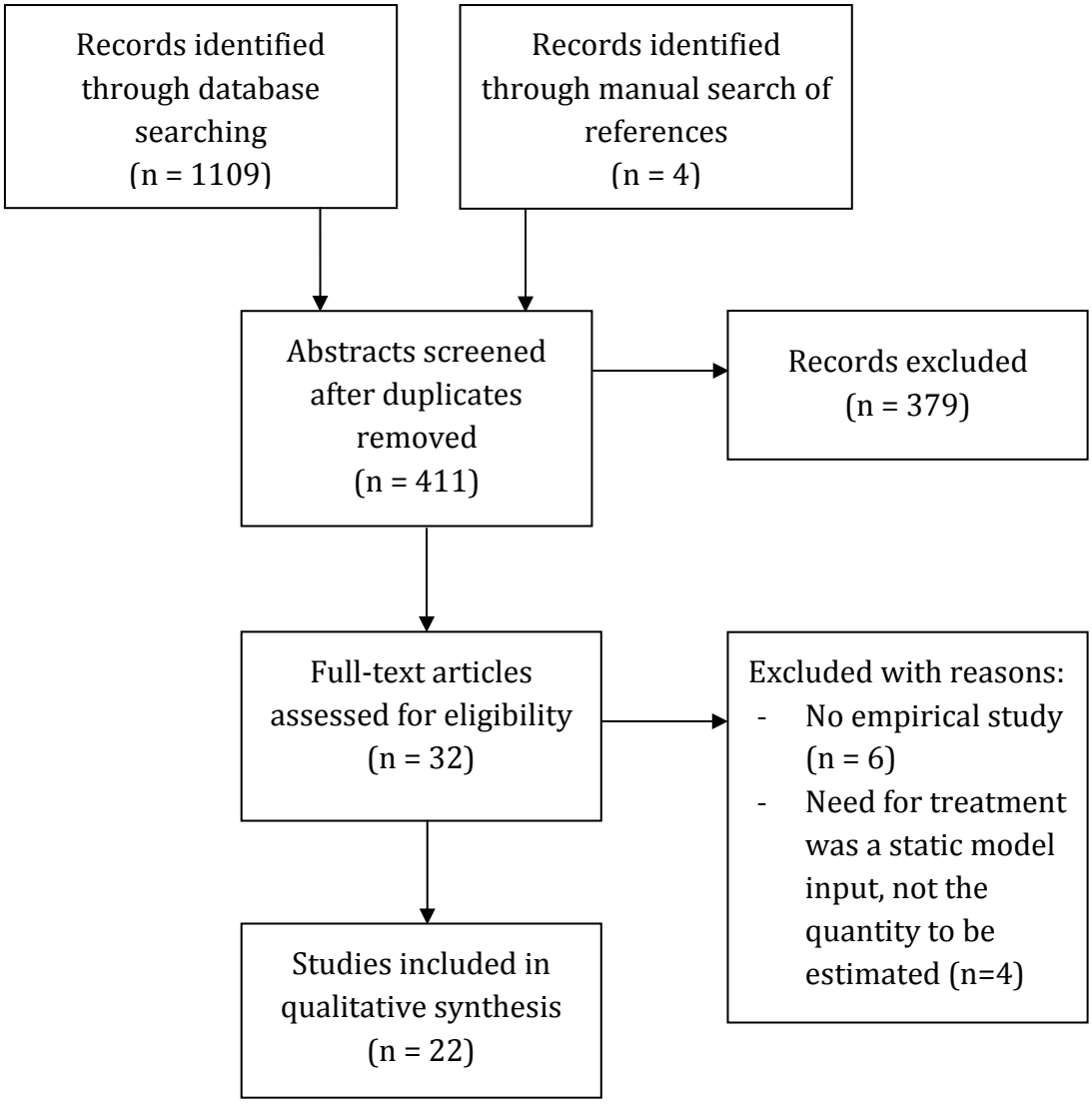


Table 5-2 Focus and origin of studies

CLINICAL AREA → INTERVENTION		STUDY
Osteoarthritis → hip/ knee replacement	8	Tennant et al. (1995); Fear et al. (1997); Frankel et al. (1999); Hawker et al. (2001); Jüni et al. (2003); Milner et al. (2004); Yong et al. (2004); Judge et al. (2009)
Cancer (various sites) → radiotherapy	6	Tyldesley et al. (2001); Foroudi et al. (2003); Delaney et al. (2005); Usmani et al. (2005); Jacob et al. (2010); Fong et al. (2012)
Urinary symptoms → prostatectomy	2	Sanderson et al. (1997); Treagust et al. (2001)
Cataract → cataract surgery	1	Frost et al. (2001)
Coronary artery disease → revascularisation	1	Martin et al. (2002)
Dental disease → extractions, dentures, restorations	1	Guiney et al. (2012)
Stroke → preventive, acute, rehabilitative services	1	Hunter et al. (2004)
Stroke → carotid endarterectomy	1	Ferris et al. (1998)
Otitis media with effusion → ventilation tube surgery	1	Schang et al. (2014a)
COUNTRY		**STUDY BY FIRST AUTHOR**
United Kingdom	13	Frankel Frost Schang Jüni Judge Tennant Treagust Sanderson Martin Ferris Milner Fear Yong
Canada	5	Hawker Hunter Usmani Foroudi Tyldesley
Ireland	1	Guiney
Australia	2	Delaney Jacob
Cross-national: Australia, Canada, Scotland	1	Fong

5.4.1 Defining criteria of capacity to benefit

Studies from the 1990s and early 2000s struggled with the lack of agreed criteria to predict patient benefit from an intervention. Three studies sought to address this by using clinical consensus panels (Table 5-3). For instance, Martin et al. (2003) asked panellists to rate the likelihood of benefit for different indications for coronary

revascularisation based on a nine-point scale in relation to existing trial evidence. Nine studies used multi-dimensional composite indices such as the New Zealand score which ranges from 0 to 100 and summarises subscores representing degree and occurrence of pain experienced by patients, functional limitations, movement and deformity, and threats to independent daily living. Eight studies relied on guidelines published by professional associations. Two of these included as an additional criterion the judgement by clinical examiners conducting the assessment that the individual in front of them would likely benefit. Only the two most recent studies from 2012 and 2014 (Fong et al., 2012, Schang et al., 2014a) applied criteria of appropriateness recommended by an independent HTA institute, the National Institute for Health and Care Excellence (NICE) in England.

The criteria of capacity to benefit differed between studies for a given clinical area and in part evolved over time. For example, while earlier studies excluded individuals with co-morbidities from hip replacement, a more recent study by Judge et al. (2009) argued that progress in modern anaesthesia and surgical techniques enabled these individuals to in principle benefit from surgery as well.

Due to uncertainty about the "correct" criteria of capacity to benefit, six studies undertook sensitivity analyses to examine the impact of different model assumptions on the results. One of these studies (Sanderson et al., 1997) applied alternative sets of criteria. Among five criteria used to define need for prostatectomy (history of retention, comorbidity, symptom type, symptom severity and symptom bothersomeness), the authors examined the impact of excluding "symptom bothersomeness" from the list of criteria. Due to uncertainty about the threshold for intervention on a given criterion, six studies applied a range of plausible values. For instance, as no agreed threshold values existed for the New Zealand score, Frankel et al. (1999) chose cut-offs of 43 and 55 points to reflect moderate and severe disease, respectively.

5.4.2 Estimating the number of incident or prevalent cases

Only seven studies empirically assessed treatment needs directly in the population under study. The rest extrapolated epidemiological information from other populations (Table 5-4). For instance, Frost et al. (2001) assessed population requirements for cataract surgery in the Avon and Somerset region in England and then applied age-sex-specific rates of need to the national level. Ferris et al. (1998) applied age-sex-specific rates of ischaemic stroke incidence from the Oxford community stroke project and a set of conditional probabilities that would indicate the subgroup of people with capacity to benefit from carotid endarterectomy (e.g. the proportion of strokes with stenosis given that they are in the carotid territory) to estimate district-level rates of need in the former Wessex Regional Health Authority area.

The variables used to extrapolate model parameters or estimated rates of need were limited to age and sex in all studies, with one exception: Judge et al. (2009) employed a nationally representative survey to identify predictors of need for hip replacement, as measured by a modified version of the New Zealand score. Statistically significant predictors (age, sex, Index of Deprivation quintiles, rurality and ethnic mix of the area) were then replicated in local census data to predict need for hip replacement surgery at district level in England. For some significant predictors of need (obesity, an individual's social class), however, no identically defined variables were available at district level, thus limiting the use of a more comprehensive set of predictors of need.

Five of the fifteen studies that did not assess needs directly performed sensitivity analyses to estimate the impact of consequent uncertainty. Methods included using a range of highest and lowest plausible estimates, Bayesian simulation and Monte Carlo simulation to arrive at credible intervals of population need.

Fifteen studies estimated incident requirements of need for an intervention (the rest assessed prevalent need only). Ten of these used longitudinal data from cohort studies or registries. Five studies converted prevalence rates obtained from cross-

sectional studies into incidence rates, using cohort simulation (Sanderson et al., 1997) or mathematical models of the relationship between incidence and prevalence (e.g. Frankel et al. 1999).

All studies used data primarily from the general population, by means of population-based screening studies, population registries or surveillance systems, national or local surveys. Based on either full population coverage or a stratified random sample, estimated incidence or prevalence rates were intended to be representative of the general population's age and sex distribution. The national survey used by Guiney et al. (2012) was also representative in terms of means-tested income. However, response rates of less than 100% with the possibility of data missing not at random created a potential threat to the representativeness of the sample. Most studies discussed the socio-demographic representativeness of the sample and possible sources of bias due to over- or underrepresentation of some population groups. One study (Jüni et al., 2003) imputed New Zealand (NZ) scores (which were used as the criterion of capacity to benefit from knee replacement) for participants with incomplete data. This means that each missing NZ score was replaced by a predicted score. The predicted score was obtained from a linear regression model which estimated the final NZ scores among participants with complete data from the NZ subscores on disability, pain, ability to live independently and multiple joint disease as predictor variables. To mitigate distortions in estimates of population need for knee replacement, calculations then included both the reported scores of participants with complete data and the predicted scores of participants with incomplete data.

In the absence of population data for some model parameters, eight studies relied on utilisation data. For instance, Martin et al. (2002) extrapolated the incidence of unstable angina from hospital admission rates, assuming a 100% referral rate.

5.4.3 Comparing estimates of need and utilisation

Eleven studies compared estimates of incident need with actual utilisation (the rest estimated need only; examined the prevalent "backlog" of need in the system; or examined primary and specialist consultations among the population with capacity to

benefit from an intervention; Table 5-5). Where guidelines recommended the intervention as one option among others, eight studies sought to adjust for patient preferences in determining the appropriate level of utilisation. These studies focused on classic preference-sensitive conditions (Wennberg, 2010) including hip or knee replacement, prostatectomy and radiotherapy. Five studies, which were based on primary survey data, asked participants about their willingness to receive the intervention if eligible. Two studies applied proportions of women preferring radiotherapy over mastectomy or breast conserving surgery from previously published Canadian research to the authors' Australian study context. One study adjusted for the proportion of eligible patients who had actually refused surgery as recorded in the same cancer registry that was also used to estimate the number of incident cases.

If one were to allow for a margin of error of 20% (chosen here merely for illustrative purposes), then the studies show a range of findings where the PCB estimate exceeds, falls short of or is roughly equal to the actual utilisation rate. Some studies fall in more than one category for different sets of modelling assumptions. For instance, Frankel et al. (1999) found an excess of 12% to 49% in incident indications for primary hip replacement (for a New Zealand score of 55 vs 43, respectively) over annual rates of surgery in England. However, adjusting for estimates of patient willingness to undergo surgery reduced the estimated excess of potential need over actual utilisation to 3% and 11%, respectively. Considerable reductions in estimated levels of population need following an adjustment for patient preference were also found by Hawker et al. (2001) for arthroplasty and Sanderson et al. (1997) for prostatectomy.

Hunter et al. (2004), the only study that modelled the entire spectrum of preventive, diagnostic, acute and rehabilitative services for stroke, found a potential underuse of prevention programmes alongside apparent overuse of carotid endarterectomy for acute stroke and a close match between need and utilisation of pharmacological treatment for hypertension (a risk factor for stroke) in Ontario, Canada.

Table 5-3 Defining criteria of capacity to benefit

2.1 BASIS FOR CRITERIA OF CAPACITY TO BENEFIT		STUDY BY FIRST AUTHOR
Evidence-based guidance developed by an independent HTA agency	**2**	Fong* Schang
Guidance endorsed by professional associations or national clinical networks	**8**	Delaney Ferris Fong* Foroudi Guiney* Hunter Jacob Treagust* Tyldesley
Additive composite indices related to osteoarthritis (New Zealand Score, WOMAC, Lequesne index) or urinary symptoms (North West Thames Symptom Index)	**9**	Fear Frankel Hawker Judge Jüni Tennant Sanderson* Milner Yong
Threshold levels chosen to reflect moderate and severe disease (modelling assumption)	**8**	Fear Frankel Judge Jüni Milner Tennant Sanderson Yong
Threshold levels derived from current practice to reflect severe disease	*1*	Hawker
Clinical panel rating the appropriateness of intervention for individual patients, in relation to existing trial evidence	*3*	Martin Usmani Sanderson*
Judgement by clinical examiners conducting the epidemiological assessment that individual would likely benefit	**2**	Guiney* Treagust*
Not specified in the paper	**1**	Frost
2.2 SENSITIVITY ANALYSIS	**6**	
Use of alternative sets of criteria	*1*	Sanderson
Use of different threshold levels	*6*	Frankel Frost Judge Jüni Martin Sanderson

* Study falls into more than one category.

Table 5-4 Epidemiological assessment

3.1 OVERALL STUDY DESIGN		STUDY BY FIRST AUTHOR
Needs assessment within the population under study	**7**	Fear Hawker Guiney Milner Tennant Treagust Yong
Extrapolation	**15**	
Of estimated rates of need: Age-sex-specific population requirements were assessed directly in a local population and then extrapolated to the national level	*2*	Frankel Frost
Of model parameters:		
Age-sex-specific rates of disease incidence and conditional probabilities defining the group with capacity to benefit from treatment were taken from other populations	*12*	Delaney Ferris Fong Foroudi Hunter Martin Jacob Jüni Schang Sanderson Tyldesley Usmani
Multiple predictors of need for hip replacement were estimated using a nationally representative survey, then replicated at district level using census data	*1*	Judge
3.2 SENSITIVITY ANALYSIS IN STUDIES USING EXTRAPOLATED DATA		
Deterministic scenario or multiway sensitivity analysis using a range of highest and lowest relevant estimates	3	Foroudi Sanderson Usmani
Bayesian simulation to estimate credible intervals for small-area predictions of need for hip replacement	1	Judge
Monte Carlo simulation of joint uncertainty in parameter estimates	2	Delaney Schang
3.3 CASE ASCERTAINMENT IN ORIGINAL DATA SOURCE		
General population study (stratified random sample or full population coverage):		
A. Clinical screening of all people in defined age groups or of an age-stratified random sample living in an area	1	Schang
B. Mandatory population registry or	9	Delaney Fong Foroudi Jacob Tyldesley Usmani

surveillance system (notifiable cancers, stroke, myocardial infarction)		Ferris Hunter* Martin*
C. Nationally representative survey: self-reported risk factors, symptoms, co-morbidities	2	Hunter* Judge
D. Nationally representative survey: standardised interview and clinical assessment	1	Guiney
E. Locally representative survey: self-reported symptoms followed by targeted clinical assessment (e.g. of people reporting pain in their hip)	5	Frankel Frost Hawker Jüni Treagust
G. Locally representative survey: self-reported symptoms only	6	Fear Milner Sanderson Tennant Treagust Yong
Some parameters estimated from	**7**	
H. General practice statistics (incidence of stable angina)	*1*	Martin*
I. Hospital admission rates (incidence of unstable angina)	*1*	Martin*
J. Hospital-based registries (incidence of cancer sub-groups)	*5*	Delaney Fong Jacob Tyldesley Usmani Foroudi

3.4 TIME INTERVAL STUDIED

A. Incident need in a population	**14**	
A.1 Data from prospective cohort studies or disease registries	*10*	Delaney Ferris Foroudi Fong Hunter Jacob Martin Schang Usmani Tyldesley
A.2 Data from cross-sectional studies, with prevalence converted into incidence rates using a cohort simulation or the method by Leske et al. (1981)	*4*	Frankel Jüni Judge Sanderson
B. Prevalent need in a population (cross-sectional assessment only)	**8**	Fear Frost Guiney Hawker Milner Tennant Treagust Yong

3.5 REPRESENTATIVENESS OF THE ORIGINAL DATA SOURCE/ TREATMENT OF MISSING DATA

Study or original data source (for studies using extrapolated data) reported on demographic characteristics of the sample and response rates.	**22**	All studies

A. Authors did not comment on threats to representativeness.	3	Ferris Hunter Martin
B. Authors (or the original studies they used) concluded the sample was sufficiently representative of the general population studied in terms of age and sex distribution.	10	Delaney Fear Frost Guiney Jacob Judge Milner Tennant Schang Yong
C. Authors provided a qualitative discussion of areas where the sample was under- or overrepresenting particular groups and the likely impacts on the results.	8	Fong Foroudi Frankel Hawker Sanderson Treagust Tyldesley Usmani
D. Authors performed a quantitative correction using imputation to account for individuals with incomplete data.	1	Jüni

* Study falls into more than one category.

Table 5-5 PCB-use comparison

4.1 ADJUSTMENT FOR PATIENT PREFERENCES	8	STUDY BY FIRST AUTHOR
Based on self-reported willingness to undergo the procedure (hip or knee replacement, prostatectomy, cataract surgery) among patients surveyed by the study authors	*5*	Frankel Frost Hawker Jüni Sanderson
Based on a retrospective application of an estimate of the proportion of women willing to undergo radiotherapy rather than mastectomy or breast conserving surgery from published research	*2*	Fong Tyldesley
Based on the number of patients who were actually offered and refused the procedure (radiotherapy) in a clinical setting as recorded in a population cancer registry	*1*	Usmani

4.2 FINDINGS		
Incident need to actual utilisation	**11**	
PCB > actual utilisation by 20 % or more	*10*	Delaney Ferris* Fong Guiney Jacob Martin Schang Frankel* Jüni Hunter*
PCB < actual utilisation by 20 % or more	*2*	Hunter* Jacob*
PCB ≈ actual utilisation within a margin of less than 20%	*2*	Ferris* Frankel* Hunter*
Prevalent need to actual utilisation	**4**	
Potential "backlog": prevalent PCB > annual number leaving the prevalence pool through operation or death		Frost Hawker Milner Yong
Incident and/or prevalent need and health service use other than the intervention for which need was assessed	**5**	
Patients were asked if they saw a general practitioner and/or specialist this year		Fear Jüni Milner Tennant Yong
Estimation of prevalent or incident need only	**6**	Foroudi Judge Sanderson Treagust Tyldesley Usmani

* Study falls into more than one category.

5.4.4 Comparing estimates of need and age-standardised rates: example for otitis media

Figure 5-3 illustrates the insights to be gained from estimating PCB over conventional age-standardised rates alone. The data come from a study by the author which sought to estimate the number of children with capacity to benefit from ventilation tubes (VTs) for otitis media with effusion (OME) in England if intervention criteria recommended by NICE were being followed (Schang et al., 2014a). VTs are widely thought to be overused (Audit Commission, 2011) and clinical audits confirm that a large proportion of VTs are not provided in compliance with NICE criteria (Daniel et al., 2013). The study however found that, at the same time, there appeared to exist a potential net underuse of VTs at the population level, meaning that children who had a capacity to benefit from VTs (as defined by NICE criteria) were not treated.

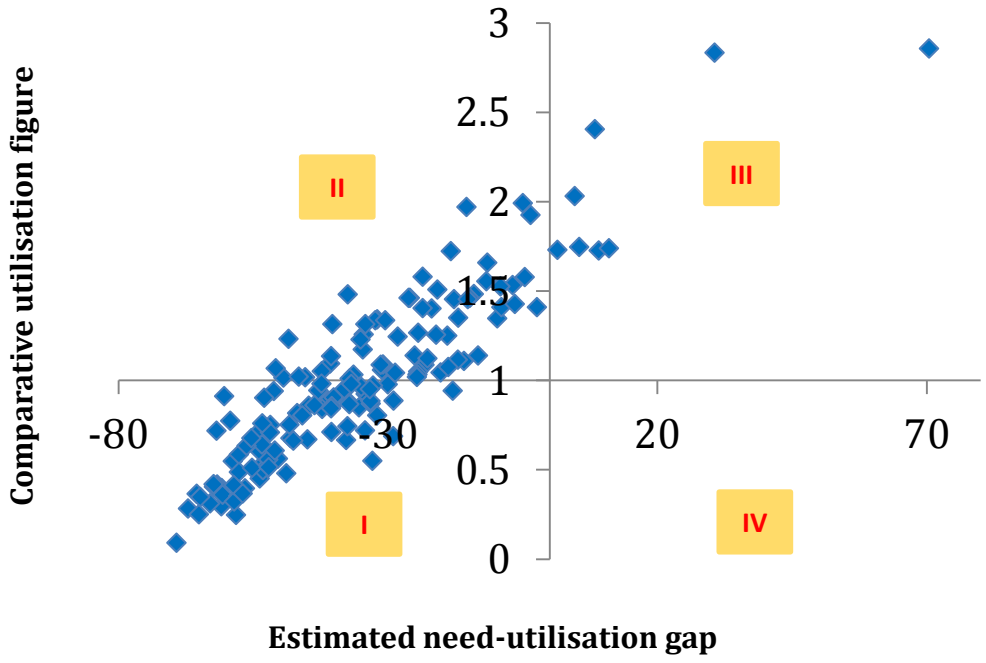
The y-axis shows the comparative utilisation figure (CUF), which is calculated by dividing the directly age-standardised rate of VT surgery of each of the 151 Primary Care Trusts (PCTs) by the national average as the standard population. Given their age composition, PCTs with a CUF of "1" thus have a comparable rate of VT surgery as the national average. The x-axis shows the difference between observed rates of VT surgery and expected capacity to benefit from VTs for each PCT population (see Appendix 5-A for the calculations and Schang et al. (2014a) for details of the underlying methods). PCTs in Quadrants I and II exhibit a deficit in relation to their local population's capacity to benefit, suggesting underuse. PCTs in Quadrant III display an excess of VTs over the local estimate of need, indicating net overuse.[6]

The positive correlation between the CUFs and the magnitude of the need-utilisation gaps ($r = 0.693$; $p < 0.001$) suggests that PCTs with a low age-standardised rate of VT insertions are also likely to have a larger need-utilisation gap. If the PCB model provides a valid estimate of population need, then this finding is reassuring because it implies that, in this particular case, standardised utilisation rates and the independent measure of population need point into the same direction. However, there are two important caveats. First, for reasons discussed in Schang et al. (2014a),

[6] Note: here only the mean estimates are shown for illustrative purposes.

age was used as the sole predictor of regional incidence of OME. This is a limitation of the PCB model because it means that other potential risk factors for OME were not taken into account. The consequent uncertainty was, however, examined through Monte Carlo simulation. Second, even though the PCB model is imperfect, it does provide additional information. PCTs in Quadrant II are above the national average (based on which one might be tempted to suspect overuse), but the PCT-specific estimate of capacity to benefit nonetheless suggests net underuse at a population level. While the CUFs might lead one to focus solely on the vertical distribution of top and bottom outliers, the model of need reframes the reference point for identifying overuse and underuse and shows that even in PCTs which are close to the national average there may be a gap between utilisation and need.

Figure 5-3 Comparative utilisation figures and estimated need-utilisation gaps by Primary Care Trust, England, four-year average 2007-2010



Legend:

y-axis: Directly age-standardised rate of ventilation tube insertions for OME divided by the England national average as the standard population.

x-axis: Locally-specific difference between the rate of ventilation tube insertions to expected population capacity to benefit (PCB) at Primary Care Trust level, per 10,000 children aged 2-8 years.

Source: author's work based on the methodology described in Schang et al. (2014a).

5.4.5 Practical relevance for health service planning: comparing PCB with a "simple" model of need

PCB models require considerable information which must be collected from the literature or from other data sources. This raises the question whether estimating a full PCB model adds sufficient value and gains in information to aid decision-making, or whether a "simple" needs index might also be adequate.

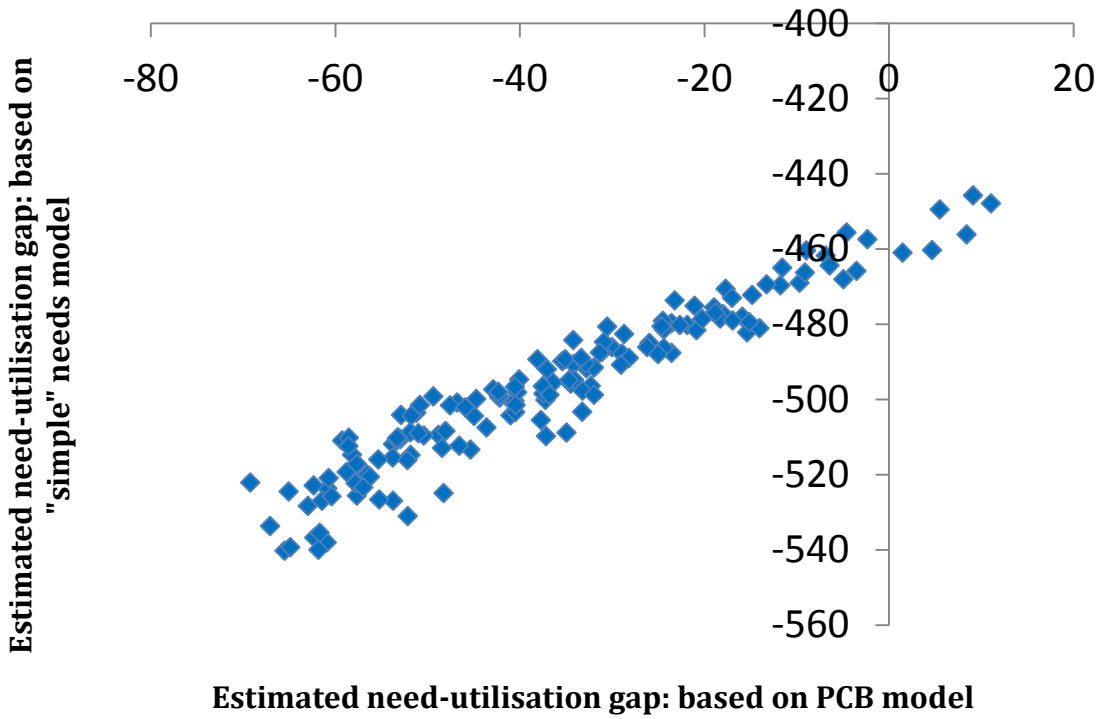
Figure 5-4 provides a comparison of estimated need-utilisation gaps based on the PCB in comparison to a "simple" needs model. This "simple" needs model is essentially based on the age- and region-specific incidence of OME which persists five months after the initial onset (Appendix 5-1). In other words, this model does not account for the additional criteria of benefit specified by NICE guidance and their epidemiological distribution in a specific population (hearing loss at a specified threshold level and bilateralism of the disease).

As Figure 5-4 shows, both models of need are strongly correlated ($r = 0.964$; $p < 0.001$). This is not surprising since the full PCB model includes all the parameters of the simple model but, in addition, accounts for the distribution of hearing loss at a specified threshold level and bilateralism of the disease. The probability distributions assigned to these two parameters (further explained in Chapter 4, Table 4-2) have a narrow spread, large emphasis on the mean estimate and are assumed to be identical for all regions considered here. This means that the additional two parameters act, essentially, as a scaling factor.

The comparison of the two models shows that "simple" model of need will grossly overestimate the number of people with a capacity to benefit from VTs for OME. While the mean utilisation-needs gap across all PCTs for the full PCB model is estimated to be about -35 (indicating that 35 per 10,000 children in an average PCT would have benefitted from VTs but did not receive them), this gap is -535 for the "simple" needs model. Thus, the "simple" needs model overestimates considerably the number of children in "need" of VTs for OME by a factor of more than 15-fold.

The critical contribution of the additional two parameters included in the PCB model is, therefore, to help narrow down the population capacity to benefit. For health service planning, such information is essential since the "simple" model of need appears to produce blatantly unrealistic estimates.

Figure 5-4 Estimated need-utilisation gaps based on the PCB in comparison to a "simple" needs model, by Primary Care Trust, England, four-year average 2007-2010



Legend:

y-axis: Locally-specific difference between the rate of ventilation tube insertions to a "simple" model of need at Primary Care Trust level, per 10,000 children aged 2-8 years.

x-axis: Locally-specific difference between the rate of ventilation tube insertions to expected population capacity to benefit (PCB) at Primary Care Trust level, per 10,000 children aged 2-8 years.

5.5 Discussion

The conventional approach to account for population needs in research on geographic variation in healthcare – standardisation of rates for variables associated with need – is easy to implement, but fails to indicate whether the range and intensity of services expected to be beneficial for a population have been met. Grounded in a health economic conception of need for healthcare, the concept of population capacity to benefit provides such a benchmark, but its measurement is challenging. The following sections discuss the feasibility and potential usefulness of estimating PCB in light of the literature review.

5.5.1 Uncertainty about criteria of capacity to benefit

As noted in section 5.2, the estimation of PCB requires measurable criteria indicating the groups of patients for whom an intervention is likely to do more good than harm. For interventions that are known to be highly effective, whose target groups can be clearly defined (e.g. in terms of age) and where individual choice is problematic from a public health perspective, this may be straightforward. A good example would be measles vaccination, since vaccination opponents endanger not only their own but also other people's health as a result of reduced protection afforded by herd immunity (Anderson and May, 1990).

Where criteria of capacity to benefit are absent or controversial, there seem to be three possible routes. A first route is to introduce a more holistic, patient-centred element into the assessment of healthcare needs. In two studies reviewed here, focused on dental care (Guiney et al., 2012) and prostatectomy (Treagust et al., 2001) clinical examiners were asked to interpret guideline criteria in the light of individual cases when assigning the label of a "capacity to benefit" to a study participant. This seems to reflect an attempt to allow for health professionals' tacit knowledge (Polanyi, 1966) which cannot be expressed by explicit criteria. It is possible where information on healthcare needs is collected directly in the population studied

(rather than extrapolated from other populations). However, this deliberate lack of standardisation limits transparency and introduces non-random bias in achieving the policy goal of equal opportunity of access for equal need. This route also follows a more paternalistic approach to defining capacity to benefit which does not necessarily account for a patient's preferences.

A second route, taken by eight of 22 studies in this review, is to acknowledge the uncertainty in our current state of knowledge and to examine the sensitivity of the results to different modelling assumptions. For many surgical procedures, the threshold above which the intervention produces net benefit is uncertain (McKee, 1996). Where cases just below or above the threshold may get marginal benefit, it seems reasonable to aim for a range of PCB estimates depending on different choices of criteria and thresholds rather than a single "right rate". This may provide valuable information for the planning and evaluation of health services if it is performed for purposes of learning (using PCB estimates as a "tin opener") rather than for command-control styles of management (using PCB estimates as a "dial"; see Carter (1989)).

A third route – and medium to long-term task – is to invest in structures that foster the development of better predictors of benefit from health interventions for a range of population subgroups. Opportunities to achieve this arise from the growing remit of HTA agencies in many countries to conduct comparative and cost-effectiveness assessments (Sorenson et al., 2008). Although HTA agencies have so far largely focused on pharmaceuticals, they have also accelerated the development of evidence-based guidance for surgical procedures, medical technologies and, at a much slower pace, public health interventions (Sorenson et al., 2008). In some countries, this has widened the range of methods used beyond classic RCTs, and their well-known limits in ascertaining the real-world effectiveness of interventions (Teutsch et al., 2005), towards pragmatic RCTs for complex interventions (Campbell et al., 2000) and cohort studies in primary care (Whitehurst et al., 2015). These developments create scope to overcome some of the hurdles faced by studies in the 1990s and early 2000s.

Clearly, the criteria of PCB depend on current medical knowledge about the benefit from and the effectiveness of specific procedures for defined conditions. As the state of knowledge evolves, so may criteria of PCB change. It is therefore not surprising that, given advances in evidence-based medicine and HTA, criteria of PCB differ over time (and also across the studies examined which cover a time span from the mid-1990s to 2014). This is not necessarily a drawback of this specific method but rather a feature of healthcare in general.

5.5.2 Morbidity statistics: accuracy, completeness and routine availability

Assessing healthcare needs directly in the population studied clearly has the highest validity. However, as reflected in the small number of studies which did this, well-designed epidemiological studies are expensive. The extrapolation of prevalence and incidence rates from other populations, adopted by the majority of studies in this review, may offer a pragmatic alternative but raises questions about external validity. That is, to what extent can one be confident that age-specific incidence rates of, say, osteoarthritis in Avon and Somerset in 1999 also apply to other settings or time periods? While recommendations exist to judge the transferability of economic evaluations across jurisdictions (Drummond et al., 2009), the STROBE statement (which seeks to strengthen the reporting of observational studies in epidemiology) has been criticised for its lack of detail about how to report on external validity (Burchett et al., 2011). The STROBE editors have recently started to address this by elaborating on requirements and examples for providing contextual information, but they suggest that external validity ultimately remains a matter of judgement (Vandenbroucke et al., 2014). Further research on this problem is important because the relationship between health, age and other factors may vary between populations and is unlikely to be fixed over time (Mason et al., 2015).

Most diseases are multifactorially determined. While some authors of studies that extrapolated epidemiological information noted they had taken data from similar populations, the underlying components of “similarity“ were rarely defined. Studies that applied age-sex-specific incidence rates thus made the implicit assumption that

no substantial distortion was caused by known and unknown confounders such as differences in ethnic and socio-economic mix, climate or genetic disposition. This assumption may not be true. Only five studies attempted to gauge the nature and direction of potential bias and its impact on the results through sensitivity analysis. This seems a missed opportunity. While modelling results can only be as good as the underlying data, methods to handle uncertainty exist (for instance in the quantitative risk analysis literature, see Morgan and Henrion, 1990, Frey and Patil, 2002, Cox, 2007, Cox, 2012) and merit adoption in future studies.

The only study that represented multiple predictors of need beyond age and sex (Judge et al., 2009) is part of a growing body of small-area estimation methods in the geography literature. In healthcare, these methods seem to have largely been employed for the spatial microsimulation of health behaviours and health needs (rather than healthcare needs) (Gibson et al., 2002, Smith et al., 2011, Whitworth, 2013). Where nationally representative surveys and identically defined local census data are available, these methods may enable the use of a wider set of predictors of healthcare needs.

Self-reported morbidity from surveys, employed by ten studies in this review (categories 3.3 C, E and G in Table 5-3) can be useful where criteria of capacity to benefit include patient-perceived symptoms such as the degree of joint-related pain as a factor in determining the appropriateness of hip replacement. However, self-reports have well-known limitations where underreporting is an issue, such as with hearing problems among children (Schang et al., 2014a). Surveys will also tend to indicate only the prevalence of symptoms indicative of need for treatment rather than their incidence over a specified time period, which is necessary for evaluating the provision of services over the same interval. Although it is possible to estimate incidence from prevalence (Leske et al., 1981) and this was done by five studies in this review, the underlying models rely on several assumptions about the irreversibility of the disease and the constancy of disease and death risks over time (Frankel et al., 1999).

Potential sources of routinely available information on the incidence of diseases are notifiable disease surveillance systems and registries. In Europe, their development has however largely been limited to selected diseases including infectious diseases and cancer, respectively (Rosenkötter and van Bon-Martens, 2015). In the absence of data from the general population, some studies in this review derived incidence and prevalence rates from diagnosis and prescription data or from rates of hospital admission. Until complete population-based morbidity statistics become available, triangulation of the consistency of morbidity data held by healthcare providers, insurance funds and disability allowance registers may seem a pragmatic way forward (Eurostat, 2014). However, utilisation-based data remains inherently biased where people who would benefit do not access healthcare settings in the first place, are misdiagnosed or not diagnosed at all (Asthana and Gibson, 2011). "Person consulting prevalence" (the number of people consulting at least once with a specific morbidity during a defined period of time such as a quarter; Jordan et al. (2007)) is therefore unlikely to represent true population prevalence of a specific disease.

5.5.3 Need-use discrepancy analysis: towards a health system perspective?

The studies reviewed here provided little to no information on the context surrounding an apparent need-utilisation gap. Because overuse and underuse may co-exist, one important way to detect a mismatch of need and utilisation at the individual level is to conduct audits of the appropriateness of care provided in tandem with population-based studies (Schang et al., 2014a).

As shown in Figure 5-1, patient preference is conceptually distinct from need for healthcare as measured by the capacity to benefit. However, the observation in several studies (Sanderson et al., 1997, Hawker et al., 2001, Jüni et al., 2003) that adjusting for an estimate of the proportion of patients willing to undergo elective surgery (here, arthroplasty or prostatectomy) led to considerable reductions in the level of care that would be both clinically indicated and wanted shows that PCB estimates can hardly be interpreted in isolation of prevailing values among patients. If one seeks to evaluate the degree to which service provision is aligned with both population needs and patient preferences, then future research should widen the

analysis beyond a single procedure towards all treatment options for a given health state. PCB estimates would thus indicate the potential maximum workload for each setting of care (specialist, primary or community care) while an estimate of the distribution of patient preferences (alongside other demand- and supply-side factors) would help to understand where and why rates of utilisation exceed or fall short of expected levels of population capacity to benefit.

From a methodological perspective, most studies that sought to adjust for patient preference did so by asking survey respondents about their stated willingness to receive the intervention. A range of other methods exists to measure stated or revealed patient preferences, each with distinct challenges in terms of internal consistency, ability to represent "true" preferences beyond hypothetical choice scenarios, and ability to understand the determinants of patient preferences (Bridges et al., 2007). Preferences may depend on various socio-demographic characteristics, psychological and cultural dispositions, and the ways in which a medical choice problem is framed (Edwards et al., 2001). Preferences are also likely to evolve with changes in technology, in wider society and in an individual's health state (Ditto et al., 2006, Fried et al., 2006). One hence needs to be cautious about transferring estimated proportions of patients willing to undergo a procedure from different contexts and time periods, as it was done by two studies in this review (Fong et al., 2012, Tyldesley et al., 2001).

5.6 Conclusions and policy implications

Returning to the question asked in this review – can estimates of capacity to benefit in populations become a feasible and useful benchmark to quantify the discrepancy between actual utilisation and population need for a defined intervention? – two main conclusions can be drawn.

First, several peer-reviewed studies have attempted to translate the elusive concept of "population need for healthcare" into an empirical measure of population capacity to benefit. While each model is a simplified representation of reality (Pidd, 1996),

each of the studies is based on assumptions that can be questioned and revised. The synopsis contributed by this literature review suggests a range of feasible routes how population benchmarks may be approximated.

Second, PCB estimates may usefully complement conventional standardised rates in providing a substantively meaningful reference point for comparative performance assessments. While gaps remain in terms of credible criteria of capacity to benefit and valid population-based incidence rates, the concept of PCB is operationalised using established methods from HTA and epidemiology and both disciplines have made great strides over the past years. In health systems which strive to evaluate the effectiveness and efficiency of care ultimately in terms of the benefit conferred on populations, there is scope in bringing these two fields of health services research closer together in the future.

This suggests several implications for policy. To enhance the robustness of estimates of PCB, it will be important to strengthen the mandate of HTA agencies and independent scientific associations so that standards of good practice can be established. Accurate and complete data from the general population is required to estimate the incidence of population need for defined interventions. National agencies should consider focusing the development of population benchmarks on top resource-intensive conditions and use these to support regional healthcare planning and surveillance. An example of how this might be done comes from England, where NICE has started to develop population benchmarks for some major conditions such as care for chronic obstructive pulmonary disease (NICE, 2011) and cardiac rehabilitation (NICE, 2013). Finally, computations of PCB should be coordinated with clinical audits and with patient preference assessments in order to improve the match between population needs, individual preferences and healthcare utilisation.

5.7 Acknowledgements

The author gratefully acknowledges helpful comments from Alec Morton and Thomas Schang on an earlier version of this Chapter.

5.1 Commentary in relation to the organising matrix of strategies to address ambiguity about the standard for evaluation

This Chapter has explored how policy-makers and managers might establish meaningful standards for evaluation with a technical-evidential approach. To this end, the Chapter has defined the concept of population capacity to benefit (PCB) as a potential benchmark for population need for defined interventions. On this basis, the Chapter has critically reviewed the feasibility and utility of measuring PCB, its generalizability across conditions and persisting challenges.

5.2 Appendix 5-A. CUF and PCB calculations

The comparative utilisation figure is the ratio of the number of operations that would be expected in a local population if it had the same age structure as the standard population, applying the stratum-specific local population rates to calculate the expectation, divided by the number of operations in the standard population (adapted from the more commonly known comparative mortality figure, see Breslow and Day, 1987: 53-63):

$$CUF = \frac{\sum_{j=1}^{J} n_j^* \frac{d_j}{n_j}}{D^*}$$

Where d_j is the number of operations in the *j*th of *J* age groups of the local population
n_j is the number of people in the *j*th age group of the local population
n_j^* is the number of people in the *j*th age group of the standard population and
D^* is the number of operations in the standard population.

All calculations use a four-year average to reduce the impact of random fluctuations at a small area level.

Some patients may have been treated in a Primary Care Trust (PCT) region other than the one in which they were resident. To take into account patient mobility, VT utilisation rates are based on postcode level data rather than hospital level data. This means that a patient who is resident in PCT A (as identified by the postcode) but was treated in a hospital located in PCT region B is nonetheless assigned to the utilisation rate of PCT A. This is done because we are interested in the performance of PCTs, as the "stewards" responsible for ensuring an equitable provision of services in relation to medical need (Boyle, 2011), not in the number of operations performed by a particular hospital.

Population capacity to benefit model: The epidemiological model (see Schang et al, 2014a for details) starts from NICE guidance which recommends offering VTs to

children who suffer from bilateral OME at a hearing level of +25 dB and for whom diagnosis is confirmed after a period of three months. The model uses data from population-based longitudinal studies, stratified by age and the cumulative incidence of otitis media. The number of new cases of otitis media in any given year $N(OME)$ is determined by the annual age-specific cumulative incidence (risk) I_j of OME multiplied by the susceptible population in a given age group S_j, summed over J eligible age groups (2, 3, 4 … to 8 years). The subgroup of cases with bilateral OME and a hearing impairment at a threshold level of +25 dB is expressed by

$$N\ (OME) = \sum_{j=2}^{8} (S_j * I_j * P(HL|Bilateral\ OME) * P(Bilateral\ OME|OME))$$

Where

P (Bilateral OME | OME) is the conditional probability of bilateral OME given a diagnosis of OME

P (HL | Bilateral OME) is the conditional probability of a hearing level of +25dBgiven a diagnosis of bilateral OME

The probability of OME persisting at time t from the onset of OME is modelled as an exponential process of the form

$$P\ (OME \mid t) = \frac{1}{2^{\frac{t}{m}}}$$

Where

m is the median time to recovery

t is the total waiting time from the onset of OME

As OME is transitory, the population with capacity to benefit will diminish as time passes since the onset of OME. Population capacity to benefit from ventilation tubes for OME at five months since the onset of OME is estimated as

$$PCB\ (t) = P\ (OME \mid t) * N\ (OME)$$

The "simple" model of need is defined as

$$N\ (OME, t) = \sum_{j=2}^{8} (\boldsymbol{S_j} * \boldsymbol{I_j})$$

Where

N (OME, t) is the number of children defined to be in "need" of VTs for OME at t=5 months after onset

I_j is the annual age-specific cumulative incidence (risk) of OME

S_j is the susceptible population in a given age group

6 COMPLEMENTARY LOGICS OF TARGET SETTING: HIERARCHIST AND EXPERIMENTALIST GOVERNANCE IN THE SCOTTISH NATIONAL HEALTH SERVICE

Authors:

Laura Schang[a] and Alec Morton[b]

Author information:

[a] Department of Management, London School of Economics and Political Science, London, United Kingdom

[b] Department of Management Science, Strathclyde Business School, University of Strathclyde, Glasgow, United Kingdom

Correspondence to:

Laura Schang
Department of Management
London School of Economics and Political Science
Houghton Street | London | United Kingdom
Email: L.K.Schang@lse.ac.uk

Key words: performance management; experimentalist governance; hierarchist governance; healthcare targets.

Abstract

Where policy ends are contested and means for change are ambiguous, imposing central targets on local organisations – what we call hierarchist governance – is problematic. The concept of experimentalist governance (Sabel and Zeitlin, 2012) suggests that target setting should rather be conceptualised as a learning process and as a dialogue between central government and local organisations. However, it is unclear how a constructive dialogue about improvement might be fostered alongside attempts to strengthen accountability for results. Drawing on experiences from the Scottish HEAT target system, we argue that complementary use of hierarchist and experimentalist ideas is possible. We show that the emphasis on experimentalist ideas was stronger where ends and means were contested (the case of shifting the balance of care for older people) than where both ends and means seemed obvious initially (the case of healthcare-associated infections). However, management drifted towards the experimentalist realm when rising rates of community-acquired infections decreased clarity about effective interventions.

6.1 Introduction

Although few would doubt the value of making explicit the priority areas where urgent improvement is needed, translating these priorities into precise, time-bound targets that are imposed by central government on local organisations has attracted much criticism (Carter, 1989, Greenhalgh et al., 2010, Lawton et al., 2000). This form of governance by targets – what we call hierarchist governance – requires "dials" (Carter, 1989): accurate measures of performance which unambiguously represent desired policy ends (Bevan and Hood, 2006) and whose means of attainment are known and available to the organisations under scrutiny (Jacobs et al., 2006a). Many performance indicators in social policy, however, are mere "tin openers": measures which "do not give answers but prompt interrogation and inquiry, and by themselves provide an incomplete and inaccurate picture" (Carter, 1989: 134). This holds in particular for "wicked" problems where goals are contested and means for change are ambiguous (Rittel and Webber, 1973). Some have therefore argued that setting delusively exact targets for wicked problems such as health inequalities obscures complex causal networks and necessary value judgements in determining desired levels of achievement (Blackman et al., 2009).

The concept of experimentalist governance (Sabel and Zeitlin, 2012) suggests that in a context of ambiguity over the "correct" targets and means, target setting should rather be conceptualised as a learning process and as a dialogue between central and local organisations. However, it is still unclear how a constructive dialogue about measurement for learning and improvement can be fostered alongside demands for accountability for results (Freeman, 2002). The conventional performance measurement literature has tended to argue the purposes of "measurement for accountability" and "measurement for improvement" ought to be kept separate since the former is premised on a culture of judgment against fixed objectives, with a consequent need for accurate data, while the latter requires a culture of learning and openness, using data that is "good enough" to diagnose and remedy problems (Solberg et al., 1997, Freeman, 2002, Davies, 2005). However, while these two

purposes of performance measurement have undeniable differences, keeping them entirely distinct would ignore the potential value of targets as a policy tool to track performance improvements on those wicked policy issues where clear agreements on change and measurement of progress might be most needed.

In this paper, we therefore ask how and to what extent a more learning-oriented experimentalist logic of setting performance targets might complement a more hierarchist philosophy focused on accountability. We add to a stream of literature in public administration that considers the potential for complementarity of seemingly dichotomous models of governing public services through deterrence and sanctions or through persuasion and support (McDermott et al., 2015). We examine two research questions:

1. Is it possible to disentangle and examine empirically the co-existence of hierarchist and experimentalist elements in the same performance management regime?

2. Does the relative emphasis on experimentalist as opposed to hierarchist logics differ between policy issues depending on the degree of perceived ambiguity over ends and means?

The next section contrasts, in a stylised way, theoretical assumptions underpinning hierarchist and experimentalist approaches to target setting. The empirical analysis draws on experiences from the Scottish HEAT target system in the National Health Service (NHS) which transformed the earlier model of "trust and altruism" (Bevan et al., 2014) and enables exploration of experimentalist ideas in a hierarchical yet collaboration-oriented context. We then compare two policy issues. Where ends and means were contested (the case of shifting the balance of care for older people; a typical "tin opener"), we find a stronger focus on experimentalist ideas in the form of locally agreed targets and a focus on local innovation. Where both ends and means seemed obvious (the case of healthcare-associated infections; an apparent "dial"), hierarchist elements dominated initially. However, management style drifted towards the experimentalist realm when rising rates of community-acquired infections

decreased clarity about effective interventions. We close with implications for policy and research.

6.2 Hierarchist and experimentalist assumptions about setting performance targets

Drawing on principal-agent theory and institutional economics, the hierarchist logic of setting performance targets involves a sovereign principal (e.g. a central government or regulator) who specifies a set of fixed targets for an agent (e.g. a subordinate agency), rewards achievement and sanctions failure (Hood, 1991, Laffont and Martimort, 2002). In the public sector, hierarchist target setting became a key policy instrument under New Public Management reforms pursued in various countries since the 1980s (Hood, 2012, Hood, 2007). It is vividly illustrated by the model of "targets and terror" (Bevan and Hood, 2006, Propper et al., 2008) adopted by the English Labour Government between 2001 and 2005 where public sector organisations were subject to a set of strict performance targets with severe consequences for failure.

Experimentalist governance (Sabel and Zeitlin, 2012) has evolved as a critique by scholars of the assumptions underlying the hierarchist perspective (Table 6-1) in parallel to calls for more deliberation in public management (e.g. Barzelay, 1992, Hood and Jackson, 1994). Experimentalist governance has been defined as a "recursive process of provisional goal-setting and revision based on learning from the comparison of alternative approaches to advancing them in different contexts" (Sabel and Zeitlin, 2012: 169). Performance management is conceived as a "learning process" whose four elements are linked in an iterative cycle: (i) Centre and local actors agree on broad goals and metrics to ascertain their achievement; (ii) local actors pursue these goals in their own way while the Centre provides support and infrastructure; (iii) local actors report their performance regularly, engage in peer review and share learning about "what works"; (iv) goals, means, and decision-making procedures are periodically revised by a widening circle of actors in response

to the problems and opportunities identified in the review process (Sabel and Zeitlin, 2012: 170).

Experimentalist governance was developed to understand how multilevel governance structures in uncertain and heterogeneous contexts can improve performance and unblock reform stalemates. This includes policy-making in the European Union on areas such as social protection, electricity, telecommunications, occupational health and safety, and drug and food safety where the interdependency of member states, the European Commission and other stakeholders often precludes formal rule-making and has led to alternative structures such as the Open Method of Coordination, Reference Networks and Joint Action Strategies (Sabel and Zeitlin, 2008, Fierlbeck, 2014). Experimentalist governance has also been used to analyse reforms of complex regulatory settings such as child protective services (Noonan et al., 2009) where norms and standards cannot fully be determined ex ante, but require revision in the light of individual cases.

The limitations of hierarchist target governance in a context of ambiguity over ends and means, and the rationale for an experimentalist alternative, can be summarised as follows. First, where ideal ends of policy are contested, scholars claim that argument offers a better basis for decision-making than authority (Pires, 2011). Because local organisations have an insight into frontline problems that national oversight bodies lack, it is argued that "global and local knowledge are mutually corrective, not hierarchically ordered" (Sabel, 2004: 181). Setting goals and metrics to gauge their achievement is therefore seen as a joint process where official authorities "must be prepared to learn from the problem-solving activities of their 'agents' " (Sabel and Zeitlin, 2012:175).

Second, where means for implementation are ambiguous, the hierarchist approach provides no indication what local organisations are to do. As a consequence, local organisations tend to develop various coping strategies (Lawton et al., 2000) but there is not necessarily an attempt to make these strategies explicit or learn from them at a national scale. Experimentalist governance, in contrast, provides a management process which reframes local variation into an experimentalist

"laboratory" and into an opportunity to promote innovation. The underlying assumption is that where the "best" interventions are not known, learning from the range of approaches taken in different localities may help to develop and scale-up good practice (Sabel and Zeitlin, 2012).

Third, the hierarchist view of accountability in relation to results implies the need to comprehensively specify goals. Since most public services have multiple and potentially conflicting goals, incomplete contracts combined with divergent interests and informational asymmetries between regulatory authorities and organisations under scrutiny risk encouraging gaming on the side of the latter (Bevan and Hood, 2006) and misinterpretation of complex local production processes on the side of the former (Smith, 1995). In search of complete contracts, reforms in England and the Netherlands have led to a multiplication of quantitative indicators so as to cover "every" aspect of performance, thereby overwhelming the capacities of both central and local organisations (Power, 1999, Pollitt et al., 2010).[7] Experimentalist governance, in contrast, reframes accountability towards the validity of underlying processes (Table 6-1). Opening the "black box" of service delivery by seeking to understand how measured indicators are implemented is seen to enable a more rounded view of performance and limit the need for a variety of indicators (Noonan et al. 2009; Sabel and Zeitlin, 2012).

The stylised comparisons in Table 6-1 and above suggest that hierarchism and experimentalism are diametrically opposed in terms of their underpinning assumptions about target setting, implementation, and monitoring and accountability. Indeed, prior research (Pires, 2011, Sabel and Zeitlin, 2012) has tended to portray these logics as competing. However, as Fossum (2012) points out, it is not clear yet if experimentalism is mutually exclusive, complementary or transformative in relation to hierarchism.

[7] In England, for instance, this entailed a cascade effect where 12 central health Public Service Agreement targets were translated into 44 targets in the Department of Health's planning framework and finally into 300 targets for local organisations (Collins et al, 2005).

Since central government has a democratic mandate to define priorities and hold subordinate administrations accountable for the use of resources (Mays, 2006), the experimentalist proposal of joint target setting seems at odds with vertical lines of accountability as they exist between central government and administrations with delegated decision-making powers. However, this does not mean that experimentalist governance has nothing to add to hierarchically organised systems.

In this paper, we examine the relationship between hierarchist and experimentalist approaches to managing the performance of a publicly financed service that is subject to parliamentary scrutiny. We investigate two hypotheses. First, we suggest that experimentalist ideas may complement hierarchist target setting so as to address key limitations of the latter. While a full adoption of experimentalist governance may not be feasible or desirable, we hypothesise that it is possible to examine empirically both experimentalist and hierarchist elements in the same performance measurement regime. Second, we examine the choice of logic contingent on the degree of perceived ambiguity in a policy issue. If experimentalist governance is a valid descriptive theory of how systems and organisations manage ambiguity over ends and means, then one would expect a stronger focus on experimentalist ideas in a context of ambiguity while in a context of (relative) certainty one would expect a more hierarchist approach to target setting.

Table 6-1 Hierarchist and experimentalist assumptions about the target setting process

	Hierarchist target setting	Experimentalist governance
Assumptions regarding		
Ambiguity about ends: **Target setting**	Central government has a legitimate mandate to set targets for subordinate administrative bodies.	Knowledge about goals is contested, provisional and distributed between central and local actors. Therefore, target setting should be a joint process between central and local actors.
Ambiguity about means: **Implementation of targets**	• Local agents have the necessary (and have in fact been chosen for their) specialist expertise to implement targets. • The role of the state is to design contracts that incentivise agents to meet the targets and that control the effects of asymmetric information about the effort of agents.	• Means for change are ambiguous. • The role of the state is to provide support and infrastructure that encourages mutual learning and exchange about diagnosing problems, coaching and spread of successful models.
Monitoring and accountability	Accountability against fixed rules: Inspection of results and application of appropriate rewards and punishments.	"Learning-by-monitoring" (inspection of validity of processes): frontline organisations repeatedly explain the choices they make for running a programme, thus enabling oversight bodies to consider how to correct flaws in service delivery at the local level. Dynamic accountability: Results may deviate from initial goals if justified to be a better way to meet the overarching purpose of the system.

Sources: authors' display based on Sabel and Zeitlin (2012); Noonan et al. (2009); Laffont and Martimort (2002).

6.3 Methods

6.3.1 System context and study design

Scotland offers an interesting testbed to examine experimentalist governance ideas within a hierarchical and diverse context. The planning and delivery of health services is delegated to 14 territorial NHS Boards who are responsible for £10.9 billion (of £11.9 billion Government spending on health in 2012/13; Audit Scotland (2013: 5)). But while NHS Boards are major budget-holders and have considerable powers to shape patterns of service delivery, they remain directly accountable to Scottish ministers and subject to central constraints such as the requirement to break even in each financial year (Steel and Cylus, 2012). Boards differ widely in terms of populations covered (between about 21,500 in NHS Orkney and over 1 million in NHS Greater Glasgow and Clyde in 2013; Scottish Public Health Observatory (2014)); rurality (ranging from highly urbanised over mixed urban/ rural economies to rural and remote islands); deprivation (from deprived inner city to wealthy suburban); and geographic size (half of Scotland's landmass being covered by a single Board, NHS Highland).

The model of governance of the Scottish NHS immediately after devolution in 1999 has been described as one of trust and altruism since local organisations were trusted to deliver a high-quality service (Bevan et al., 2014). Ministers have long eschewed targets or rankings that inflict reputational damage, rejecting top-down performance management in favour of a consensual approach. A strong policy discourse on partnership has evolved in relation to a well-organized, powerful medical elite (Greer 2004) and a lower relational distance between central and local organisations than in England (Hood, 2007). In Scotland, senior managers from the different regions "meet regularly and have easy access to ministers and officials in the Scottish Government" (Steel and Cylus 2012: 26).

Targets for territorial Boards were first introduced in 2002, in the form of the Performance Assessment Framework (PAF) (Scottish Executive Health Department

2003). The exact number of targets was not known as the PAF referred to other policy frameworks; estimates range between a hundred and over two hundred targets. An evaluation concluded that there was an "overload from the data collection (…) and the risk that PAF might become an end in itself" since it lacked incentives for Boards to improve and share good practice (Farrar et al., 2004: ii).

This changed in 2006 when, as Steel and Cylus state, "[u]nfavourable cross-border comparisons (…) about performance, particularly on waiting times" (2012: 113) and a change in minister led to the introduction of a "tougher and more sophisticated approach to performance management" (2012: 114); known as the HEAT (Health improvement, Efficiency, Access and Treatment appropriateness) target system. Within a hierarchical yet consensual context, this system has now matured over almost a decade. It offers a suitable context to investigate empirically the potential co-existence of hierarchist and experimentalist logics of setting performance targets (our first research question).

Since our second research question is concerned with the balance between hierarchist and experimentalist logics contingent on the nature of the policy issue, we use a comparative embedded case study design (Yin, 2003). This enables us to compare, within the broader HEAT target system, the development of HEAT targets for two policy issues which represent opposite ends on a spectrum of ambiguity over goals and means (Table 6-2).

- For the case of healthcare-associated infections (HAI), where both ideal performance and the means for change were relatively well-known when targets were introduced, one would expect a stronger hierarchist logic.
- For the case of shifting the balance of care for older people (SBC), where both the ideal ends and the means for change were ambiguous, one would expect a stronger experimentalist logic.

The second reason for choosing these two policy issues is methodological: Targets were introduced in 2006 (SBC) and 2008 (HAI) and have evolved up to the time of writing (April 2015), thus enabling a comparison of their development over time.

Table 6-2 Case studies

Policy issue	End (ideal)	Means	Targets (examples)
Healthcare-associated infections (HAI)	Zero infections	Relatively good evidence of effective interventions (e.g. Haley et al., 1985)	Reduce by 2012/13 NHS Boards' staphylococcus aureus bacteraemia (including MRSA and MSSA) cases to 0.26 or less per 1,000 acute occupied bed days; and the rate of Clostridium difficile infections in patients aged 65 and over to 0.39 cases or less per 1,000 total occupied bed days
Shifting the balance of care for older people (SBC)	Unknown balance between hospital and community care	Service redesign – complex and little evidence (Johnston et al., 2008)	Reduce the rate of emergency inpatient bed days for people aged 75 and over per 1,000 population, by at least 12% between 2009/10 and 2014/15

6.3.2 Data collection and analysis

To achieve a rounded perspective, the study triangulates multiple sources of data including national and local policy documents and interviews (Table 6-3). Interviewees were invited following a purposive strategy (Patton, 2002) to capture national and local experiences and represent diversity in local contexts. We started with an initial group of national and local managers and, using "snowballing" techniques, invited 33 people for interview. A total of 31 interviews were conducted between June 2014 and February 2015 (two people declined due to time constraints). We considered data saturation to be achieved when no new themes emerged after a couple of further interviews (Guest et al., 2006). Participants were informed about the aims of the research project, encouraged to ask questions and assured of the anonymity of their responses. Ethics approval was obtained from the LSE research ethics review committee.

Thematic analysis was used to analyse the documentary material and interview transcripts (Boyatzis, 1998). This systematic approach consisted of identifying

patterns in the data through a process of careful iterative reading and indexing from the data, facilitated by the NVivo software programme. The analysis followed the theoretical constructs from hierarchist and experimentalist governance and thus enabled a form of theory triangulation where the same case is examined through different theoretical lenses to see what each perspective adds or omits (Patton, 1999). To mitigate against misinterpretation, we shared the findings with the interviewees who were given the opportunity to comment on the draft and point out any factual errors.

Table 6-3 Data sources

Data source	Role in analysis
National policy documents	Policy context and developments
112 Annual Local Delivery Plans (from 2006/07 over seven years for all Boards) and other local plans	Historical, public documents agreed with the Government in which Boards set out risks and management strategies for each HEAT target
31 semi-structured interviews lasting 35-90 minutes ***National level:*** 9 with Scottish Government officials and representatives from national organisations: Quality, Efficiency & Support Team (QuEST), HAI Task Force, Health Protection Scotland, Joint Improvement Team (JIT) ***Local level:*** 22 with senior and middle managers from NHS Boards: chief executives, heads of performance management, medical directors, infection control managers, and operational managers - 8 of 14 Boards with a mix of rural/urban, small/larger Boards: Greater Glasgow and Clyde; Borders; Tayside; Dumfries and Galloway; Shetland; Grampian; Forth Valley; Lothian - 2-4 interviews per Board to obtain different perspectives	Perceptions not addressed in public documents

6.4 Findings

The next sections present the overall HEAT target framework and the comparative analysis of policy issues. The theoretical relevance of the findings in relation to hierarchist and experimentalist logics is summarized in Table 6-4.

6.4.1 The HEAT target system

The choice of HEAT targets is informed by consultations with service user groups, professional associations and NHS Boards. Some targets arise from political manifesto commitments. Dissatisfaction with the wide range of indicators in the PAF has led to an explicit articulation of criteria for selecting HEAT targets: (i) strategic fit with Government priorities; (ii) availability of baseline data; and (iii) scope for implementation by NHS Boards. The number of targets has been progressively reduced from 32 targets in 2006/07 to 14 targets in 2013/14. Each target runs over a three-year cycle, after which feedback meetings with NHS managers, health professionals and Government officials create the opportunity to revise or abolish targets. While the Scottish Government decides on targets (reflecting the hierarchist logic), stakeholders do discuss areas where targets add value and raise concerns (reflecting the experimentalist logic).

The HEAT target process is led by a Directorate in the Scottish Government which agrees annual local delivery plans (LDPs) with each Board and monitors progress against these. In LDPs, Boards explain how they plan to attain the HEAT targets and the risks they face. A head of performance noted that *"LDPs are a way to sensecheck with the Scottish Government and raise concerns we see locally".* This reflects the experimentalist logic insofar as LDPs may serve as a mechanism to gain an overview of local problems that enables the Government to reconsider its policy ambitions. Nevertheless, LDPs are drafted in relation to guidance by the Government in line with national policy, including the 2020 Vision for Health and Social Care (The Scottish Government, 2013), and are signed-off by the Scottish Government. Key informants perceived LDPs primarily as the "contract" between Government and NHS Boards rather than as tools for dialogue, reflecting the hierarchist logic.

In contrast to the PAF, the HEAT target system has institutionalised two routes to assure accountability. One is a summative assessment that reflects the hierarchist logic of accountability for results. The Government examines progress against the LDPs at NHS Board Annual Reviews. In addition, biennial Accountability Reviews are conducted by ministers in public where members of the public can ask questions. These reviews result in a formal letter from the Cabinet Secretary about areas of concern identified which Boards are expected to address. National progress against the HEAT targets is reported publicly on the Government's website *Scotland Performs* which shows comparisons with previous years and with other NHS Boards. There are no financial sanctions or forced redundancies when targets are not met.

The other route is a formative assessment that reflects the experimentalist view of accountability as valid processes. This takes the form of Mid-Year reviews to gauge if Boards are "on track". Pressure for corrective action can be applied and escalated through several mechanisms. At monthly meetings of Board Chief Executives with the NHS Chief Executive, performance is routinely discussed in an open forum. Regular informal (bilateral) meetings take also place between performance managers from the Scottish Government Health and Social Care Directorate and Board staff. If there are concerns about a major failure, the Scottish government sends a performance support team to Boards to identify problems and point out interventions for improvement. To give them a greater degree of credibility, these teams typically comprise clinicians, managerial and data management experts seconded to the Scottish Government from other organisations, rather than civil servants. The main purpose of these teams is to enable early intervention and not allow a system to fail. However, respondents highlighted a constant tension in the teams' dual role of providing genuine support and exercising Government control:

"Boards don't have a choice (…) they develop a joint action plan that is heavily scrutinised by the Scottish Government" (Government official).

"The process is not comfortable (…) but from my experience it has been respectful and positive to diagnose problems with data management we had here (…) and showing how other high-performing organisations solve these issues" (senior manager).

Government officials and local managers pointed out that accountability for results was a response to perceived opacity of the previous system. However, because this evolved within a broader context of public sector reform which emphasised the values of partnership and collaboration (see e.g. the report of the Commission on the Future Delivery of Public Services (2011)), there was also an increasing commitment by the Government to listen to and act on local feedback. Below we explore these themes further in the context of HAI and SBC.

6.4.2 Comparative analysis of policy issues

Both the rates of HAI and emergency bed days have reduced considerably over the past years (Figures 6-1 to 6-3). The change points occurred, approximately, when targets where introduced. Although it is not possible to attribute these changes solely to the introduction of targets, and there are some notable variations between Boards, the trend does indicate that actions taken around this time period have impacted on performance.

6.4.2.1 Healthcare-associated infections: "zero is best"?

The issue of healthcare-associated infections (HAI) climbed the policy agenda in 2008 when an outbreak of clostridium difficile (CDI) at the Vale of Leven hospital resulted in a major revision of infection control practices. To coordinate national strategy development and implementation, the Scottish Government's HAI Task Force set out a multifaceted approach to change, funded with £56 million ring-fenced for three years across five broad areas (HAI Taskforce, 2008): standards of practice; culture (resulting e.g. in a national campaign to promote "zero tolerance" for insufficient hand hygiene); education (e.g. on prudent use of antibiotics); surveillance and audit (e.g. since 2009, the Healthcare Environment Inspectorate has carried out regular audits of compliance with national standards); and changes in the physical

environment and processes (e.g. the introduction of MRSA screening on admission). Action on HAI was also embedded into the Scottish Patient Safety Programme, a large Collaborative aimed at improving patient safety. Driven by a political scandal, for HAI there was thus from the outset a strong emphasis on national leadership.

Starting in 2008, the HAI Task Force also recommended the introduction of a national HEAT target on *Staphylococcus aureus bacteraemia* (SABs, including MRSA and MSSA), based on the knowledge at that time that SABs represented a dominant cause of infection. A HEAT target on CDI followed in 2009, once baseline data from mandatory surveillance (since 2008) was available. Proposals for targets are inititated by the National Advisory Group, a subgroup of the HAI Task Force whose membership includes key professional groups (infection control managers, medical directors, and microbiologists employed by Boards). Proposals are passed to the Scottish Government for decision. While the Scottish Government thus retains the responsibility for setting targets (reflecting the hierarchist logic), proposals are initiated by the stakeholders who will implement them (which is closer to the experimentalist logic). The implication of this approach, as an infection control manager observed, was that targets were *"more easily accepted and we are not surprised when the target comes"*.

Target values, however, were derived from technical information rather than consensus; reflecting hierarchist ideas. Initially, targets aimed at a 30 per cent reduction of HAIs over a five-year period. This was based on a seminal study on the prevention of nosocomial infection in the United States (Haley et al., 1985) which had found that a third of hospital infections was avoidable with a defined set of interventions including surveillance, having trained infection control staff, and a system for reporting infection rates to practising surgeons. Although this study came from a different context, it seemed *"the best evidence and reference point at the time of potential for prevention"* (HAI Task Force representative).

Although an ideal rate is known for HAI (zero infections), uncertain effect sizes of interventions make it hard to ascertain what levels of quality are feasible in practice. With comparable data becoming available from standardised reporting across the

United Kingdom and other European health systems, since 2011 levels of achievement are determined by "best-in-class" benchmarking where the Government seeks to infer an attainable level of quality for NHS Boards from the best performers within Scotland and abroad. The theory underpinning this approach has been formalised as yardstick competition (Shleifer, 1985) as a strategy to overcome imperfect and asymmetric information about prices, and socially efficient levels of cost reduction. A key challenge is to identify genuinely comparable firms which "the regulator can expect to be able to reduce costs at the same rate" (Shleifer, 1985: 320). In Scotland, a HAI Task Force representative noted that no adjustments for differences in context were made to keep the system transparent. However, as some Boards argued that the lowest rate would not be achievable for all Boards, given differences in local populations, the HAI Task Force agreed on the roughly 75th percentile of the empirical distribution of performance as the minimum target. Those performing better were however expected to continue to improve to prevent regression towards the minimum target.

Since Boards have integrated responsibility for acute, primary and community care, the national target includes all cases of HAI regardless of where they have been acquired. Boards' actions have traditionally focused on hospitals: LDPs over the past years emphasise the education and training of hospital staff to prevent the transmission of infections. However, while significant reductions in HAI have led to increasingly stringent targets consistent with the "best-in-class" approach, Boards reported struggling with a rising proportion of community-acquired infections (CA-HAIs).[8] In interviews, managers commented that CA-HAIs were considerably harder to prevent than hospital-acquired HAIs as the latter arise within clear physical and managerial boundaries. Many challenged the policy ambition towards increasingly strict meanings of quality: "*We're stuck (...) How low can we go in practice? We can get lower, but we can never get to zero*" (infection control manager). The Scottish

[8] Community-acquired infections are defined as infections that develop within 48 hours after patient admission to hospital. Mentioned by three Boards in 2007/08 (when the SAB target was introduced), by 2013/14, all Boards identified the increasing proportion of community-onset SABs as a key risk to meeting national targets. For instance, NHS Tayside's LDPs record that by 2013/14, about 50% of SABs were present on admission to hospital, an increase from about 22% of SABs in 2006/07.

Government argues however that Boards' legal mandate includes public health and thus has chosen not to exclude CA-HAIs from Boards' targets. There were, however, two sets of alternative responses.

First, the approach to implementation was refined. Scottish Government leads have started to regularly attend local HAI meetings and share "good practice" from other Boards. As an infection control manager commented: *"We had some problems with community-acquired infections and the Government teams said to us, 'you might speak to Board x and y about this' (...) this was not a panacea but it was a start"*. Further, as the epidemiology of CA-HAI is poorly understood and routine national data is lacking (Health Protection Scotland, 2014), this has led to a collaborative research programme through the Scottish Infection Control Network. A rising number of Boards has also begun reaching out beyond hospitals through programmes targeted at high-risk groups such as care home residents and intravenous drug users. Central guidance on infection control now includes sections on CA-HAI (Health Protection Scotland, 2012), and Boards reported referring general practitioners, district nurses and care home staff increasingly to this national resource.

Second, the nature of accountability arrangements for HAI seems to have recognised the uncertainty in attribution and capacity for change of HAI rates. This includes an emphasis on root-cause analysis where staff seek to trace infections back to their initial source. Health Protection Scotland has adopted an approach to inspection that consists of visiting hospital wards locally to understand why they are not "on target", by observing actual practice: *"One Board told us, 'It cannot be the central lines, not the peripheral catheters we checked these' (…) we observed how audits were conducted and it turned out they were not rigorous enough (…) so there was a training to help clinicians understand their importance"* (Health Protection Scotland representative). Especially respondents from smaller Boards also commented that in discussions with the Scottish Government, HAI rates were recognised to be sensitive to unpredictable fluctuations.

In conclusion, the perception of HAI as an issue where national standards where both feasible and desirable has resulted in strongly centrally determined targets. However,

as rates approach zero, finding a normative standard seems harder than ever. The reframing of the boundaries of the problem from hospitals to the wider community has emphasised the value of more learning-oriented approaches to implementation and accountability.

6.4.2.2 Care for older people: a question of "balance"?

Shifting care out of hospitals into the community became a key policy focus in 2005, when the Scottish Government launched the Unscheduled Care Collaborative to tackle the growing rate of emergency admissions especially among older people. The Collaborative was intended as the overarching approach to engage health professionals, service users and carers and provided funding to develop local infrastructures, clinical leadership and information management. The Collaborative framework (Scottish Executive, 2005) set out the principle that: *"change will not be delivered by issuing guidance and directives (…) one size does not fit all. Solutions must meet local need and circumstance and more importantly actively engage staff in the change process"* (p.3). The Scottish Government's (2009a) strategic framework *Shifting the balance of care* re-emphasised shifting the location of care (outside hospitals), its focus (from acute to preventive care) and responsibility for its delivery (involving non-medical professionals, patients and carers following the principle of co-production).

To achieve this ambition, between 2008 and 2011, a Long-Term Conditions Collaborative promoted a variety of tools and techniques focused on improving the self-management and care pathways for people with long-term conditions and sought to support NHS Boards in adapting these tools to their own local contexts (The Scottish Government, 2009b). In 2010, the 10-year programme Reshaping Care for Older People (The Scottish Government, 2010) established a £70 million Change Fund to provide bridging finance for local partnerships involving acute, community, third and independent sectors to create joint commissioning strategies. Thus, from the outset, policy discourse emphasised local diversity and stakeholder engagement.

To serve as strategic measure in support of these policies, in 2006, a HEAT target on reducing multiple re-admissions to hospital for people aged 65 and over was introduced. After the usual three-year cycle, a national stakeholder event including Government officials, clinicians, service users and Boards concluded that this target ignored the fact that some admissions might be unavoidable. *"Clinicians were concerned this would create a perverse incentive to prevent even necessary admissions (...) based on that feedback, the target was reformulated"* (clinical lead). In 2009, an alternative target to reduce emergency bed days for people aged 65 and over (later narrowed down to the population aged 75 and over) was introduced which seemed to signal more clearly the underlying policy ambition: *"The target is about minimising the time spent in hospital for older people (...) patient experience of care and health status often suffer as a result of long hospital stays"* (medical director). This target was reaffirmed at the subsequent stakeholder review and has continued up to the time of writing.

Rather than imposing a uniform target, the Government agreed levels of achievement individually with each Board. This model was adopted because local variations in the availability of community care, socio-demographic composition, and previous reductions in emergency bed days were perceived as making a single "right rate" untenable. A lack of comparable data between Boards also precluded the use of benchmarking as for HAI.[9]

Experimentalism would suggest that setting targets through dialogue requires a mutual interest in obtaining challenging yet feasible targets. Since HEAT targets are publicly reported and frequently cited to underpin political achievements, the Scottish Government has an interest in realistic targets that can be met. However, emergency care puts substantial strain on the NHS budget and rising demand due to demographic changes challenges the Government's pledge to protect universal coverage (Barbour et al., 2014). This has led to a perception that *"the current situation*

[9] For historic reasons, NHS Boards classify inpatient beds, especially for long stay treatment, in different ways. Boards are at the time of writing in discussion with the Information Services Division (ISD) to understand these differences. Consequently, however, one can only compare trajectories i.e. relative changes but not absolute levels of achievement.

is unsustainable and there is a real need to reorient health services (Government official), suggesting a commitment to move beyond merely symbolic targets. The incentive for Boards to identify a credible trajectory was, according to our respondents, both improved patient experience and financial sustainability, which is heavily scrutinised by ministers, Audit Scotland and the Scottish Parliament. The Joint Improvement Team (JIT; a partnership involving the Scottish Government, NHSScotland, the Convention of Scottish Local Authorities (COSLA), and the Third, Independent and Housing sectors) played a key role in building this argument: JIT provided Boards with estimates of future needs for hospital beds given demographic projections to show how reducing bed days now would mitigate the creation of expensive new hospital beds in the future.

The national target of achieving a 12 per cent reduction in emergency bed days between 2009/10 and 2014/15 was derived from the aggregation of Board-specific targets. These ranged from 0 per cent (for Boards with a relatively low rate of bed days who felt further reductions were not feasible) to about 20 per cent over a five-year period (in Grampian, where larger-scale service redesign was under development). Notably, both national and local respondents perceived the process as dialogical rather than adversarial:

"We cannot just set the trajectory as we like (...) we look at our historical data and suggest what we can do (...) then the Government says 'we think you can do more here' or 'you are too ambitious' and then we go back to the data (...) it is a dialogue really" (planning manager).

"Some Boards seemed very ambitious (...) but some also had an ambitious improvement programme so the trajectory was backed up (...) it was a lot about speaking to Boards." (Government official).

Two fundamental problems in SBC – and initial differences to HAI – were the limited evidence of (cost-)effective interventions and Boards' partial ownership of targets. In particular the capacity and quality of social care, funded by local authorities, strongly influences Boards' ability to reduce emergency admissions and achieve timely

discharges into community care. In their LDPs, all Boards emphasised the National Change Fund as a catalyst to test new models of care, typically in collaboration with local authorities. Boards state in their LDPs that the complexity of drivers of emergency bed days motivates diversified portfolios of interventions, so as to reduce the risks from "putting all eggs into one basket" (see e.g. NHS Tayside (2011: 27)). Packages of interventions differ widely also between Boards. Examples include "hospital at home" and rapid response services; creating a single point of contact for patients; use of a national risk prediction tool (SPARRA) to predict emergency re-admissions; anticipatory care plans for vulnerable patients; telehealth and telecare; intermediate models of care to provide specialist assessment; and improved discharge planning.

Several national organisations – including the Quality, Efficiency & Support Team (QuEST) within the Scottish Government, Healthcare Improvement Scotland and JIT – have started to foster the scaling-up of good practice through national events, benchmarking initiatives, longer-term programmes geared towards joint strategic commissioning and integrated resourcing, and sharing of case studies (see e.g. QuEST (Quality and Efficiency Support Team), 2014, JIT (Joint Improvement Team), 2014). Although attributing changes to specific interventions remains hard (Steel, 2013), many case studies suggest benefits are being realised. For instance, an immediate discharge service (IDS) in NHS Tayside that streamlined the referral process to reablement services and fostered daily telephone conversations between occupational therapists and the IDS coordinator saved over 1,600 bed days in 2011/12 compared to 2010/11, at an estimated cost saving of £100,562 net of the cost of the IDS team (JIT (Joint Improvement Team), 2015). Local managers we interviewed noted that these case studies were increasingly valued as a central repository of options for local action.

An interesting case study is the trajectory of NHS Borders (Figure 6-3). The data suggests that this Board was transformed from a negative outlier in terms of rates of emergency bed days in 2007/08 to a positive outlier by 2012/13. The Local Delivery Plans over this period suggest that there was an ambitious healthcare improvement programme under development. This programme included the following elements: (i)

introduction of an anticipatory care planning service and self-management programmes to support people to stay at home where possible; (ii) telehealthcare home monitoring; (iii) a new contract for GPs focused on patient pathways and intended to incentivise reductions in length of stay in community hospitals; (iv) LEAN service redesign projects to standardise treatment pathways for common chronic conditions (e.g. COPD, heart failure); (v) a discharge transfer policy signed off by NHS Borders and the local Community Health partnership to minimise delayed discharges; and (vi) the introduction of intermediate care options and improved training of the health and social care workforce. These interventions indicate a range of actions that have been taken in NHS Borders. Nevertheless, more detailed evaluation of these interventions will be needed to understand what exactly made the difference in reducing emergency bed days, compared with other Health Boards.

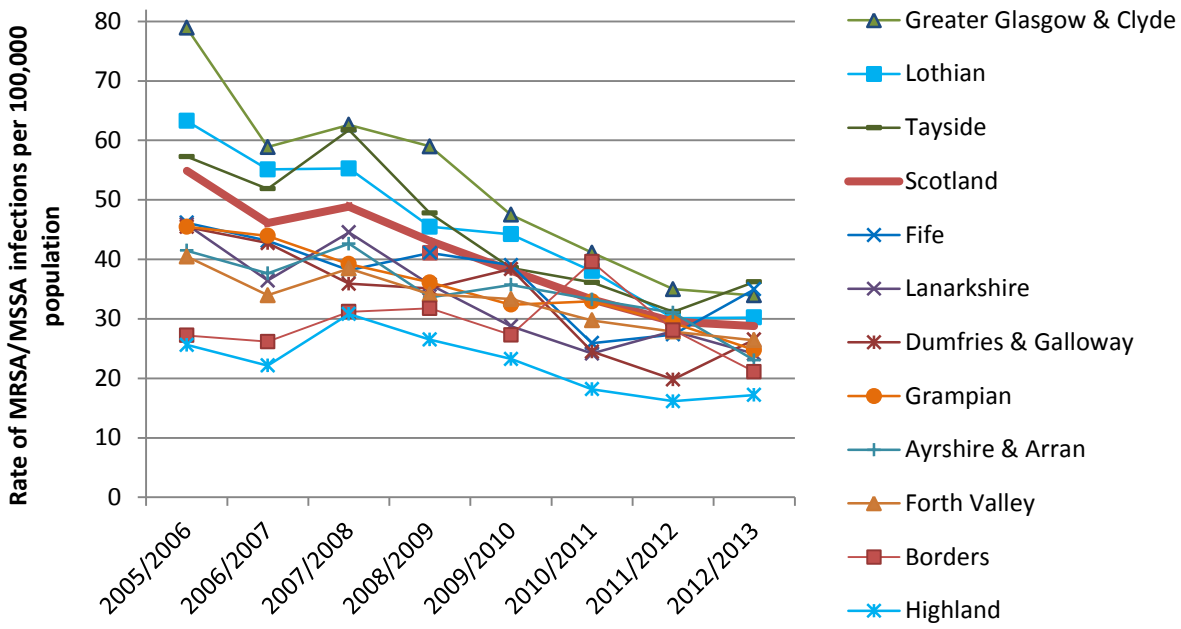
Encouraging local freedom to innovate, with requisite financial and managerial support, and then generalising successes across the system, lies at the heart of experimentalist governance. However, as highlighted by Audit Scotland (2014), various challenges remain: projects tend to be small-scale, often without in-built evaluation or plans how to sustain successful projects after initial funding has ended; and so far there has been less systematic coordination across NHS Boards to understand the reasons for variation in activity and expenditure and to monitor the impacts of interventions for older people across Scotland. In interviews, managers also commented on the difficulties in adapting interventions from elsewhere to their own contexts: *"There is still much thinking that 'we are different in terms of capacity, workforce, money' and clinicians don't always accept even projects that have shown success in other Boards*" (planning manager).

In conclusion, because shifting the balance of care for older people was from the outset recognised as an issue of local diversity and uncertainty about effective interventions, the performance management system focused on the development of Board-specific targets and fostering change through local partnerships. While the generalisation of local "lessons" remains challenging, this approach seems to have enabled the Scottish NHS to make progress on a complex policy issue.

Table 6-4 Hierarchist and experimentalist governance elements in the Scottish HEAT target system

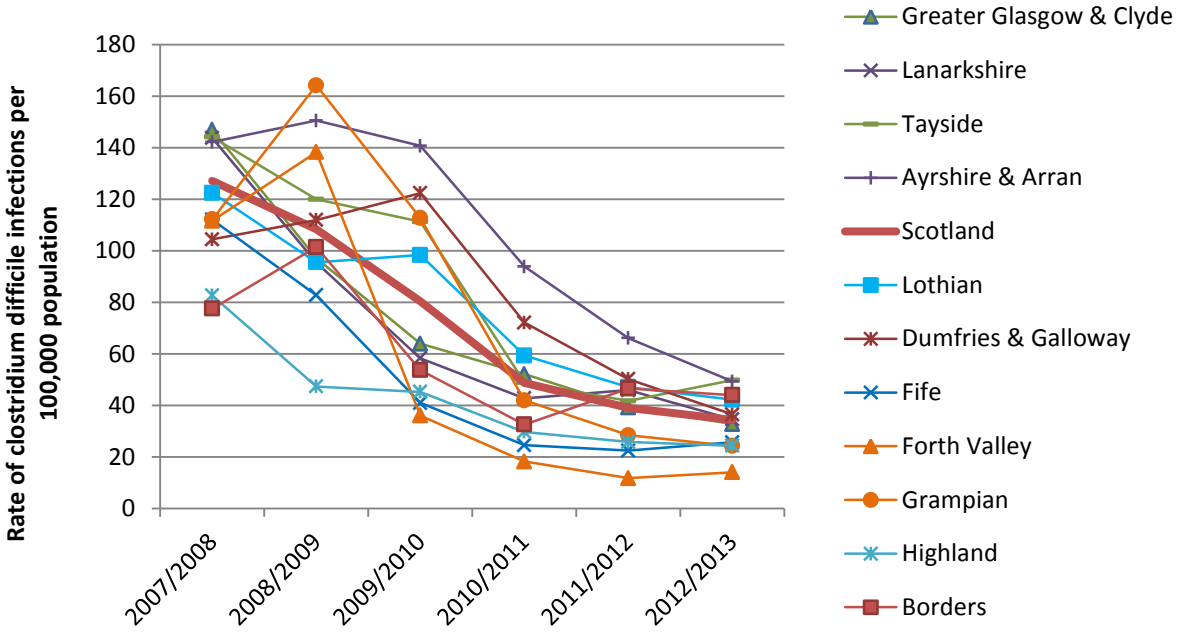
	System level: HEAT framework		HAI	SBC
	Reflects hierarchist logic	**Reflects experimentalist logic**		
Ambiguity about ends: Target setting	The Scottish Government decides on target indicators based on Government priorities.	Consultation with Boards and service user groups for target setting and revision after a three-year cycle	Target values were set centrally based on external (until 2010) and comparative (from 2011) information reflecting a more hierarchist orientation.	Target values were developed through dialogue between Boards and the Scottish Government reflecting a more experimentalist orientation.
Ambiguity about means: Implementation of targets	LDPs are negotiated as "contract" between Government and Boards and signed-off by the Government.	LDPs serve to identify risks and management strategies locally.	National Taskforce Delivery Plans and guidance on infection control define standards for local action on HAI reflecting a more hierarchist orientation. ***With the rise of CA-HAIs***: Increasing emphasis on networks to share good practice, collaborative research, and widening involvement of community health professionals beyond the traditional focus on hospital staff reflecting a more experimentalist orientation.	Focus on local innovation funded through the Change Fund, development of an Improvement Collaborative to provide an infrastructure for improvement, use of case studies to share learning reflecting a more experimentalist orientation.
Monitoring and accountability	Accountability against target achievement is done at Annual Reviews and through public reporting over a Government website.	Diagnostic monitoring throughout the year serves to identify unusual trends and to remedy the underlying reasons.	***With the rise of CA-HAIs***: Increasing emphasis on root-cause analysis and inspection of local clinical processes reflecting a more experimentalist orientation.	

Figure 6-1 Rates of staphyloccocus aureus bacteraemia per 100,000 population (SAB, including MRSA and MSSA)



Note: The target was introduced in 2008. Shetland, Orkney and Western Isles are excluded due to small numbers.

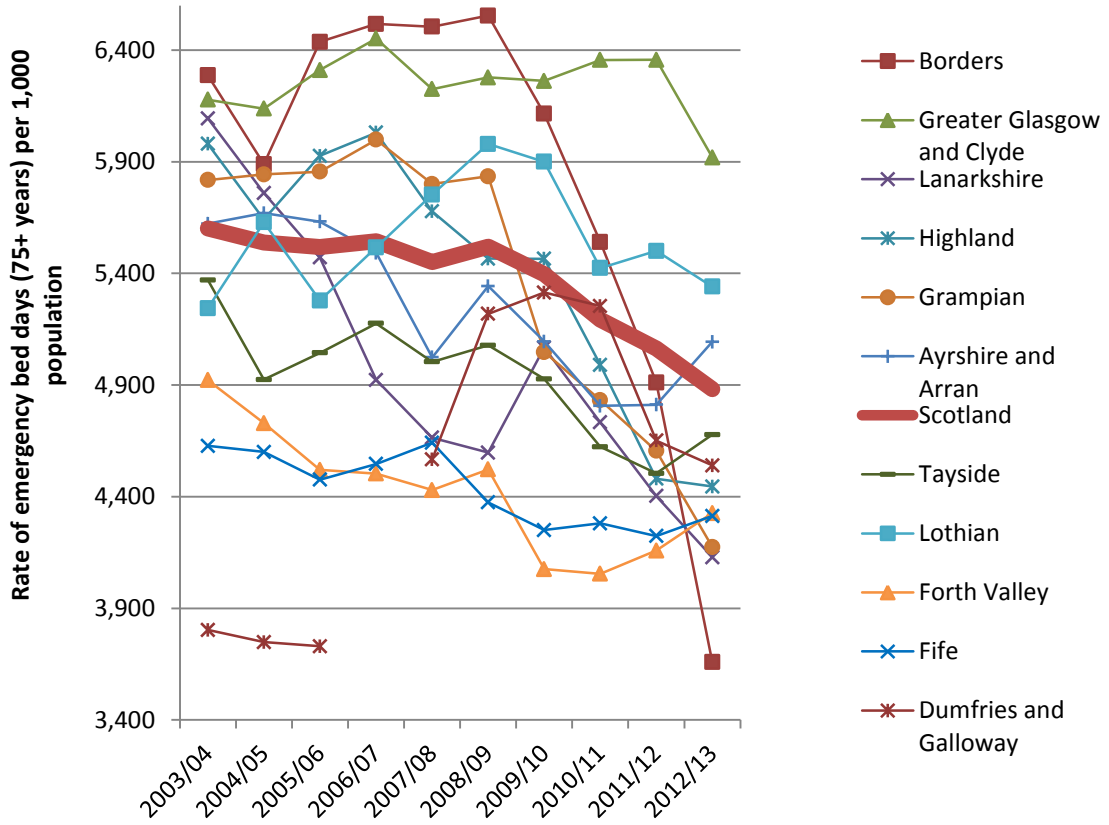
Figure 6-2 Rates of clostridium difficile infections for people aged 65+ per 100,000 population



Source: own display based on Health Protection Scotland (2014).

Note: The target was introduced in 2009. Shetland, Orkney and Western Isles are excluded due to small numbers.

Figure 6-3 Rates of emergency bed days for patients aged 75+ per 1,000 population



Source: own display based on ISD Scotland (2014).

Note: The target was introduced in 2009. Shetland, Orkney and Western Isles are excluded due to small numbers.The stark increase in bed days in Dumfries and Galloway after 2007 is due to the re-classification of geriatric community beds for general use, so that admissions to these beds are now counted which would have previously been excluded. The data for this Board are therefore presented as broken series.

6.5 Discussion

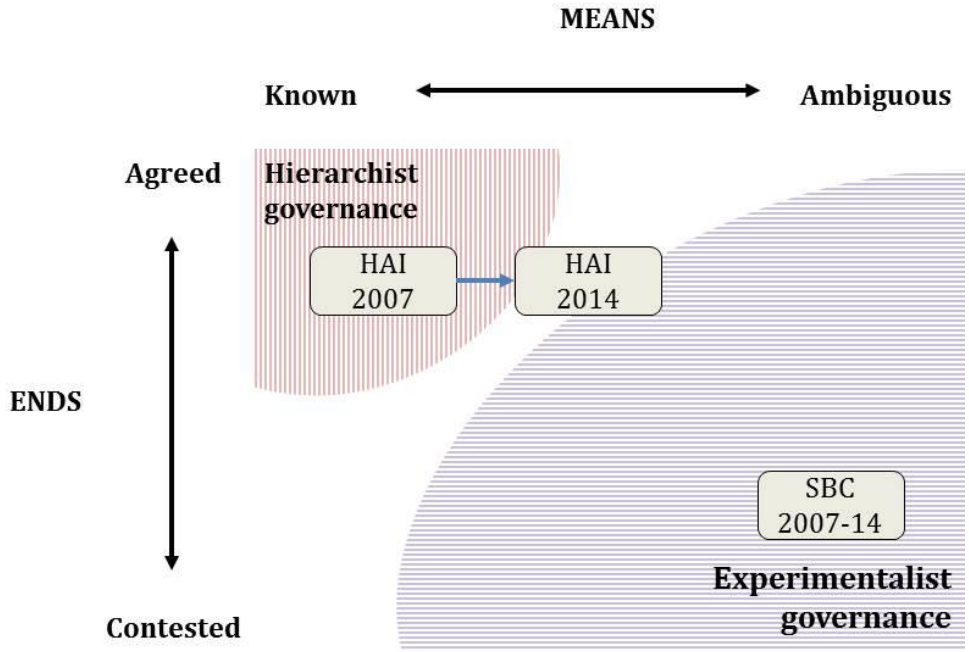
Prompted by the limits of a hierarchist approach to target setting for policy issues whose "ideal" ends and means for change are ambiguous, this paper has examined how more learning-oriented strategies as proposed by experimentalist governance scholars might complement the target setting process. In relation to the two questions stated in the introduction, the findings suggest the following.

First, hierarchist and experimentalist elements can be shown to exist in the same performance management regime (Table 6-4). The experimentalist elements add distinct aspects in relation to target setting, implementation, and monitoring and accountability that are missing from a purely hierarchist approach (Table 6-1). In Scotland, this has led to a performance management system where: central and local actors contribute to setting targets; central bodies support local attempts to implement change; and local actors are held accountable both for processes and for results. This suggests a complementary role of experimentalism (Fossum, 2012). Nevertheless, there is also some evidence of tensions or "competition" between the two logics. This is illustrated by the ambivalent perception of Government performance support teams with their dual mandate for central control and local support.

Second, while both logics influenced the management of each policy issue, their relative emphasis differed between policy issues and also over time within a policy issue. Where targets were informed by the vision of an optimal balance between community and hospital care and means for change were ambiguous (care for older people; Table 6-2), target setting reflected experimentalist ideas more strongly. Yet, central-local discussions were anchored by estimates of demographic projections provided by JIT and thus not data-free. Where ends and means were known initially (HAI; Table 6-2), scope for avoidability was determined centrally by the Government from technical information. However, the National Advisory Group – with representatives from Boards – initiated proposals for targets. When the rise of

community-acquired infections decreased the initial clarity about the causal mechanism and effective interventions, the ideal level of HAI (zero) and the model for target setting remained constant. However, a more learning-oriented approach to implementation and accountability ensued (Table 6-4). This can be interpreted as a partial drift in management style to the experimentalist realm (Figure 6-4).

Figure 6-4 Ambiguity over goals and means in relation to governance style



The main implication of these findings is that the choice of logic (or the distinct combination of logics) can be made on a target-by-target basis. Thinking of hierarchism and experimentalism as a property of the performance management system as a whole may be analytically too coarse (from a research perspective) and neglect opportunities that arise from drawing on the best of both logics (from a policy perspective). As this study shows, it is important to look deeper at the level of individual targets and at changes within a target over time. Returning to the idea of "dials" and "tin openers" (Carter, 1989), this means one can and, arguably, should tailor models of governance to the nature of the policy issue.

Our findings highlight the need to reconsider the strict separation between purposes of measurement for improvement and measurement for accountability advocated in the literature on performance measurement in public services (Solberg et al., 1997, Freeman, 2002, Davies, 2005). Empirically, it seems possible to combine both purposes within a system. Prescriptively, however, should regulators seek to integrate experimentalist and hierarchist logics? We have explored the rationale and context for adoption of different logics of target setting for two policy issues but not their relative effectiveness in facilitating or hindering progress towards achievement of targets.

Our study suggests, however, that an integrative model of governance may help to recognise the inherent differences between policy issues and also their potential dynamics over time. Combining different models of governance may thus serve as a strategy to address the unintended consequences of adopting a single model only (Goddard et al., 2000). Moreover, within a publicly funded healthcare system where government has a legitimate mandate to set priorities and targets (Mays, 2006), ignoring demands for accountability in efforts to improve public service performance seems unrealistic. Sabel and Zeitlin (2012) suggest that working in the ”shadow of hierarchy” within an otherwise experimentalist system retains the possibility of hierarchical intervention if actors fail to collaborate effectively. Nevertheless, prior research also highlights the danger for more learning-oriented modes of governance to be ”crowded out” if regulators deploy sanctions too frequently (Fischer and Ferlie, 2013, McDermott et al., 2015).

How, then, can different logics be integrated so as to limit any consequential damage? While it was outside the scope of our study to examine the preconditions for successful integration of different models of governance, the socio-historical and institutional context with regard to existing legislative frameworks and prevailing norms (Scott et al., 2000) is likely to be very important. Clearly, the choice and effective combination of different models of governance will not solely depend on the nature of the policy issue in terms of perceived degree of ambiguity over ideal ends and means for implementation. The Scottish approach to quality improvement evolved within a unique socio-historical context which included an early emphasis on

professionally led development of standards and longstanding collaborative relationships between central and local organisations (Legido-Quigley et al., 2012, McDermott et al., 2015), making it receptive to experimentalist ideas. Our finding that it is possible to combine experimentalism and hierarchism may thus not be generalisable to a system which is habituated to hierarchist ways of working.

It seems plausible to suggest that experimentalist governance requires a system (i) where actors share at least moderate levels of trust if there is to be an open dialogue; and (ii) where actors are able to build an infrastructure that fosters deliberation and mutual exchange. As Reay and Hinings (2009) argue, the development of collaborative relationships can serve as an effective mechanism to manage the rivalry of competing logics. However, since in this study we have focused on one system only (the Scottish HEAT target system), it makes sense to examine these issues further through comparative research; examining for instance Scotland and England where target setting has tended to follow a more hierarchist logic focused on accountability and deterrence (Bevan and Hood, 2006).

In terms of providing guidance for the design of national performance management systems, it is worthwhile to note that experimentalist and hierarchist logics each embody potential trade-offs. Locally agreed targets might engender stakeholder support and thus mitigate the negative impacts on public service motivation documented for hierarchically driven targets (Drucker, 1954, Le Grand, 2010). However, they also curiously bypass the objective of a national performance management system to enable comparative analyses of how public resources are spent. Forecasting based on historical patterns of activity may be a strategy to get started as long as no comparable data exists, yet it risks reinforcing the "fallacy of comparative difference" (Marmor, 2012: 20) that no cross-regional comparison at all is possible. Benchmarking, in turn, considers the practical feasibility given what top-performers have achieved (Bogan and English, 1994), but ignores that even top-performers may have substantial scope for improvement. Transferring estimates of the potential for prevention from scientific research may enable a rigourous external standard, but reliable estimates of effect sizes of interventions are often unavailable (Ernst et al., 2008) or not applicable to other contexts (Cartwright and Hardie, 2012).

The relative desirability of hierarchist and experimentalist logics is likely to depend on how these trade-offs are managed.

Strategies for successful implementation of experimentalist ideas also merit further research. In light of wide geographic variations in the utilisation and outcomes of public services such as healthcare (OECD (Organisation for Economic Co-operation and Development), 2014), experimentalist governance encourages a guardedly optimistic view: it sees local variation as an opportunity to learn from others. However, experimentalist governance as described by Sabel and Zeitlin (2012) does not indicate *how* actors can best learn from each other (Pires, 2011). Nor does it provide criteria to ascertain that what takes place is indeed learning, a reflective practice (Schön, 1987), or emulation only (Fossum, 2012). These challenges were evident particularly in the case of care for older people. The two core problems that experimentalist governance purports to address – ambiguity over ends and over means – will require very different strategies for learning. The former will require forms of deliberation about values (Mooney and Blackwell, 2004, Brugnach and Ingram, 2012). The latter will require deliberation about what constitutes relevant evidence in a specific context (Cartwright and Hardie, 2012).

The experiences in the Scottish HEAT target system lead us to the inspiring and optimistic conclusion that setting targets for public service performance does not have to end up in a ”targets and terror” model of governance. Targets are a valuable policy tool to focus attention on priorities and to clarify policy ambitions with reference to improving the quality and outcomes of public services. However, when problems are wicked, the hierarchist logic of target setting provides no mechanism to engage stakeholders affected by targets and to foster improvement. Local managers of the system can benefit and learn from variations between local contexts in a much more productive (and democratic) way when they are given the responsibility and capacity to search for feasible solutions for their own local circumstances. Future research should consider the impact of different models of governance on the outcomes of public services and the enabling social and political conditions for effective measurement for learning, collaboration and improvement.

6.6 Acknowledgements

The study was supported under the New Professors' Fund at the University of Strathclyde. We are grateful to all key informants from Scotland who generously shared their experiences. We also thank the participants of the European Health Policy Group Meeting, London 2014, for their comments and in particular Jan-Kees Helderman for his constructive discussion that helped to develop the paper. We thank Gwyn Bevan, Jan-Kees Helderman, David Steel and interviewees from the research for comments on a previous version of this paper. The authors are responsible for the conclusions reached and for any mistakes.

6.7 Commentary in relation to the organising matrix of strategies to address ambiguity about the standard for evaluation

This Chapter has explored how policy-makers and managers might establish meaningful standards for evaluation with a socio-political approach. To this end, the Chapter has examined how an experimentalist governance logic focused on learning and dialogue between central government and local organisations (Sabel and Zeitlin, 2012) might complement a more hierarchist philosophy focused on accountability when setting performance targets. The Chapter has contributed an empirically-based characterisation of the co-existence and potential complementarity of these logics in the Scottish HEAT target system.

7 CONCLUSIONS

Geographic variations in rates of hospital admissions, surgical procedures and other types of health service utilisation have been widely documented for a range of conditions and in various countries (Corallo et al., 2014, OECD (Organisation for Economic Co-operation and Development), 2014). However, due to the absence of a clear standard for evaluation, the practical and policy significance of this information is often unclear. This ambiguity about the standard for evaluation is problematic because it risks foregoing potential insights about health system performance, defined as the degree to which objectives such as the provision of safe, effective and cost-effective services in relation to medical need have been met (Evans, 1990, Hurst, 2002); because it risks encouraging misinterpretation and causing harm (Tanenbaum, 2012); and because reporting data without clear managerial implications seems futile and a waste of resources (Spiegelhalter, 1999, Goddard et al., 2000).

The aim of this thesis was therefore to investigate how regulators and managers in charge of planning, auditing and improving health services might address ambiguity about the standard for evaluation. Essentially the thesis argues that ambiguity about the standard is a multi-faceted phenomenon that can be tackled through four categories of management strategies: managing ambiguity in the absence of a standard using a socio-political approach; managing ambiguity in the absence of a standard using a technical-evidential approach; determining a meaningful standard using a socio-political approach; and determining a meaningful standard using a technical-evidential approach.

On the basis of five empirical studies, presented in Chapters 2 to 6, the thesis has investigated how strategies within each of these categories might look like. This Chapter synthesises the main findings and implications for policy, discusses the limitations of the thesis and suggests directions for future research, and offers concluding remarks.

7.1 Main findings and implications for policy

This section synthesises the empirical findings in relation to the overarching conceptual framework. The resulting implications for policy are summarised in Table 7-1. The four types of management strategies are not mutually exclusive. A comprehensive approach to address ambiguity about the standard for evaluation will likely involve both socio-political and technical-evidential elements, and seek to determine standards where this is possible and manage ambiguity where standards cannot be established. Each of the implications suggested in Table 7-1 is briefly discussed below.

Table 7-1 Policy implications

Strategy	Policy implications from findings of this thesis
1. Manage ambiguity in the absence of a standard using a socio-political approach.	• Overcome barriers to using information on variations: awareness, acceptance, perceived applicability of the information and capacity to use the information. • Strengthen levers for using information on variations: agree responsibilities for action, involve stakeholders and develop tools to help clarify which variations are unwarranted.
2. Manage ambiguity in the absence of a standard using a technical-evidential approach.	• Avoid assigning a single ranking to each unit of assessment in the context of performance comparisons based on composite indicators. • Use ranking intervals and dominance relations to show the uncertainty in rankings and the impact of different assumptions about weight sets and population denominators on the results.
3. Determine meaningful standards using a technical-evidential approach.	• Estimate capacity to benefit in populations to help identify underuse and overuse of defined interventions in tandem with clinical audits and standardisation of utilisation rates for variables associated with need for healthcare.
4. Determine meaningful standards using a socio-political approach.	• Choose the logic for setting healthcare targets (hierarchist or experimentalist) on a target-by-target basis. • Link the evaluation of variations to a management process that focuses on the identification of causes of variations and on the sharing of good practice between central government and local organisations.

If one seeks to manage ambiguity in the absence of a standard using a socio-political approach, then one must overcome a series of practical barriers to information use including awareness, acceptance, perceived applicability and capacity of potential users. Agreeing responsibilities for action and involving clinicians in the process appear to enable the use of information on variations for strategic problem framing and communication. These findings are consistent with those of large-scale reviews of programmes in the United Kingdom focused on quality improvement (Dixon-Woods et al., 2012) and knowledge mobilisation in NHS organisations (Crilly et al., 2013). These reviews demonstrate the importance of involving stakeholders who are affected by performance indicators and whose behaviour enables or constrains progress.

The apparent tendency among healthcare managers to interpret variations with reference to the national average or the top and bottom outliers implies the need for appropriate tools to examine which variations are substantively meaningful in terms of quality of care, expenditure or both. These tools should enable building a narrative why unwarranted variation matters in a particular local context (for examples from English Primary Care Trusts, see Schang and Morton, 2012), explaining the drivers of variations in patients' care pathways across all settings of care (Porter, 2010, Porter and Teisberg, 2006) and engaging stakeholders in a process to prioritise remedial interventions in terms of their impact on population health relative to the required expenditure (Airoldi et al., 2014).

To manage ambiguity in the absence of a standard using a technical-evidential approach, the use of ranking intervals and dominance relations obtained from ratio-based efficiency analysis (REA) can help to avoid the forced assignment of a single, potentially controversial ranking based on a composite measure of performance to each organisation under scrutiny. The REA technique is able to incorporate the full set of feasible weights and different choices of population denominator. The size of the ranking intervals shows the specific positions an organisation can attain in the overall ranking and the sensitivity to different methodological assumptions. The

estimated ranking intervals for a given organisation are robust insofar as they are bounded by the maximum and minimum possible rankings (say, not better than 3rd but not worse than 6th).

There are good reasons, reviewed in Chapter 3, to abstain from developing composite measures of performance altogether and instead to report comparative performance separately for different indicators (Jacobs et al., 2005, Hauck and Street, 2006). In light of the findings of Chapter 2, the creation of composite indicators is likely to further complicate local interpretation of this information. A key strategy pursued by healthcare managers was to disaggregate data from a geographic level to a lower (e.g. provider) level of analysis. This is important as even within hospitals (between wards or at a specialty level), the quality of care may differ considerably (Zhang et al., 2013). While smaller numbers make the disaggregate data on which this approach relies more vulnerable to random fluctuations (Diehr and Grembowski, 1990, Diehr et al., 1990), it can nevertheless help reveal distinct patterns of clinical practice variation which are obscured at a higher level of analysis (NHS Right Care, 2011). If however one moves towards even further aggregation in the form of composite indicators, in order to compare performance on multiple objectives in a unified way (OECD, 2008), then methods of analysis should recognise uncertainty in consequent rankings of organisational performance.

To determine a standard using a technical-evidential approach, estimating capacity to benefit in populations provides a theoretically sound and feasible benchmark to assess the utilisation of services against population needs. Such estimates contribute to the debate on suitable tools to evaluate the appropriateness of variations in medical practice. In particular, they complement clinical audits (which can identify overuse at a patient level, but fail to detect underuse among members of the general population who do not access the health service in the first place) and standardisation of rates for variables associated with need (which helps to ensure fair comparisons between regions which differ in their composition of uncontrollable determinants of healthcare need (Nicholl et al., 2013), but does not provide a normative benchmark of a level of care that is expected to be beneficial for a specific population). To derive estimates of capacity to benefit in populations, it is important

to strengthen the development of evidence-based criteria of capacity to benefit and to target the collection of required epidemiological information.

To determine a standard using a socio-political approach, it is possible to tailor the model of governance to the degree of ambiguity over ideal ends and means inherent in a particular policy issue. As experience from the Scottish HEAT target system suggests, an experimentalist logic of governing healthcare performance through learning and dialogue may usefully complement a hierarchist logic focused on accountability for results when both ideal ends and means for change are ambiguous. This finding adds to a growing line of research that explores the potential for synergistic benefits between different models of governance (McDermott et al., 2015). Taken together, these developments challenge the conventional distinction between "measurement for improvement" and "measurement for accountability" as mutually exclusive purposes of performance measurement (Freeman, 2002). In order to move beyond the mere measurement of variations to their management, the setting of standards and targets should be linked to a process of identification of causes of variations and sharing of good practice between central government and local organisations.

This thesis has contributed an analysis of techical-evidential and socio-political approaches to managing ambiguity in the absence of a standard and to setting meaningful standards for the evaluation of unwarranted variations in healthcare. The subsequent section critically discusses the overall limitations of the thesis.

7.2 Limitations and directions for future research

The specific limitations of each study were summarised in the respective Chapters 2 to 6. This section discusses epistemological, contextual and methodological limitations of the thesis as a whole and suggests directions for further research on this basis.

7.2.1 Epistemological approach

In order to advance knowledge about how to address ambiguity about the standard for evaluation, this thesis was grounded in, and built on, different scientific traditions in the social sciences. Specifically, the nature of options within the four categories of management strategies was explored by drawing on concepts and methods from public health and epidemiology, health economics, operations research and public administration.

Adopting such an interdisciplinary perspective has advantages and disadvantages. Within the scope of this thesis, it was valuable because it enabled the development of a richer and more nuanced account of strategies to address ambiguity about the standard for evaluation. Insofar as all scientific theories and models are simplified representations of reality (Pidd, 1996), each of the concepts and methods applied helped to illuminate the research problem from a different angle. Following Smith et al. (2009), the thesis takes the view that designers of performance measurement systems in healthcare can learn from the distinct insights offered by different fields of the social sciences, justifying an interdisciplinary approach.

However, this richness of perspective comes at a cost. In this thesis, it limited the ability to analyse in more detail the different facets of the research problem and to explore more deeply what different areas of scholarship might offer to this end. Following from Chapter 6, for instance, it would be desirable to explore in more depth the contextual preconditions and the regulatory implications of adopting an experimentalist governance perspective. Following from Chapter 4, future research should seek to examine reasons for the apparent discrepancy between the use of ventilation tubes and the estimated population capacity to benefit. As identified in Chapter 5, the magnitude of these need-utilisation gaps seems to differ across regions and it would be important to explain why this is the case. To do this, one could draw on quantitative and qualitative methods to understand the actual pathways patients take and to assess the impact of and interaction between local supply structures, patterns of medical practice and commissioning policies. Answering these questions went beyond the scope of this thesis and provides fascinating and worthwhile

directions for future inquiry.

7.2.2 System context: Aligning the level of analysis with the locus of decision making

Two studies included in this thesis were set in the English NHS (Chapters 2 and 4) and two focused on the Scottish NHS (Chapters 3 and 6). The English NHS and the Scottish NHS provide a special and, in theory, ideal context for analyses of geographic variations in healthcare. Both NHS systems are characterised by a territorial administrative structure where decision-making is delegated to local organisations in charge of planning and purchasing health services. The level of analysis (Primary Care Trusts and, as from April 2013, Clinical Commissioning Groups in England and Health Boards in Scotland) is therefore consistent with the locus of responsibility for action. In other words, analyses of geographic variations are meaningful in terms of ensuring accountability for the use of resources and for instigating action for improvement.

The concepts and models adopted in this thesis are in principle generalisable to all types of health systems, regardless of their governance mechanisms for the planning, financing and provision of services. However, in systems where no authorities with a geographic basis for healthcare planning exist, it will be more difficult to achieve a close linkage between the measurement and the management of variations in healthcare.

In a recent report to the US Institute of Medicine, Newhouse et al. (2013) question the usefulness of reporting information on variations at a geographic level since, in the United States, this does not reflect the level where decisions about service planning and provision are taken. In many European countries, the mandate for decision-making tends to be fragmented between stakeholders (e.g. national and local levels of government, healthcare payers, hospitals and physicians), policy areas (primary, secondary and tertiary care, public health and pharmaceuticals) and policy tasks (implementation, provision, finance, regulation, and framework legislation) (Adolph et al., 2012). As a result, geographically based analyses may complicate problems of attribution insofar as it is unclear whose performance is being evaluated. Moreover,

payment and other incentives that are based on average outcomes in some geographic area rather than targeted at specific decision-makers (healthcare payers, hospitals, physicians, patients) are likely to be misdirected and fail to reduce unwarranted variations in the appropriateness and efficiency of service provision (Newhouse et al., 2013).

For future research, this implies the need to reconsider the policy relevance of publishing information on variations at a geographic (e.g. district) level, an increasingly popular practice in many countries in the form of Atlases of Variation (Right Care, 2011) and cross-national projects such as those by the OECD (2014). In order to align the level of analysis with the locus of decision making, future research might follow three directions.

First, research might concentrate on analyses of variations in the quality and cost of care between healthcare providers (e.g. hospitals, ambulatory physicians). This is widely done (e.g. Hauck et al., 2012, Gutacker et al., 2013, Castelli et al., 2015) and undoubtedly valuable where providers are indeed the responsible decision-makers. However, provider-level analyses are inherently limited where quality and cost of care is the result of shared efforts across multiple (e.g. primary, hospital, social) care sectors, as it is typically the case in caring for people with multiple and chronic conditions (Nolte and McKee, 2008).

Second, analysts might focus on actual patient flows between ambulatory providers and hospitals. In the United States, a response to critiques of geographically based analyses has led to a methodology intended to derive physician-hospital clusters in which patients receive most of their care (Bynum et al., 2007, Bynum and Ross, 2013). The units of analysis are hence “networks” of ambulatory physicians and hospitals which are defined empirically based on the extent to which they share patients. This might overcome the arbitrariness of spatial analyses insofar as it focuses on patterns of actual decision-making and care coordination. However, this approach is entirely empirical and provides at best an indirect means of linking analysis and action for improvement (e.g. by making providers aware of their

membership in a "network" so as to compare the quality of information exchange, patterns of utilisation and outcomes with other "networks").

A third direction lies in examining the effects of policies which result in the creation of organisations which are accountable for a broader spectrum of care for their populations, such as Accountable Care Organisations in the United States (Epstein et al., 2014, Luft, 2012). In countries where a specific geographic level of analysis does not correspond to administrative structures for decision making, these organisations provide more meaningful units of analysis and potential loci to achieve high-value care than a focus on single providers. The findings of this thesis will be more relevant to entities where the measurement of variations can be directly linked to decision making.

7.2.3 Methodological considerations

The research was conducted using a mix of qualitative and quantitative methods. The methods chosen have different strengths and weaknesses which have been summarised in the respective Chapters 2 to 6. In principle, qualitative and quantitative approaches should meet similar fundamental requirements. Both should be conducted in a systematic and transparent manner to demonstrate how conclusions were reached, and both should generate internally valid results (Creswell, 2003), defined as the "degree to which findings correctly map the phenomenon in question" (Devers, 1999: 1157). This section discusses some cross-cutting challenges encountered during the research with regard to validity and validation of the findings.

In the more qualitatively oriented Chapters 2 and 6, data collection and analysis were guided by the concept of triangulation, a core strategy in qualitative research to foster internal validity of the results (Yin, 1999, Patton, 1999, Wisdom et al., 2012). Triangulation seeks to achieve a balanced perspective by means of a systematic comparison of insights gained from applying different theoretical lenses or different methods of data collection and analysis (Yin, 1999, Patton, 1999). Chapters 2 and 6 both used forms of method triangulation. Based on a national survey followed by

local interviews, Chapter 2 was able to gain an overview of barriers faced and strategies adopted in different Primary Care Trusts and to subsequently explore these issues in more depth through interviews. Chapter 6, in turn, used a concurrent review of formal national and local policy documents and more informal, interview-based perceptions of national and local managers. Both Chapters relied, if possible, on multiple respondents per organisation. In addition, Chapter 6 triangulated stylised accounts of two theoretical logics of target setting; experimentalist and hierarchist governance. These strategies were valuable because they enabled a clearer understanding of what was gained or omitted from each perspective.

However, triangulation does not guarantee findings will be accurate; these can only be as good as the underlying data. In times of major structural reorganisation in the English NHS, this was particularly a problem for the study reported in Chapter 2 in the form of a low response rate to the national survey. Through a telephone follow-up of non-respondents, it was possible to achieve a response rate of 35% (53 of 151 of PCTs). However, while the findings showed a wide range of responses to the NHS Atlas, it was not possible to conclude whether the distribution of these responses was representative of the totality of PCTs. Finding relatively unobtrusive methods of investigation that enable a comprehensive understanding of local experiences while minimising the burden on the organisations studied should be a priority for future research.

Validation in qualitative research is usually understood in terms of scrutinising the credibility of the findings (Patton, 2002). Because qualitative research is intended to understand people's experiences in context, the group that is best positioned to judge if this intention has been are achieved are the actors within the specific research setting themselves. To mitigate against selective evidence use and misinterpretation by the researcher, the findings of Chapters 2 and 6 were therefore shared with the interviewees and survey respondents who were given the opportunity to comment on the draft and point out any factual errors. This process helped to clarify some issues raised in the initial interviews and also revealed new directions for the analysis.

In the more quantitatively oriented Chapter 3, the phenomenon to be measured was the robustness of performance rankings to alternative sets of assumptions about quality measure weights and choices of population denominator. However, there are other assumptions performance comparisons should be robust to. For instance, in a context of funding under weighted capitation, the impact of environmental influences on measured quality indicators should have already been compensated for by the resource allocation formula. In practice, however, formula funding provides inevitably imperfect compensation for uncontrollable circumstances faced in different local contexts (Smith, 2003). Further research of this problem in a performance measurement context will be important.

Another assumption concerns the choice of healthcare quality measures, measured in Chapter 3 as Boards' relative success in avoiding quality measures experienced by patients. However, performance comparisons also require some concept of the benefit or "value-added" (Goldstein and Spiegelhalter, 1996) that is produced by the different local health systems. This information was not available from existing quality indicators. Chapters 4 and 5 sought to respond to this limitation by examining a specific methodology, capacity to benefit in populations, which might help in moving towards a measure not of the *actual benefit* produced by the health system but, as an essential precondition, of the *potential for benefit* from healthcare.

The epidemiological model developed in Chapter 4 and, as Chapter 5 shows, existing attempts to estimate population capacity to benefit from a defined intervention generally, faced a number of limitations and challenges. These include, most notably, the lack of contemporary, comprehensive epidemiological data on the incidence of specific diseases and the consequent need to extrapolate such information from other populations and time periods. With one exception (Judge et al., 2009), all studies – Chapter 4 included – extrapolated rates of the incidence of a particular disease based on age. Age-specific incidence rates were then adjusted for the proportion of people fulfilling additional criteria of capacity to benefit from a defined intervention. In the case of ventilation tubes for otitis media with effusion, for instance, these criteria included bilaterality of the disease; a hearing level of +25 dB; and time elapsed of three months from the initial diagnosis (NICE Guidance, 2008).

For the epidemiological model in Chapter 4, this was a pragmatic and defensible choice insofar as the relationship between the incidence of otitis media with effusion and age, with two substantial peaks around two and five years of age and considerably lower incidence rates for other age groups, is well-established and most reliably documented (Zielhuis et al., 1990b). However, in spite of the vast literature on otitis media, it was not possible to obtain credible estimates of *how much* of the variation in rates of the incidence of OME is explained by age. Since most diseases are multifactorially determined, applying age-specific incidence rates to other populations hence risks introducing bias due to omitted variables (Mason et al., 2015). Future research should strive to take more predictors of incidence into account, examine how much of the variation in incidence rates is explained by the chosen predictors and investigate consequent uncertainty in estimates of population capacity to benefit through sensitivity analysis.

Finally, the method of validation of models to estimate capacity to benefit in populations merits discussion. Recommendations by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) define validation as the process of checking "how well the model reproduces reality" (Eddy et al., 2012: 843). However, since the purpose of the model in Chapter 4 was to estimate a standard that is independent of reality (i.e. observed rates of utilisation), empirical validation of the model against actual practice would have been fundamentally misconceived. The approach to validation therefore involved a combination of face validation and verification of the internal and technical consistency of the model (Eddy et al., 2012). This involved examining the extent to which the structure of the model and the parameter estimates corresponded to the current state of knowledge, as judged by people with expertise in the field. This was done by conducting a structured "walk-through" (Eddy et al., 2012: 846) in which the separate parts of the model were explained in detail to a panel of experts.

Even with an open and transparent approach to model validation, some uncertainty may remain. Chapter 4 addressed this by using a structured approach to expert elicitation of fractiles of subjective probability distributions (O'Hagan et al., 2006)

which were used as inputs in the model for a Monte Carlo simulation. Following the expert workshop, results of the discussion were summarised and sent to the participants for comment and revision. By their nature, these estimates were subjective. However, in the absence of relevant evidence from the literature, this approach provided a justifiable alternative.

As Box and Draper (1986: 424) lucidly put it, "Essentially, all models are wrong, but some are useful". The contribution of this thesis has been to investigate a series of technical-evidential and socio-political models and strategies intended to address ambiguity about the standard for evaluating the appropriateness of variations in healthcare utilisation. While these models are not perfect, they have the potential to provide a useful stimulus for health policy and management and for future research.

7.3 Concluding remarks and outlook

With increasing pressures on healthcare budgets, evidence of considerable and persistent variations in medical practice offers the opportunity to achieve better outcomes for patients and populations by reducing the inappropriate and inefficient use of resources, to be invested in care of higher value. A fundamental precondition to realise this opportunity lies however in defining and evaluating which variations are unwarranted. By definition, this requires a standard of "good" as opposed to "poor" performance.

Addressing ambiguity about the standard for evaluation is certainly not the only, but definitely an essential step if one seeks to move from the mere measurement of variations to their management. Both technical-evidential and socio-political strategies are required to make this information useful to regulators and managers in charge of planning, auditing and improving the performance of health systems. In the absence of such management strategies, information on variations in healthcare utilisation will likely fail to achieve the anticipated impact on improving health system performance.

In the years and decades to come, health systems will be hard pressed to deliver care that is affordable, of high quality, and accessible to those who would benefit. Using information on geographic variations in healthcare utilisation offers the prospect to help track progress towards the attainment of these objectives. How we measure unwarranted variations in healthcare and how we manage the health system's ability to improve performance in an effective, efficient and responsible manner will remain an enduring task.

REFERENCES

ACHESON, R. 1978. The definition and identification of need for health care. *Journal of Epidemiology and Community Health,* 32**,** 10-5.

ADOLPH, C., GREER, S. & MASSARD DA FONSECA, E. 2012. Allocation of authority in European health policy. *Social Science & Medicine,* 75**,** 1595-603.

AIROLDI, M., MORTON, A., SMITH, J. & BEVAN, G. 2014. STAR—People-Powered Prioritization: A 21st-Century Solution to Allocation Headaches. *Medical Decision Making,* 34**,** 965-975.

ALLEN, R., ATHANASSOPOULOS, A., DYSON, R. G. & THANASSOULIS, E. 1997. Weights restrictions and value judgements in Data Envelopment Analysis: Evolution, development and future directions. *Annals of Operations Research,* 73**,** 13-34.

ANDERSON, R. & MAY, R. 1990. Immunisation and herd immunity. *The Lancet,* 335**,** 641-645.

APPLEBY, J. & MULLIGAN, J. 2000. *How well is the NHS performing? A composite performance indicator based on public consultation,* London, The King's Fund.

APPLEBY, J., RALEIGH, V., FROSINI, F., BEVAN, G., GAO, H. & LYSCOM, T. 2011. *Variations in Health Care. The good, the bad and the inexplicable,* London, King's Fund.

ARON, D., RAJAN, M. & POGACH, L. M. 2007. Summary measures of quality of diabetes care: comparison of continuous weighted performance measurement and dichotomous thresholds. *International Journal for Quality in Health Care,* 19**,** 29-36.

ASTHANA, S. & GIBSON, A. 2011. Setting health care capitations through diagnosis-based risk adjustment: a suitable model for the English NHS? *Health Policy,* 101**,** 133-9.

AUDIT COMMISSION 2011. *Reducing spending on low clinical value treatments,* London, Audit Commission.

AUDIT SCOTLAND 2013. *NHS financial performance 2012/13,* Edinburgh, Audit Scotland.

AUDIT SCOTLAND 2014. *Reshaping care for older people,* Edinburgh, Audit Scotland.

BANKER, R. D., CONRAD, R. F. & STRAUSS, R. P. 1986. A Comparative Application of Data Envelopment Analysis and Translog Methods - an Illustrative Study of Hospital Production. *Management Science,* 32**,** 30-44.

BARBOUR, J., MORTON, A. & SCHANG, L. 2014. The Scottish NHS: meeting the financial challenge ahead. *Fraser of Allander Economic Commentary,* 38**,** 126-146.

BARZELAY, M. 1992. *Breaking through Bureaucracy: A New Vision for Managing in Government,* Berkeley, CA, University of California Press.

BERNAL-DELGADO, E., CHRISTIANSEN, T., BLOOR, K., MATEUS, C., YAZBECK, A., MUNCK, J., BREMNER, J. & ECHO CONSORTIUM 2015. ECHO: health care performance assessment in several European health systems. *European Journal of Public Health,* 25 3-7.

BERNAL-DELGADO, E., GARCÍA-ARMESTO, S. & PEIRÓ, S. 2014. Atlas of Variations in Medical Practice in Spain: The Spanish National Health Service under scrutiny. *Health Policy,* 114**,** 15-30.

BERWICK, D. M. 1996. A primer on leading the improvement of systems. *British Medical Journal,* 312**,** 619-622.

BEVAN, G. 2011. Regulation and system management. *In:* DIXON, A. & MAYS N (eds.) *Understanding New Labour's market reforms of the English NHS.* London: King's Fund, pp. 89-111.
BEVAN, G. & HAMBLIN, R. 2009. Hitting and missing targets by ambulance services for emergency calls: Effects of different systems of performance measurement within the UK. *Journal of the Royal Statistical Society. Series A,* 172**,** 161-190.
BEVAN, G. & HOOD, C. 2006. What's measured is what matters: Targets and gaming in the English public health care system. *Public Administration,* 84**,** 517-538.
BEVAN, G., KARANIKOLOS, M., EXLEY, J., NOLTE, E., CONNOLLY, S. & MAYS, N. 2014. *The four health systems of the United Kingdom: how do they compare?,* London, The Health Foundation and The Nuffield Trust.
BLACK, N. 1985a. Causes of glue ear. An historical review of theories and evidence. *The Journal of Laryngology & Otology,* 99**,** 953-66.
BLACK, N. 1985b. Geographical variations in use of surgery for glue ear. *Journal of the Royal Society of Medicine,* 78**,** 641-8.
BLACK, N. 1985c. Glue ear: the new dyslexia? *British Medical Journal,* 290**,** 1963-5.
BLACKMAN, T., ELLIOTT, E., GREENE, A., HARRINGTON, B., HUNTER, D., MARKS, L., MCKEE, L., SMITH, K. & WILLIAMS, G. 2009. Tackling Health Inequalities in Post-Devolution Britain: Do Targets Matter? *Public Administration,* 87**,** 762-778.
BOGAN, C. & ENGLISH, M. 1994. *Benchmarking for Best Practices: Winning Through Innovative Adaptation,* New York, McGraw-Hil.
BOJKE, C., CASTELLI, A., STREET, A., WARD, P. & LAUDICELLA, M. 2013. Regional variation in the productivity of the English National Health Service. *Health Economics,* 22**,** 194-211.
BOX, G. & DRAPER, N. 1986. *Empirical Model-Building and Response Surfaces,* New York, Wiley.
BOYATZIS, R. 1998. *Thematic Analysis and Code Development. Transforming Qualitative Information,* Thousand Oaks, Sage.
BOYLE, S. 2011. United Kingdom (England): Health System Review. *Health Systems in Transition,* 13**,** 1-486.
BRADSHAW, J. 1972. A taxonomy of social need. *In:* MCLACHLAN, G. (ed.) *Problems and Progress in Medical Care.* Oxford: Oxford University Press.
BRESLOW, N. E. & DAY, N. E. 1987. Rates and standardization. *Statistical methods in cancer research, volume II: The design and analysis of cohort studies.* Lyon: International Agency for Research on Cancer, World Health Organization.
BRIDGES, J., ONUKWUGHA, E., JOHNSON, F. & HAUBER, A. 2007. Patient preference methods - A patient-centered evaluation paradigm. *ISPOR Connections,* 13**,** 4-7.
BRIGGS, A. H., WEINSTEIN, M. C., FENWICK, E. A., KARNON, J., SCULPHER, M. J. & PALTIEL, A. D. 2012. Model Parameter Estimation and Uncertainty Analysis: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. *Medical Decision Making,* 32**,** 722-732.
BRUGNACH, M. & INGRAM, H. 2012. Ambiguity: the challenge of knowing and deciding together. *Environmental Science & Policy,* 15**,** 60-71.
BUCHANAN, C. & POTHIER, D. 2008. Recognition of paediatric otopathology by General Practitioners. *International Journal of Pediatric Otorhinolaryngology,* 72**,** 669-673.
BURCHETT, H., UMOQUIT, M. & DOBROW, M. 2011. How do we know when research from one setting can be useful in another? A review of external validity,

applicability and transferability frameworks. *Journal of Health Services Research & Policy,* 16**,** 238-244.
BUSATO, A., MATTER, P., KUNZE, B. & GOODMAN, D. 2010. Supply sensitive services in Swiss ambulatory care: An analysis of basic health insurance records for 2003-2007. *BMC Health Services Research,* 10**,** Doi 10.1186/1472-6963-10-315.
BUSSE, R., FIGUERAS, J., ROBINSON, R. & JAKUBOWSKI, E. 2007. Strategic Purchasing to Improve Health System Performance: Key Issues and International Trends. *HealthcarePapers,* 8**,** 62-76.
BYNUM, J., BERNAL-DELGADO, E., GOTTLIEB, D. & FISHER, E. 2007. Assigning Ambulatory Patients and Their Physicians to Hospitals: A Method for Obtaining Population-Based Provider Performance Measurements. *Health Services Research,* 42**,** 45-62.
BYNUM, J. & ROSS, J. 2013. A Measure of Care Coordination? *Journal of General Internal Medicine,* 28**,** 336-338.
CAMPBELL, M., FITZPATRICK, R., HAINES, A., KINMONTH, A., SANDERCOCK, P., SPIEGELHALTER, D. & TYRER, P. 2000. Framework for design and evaluation of complex interventions to improve health. *British Medical Journal,* 321**,** 694-696.
CAPER, P. 1987. The epidemiologic surveillance of medical care. *American Journal of Public Health,* 77**,** 669-670.
CARINCI, F., VAN GOOL, K., MAINZ, J., VEILLARD, J., PICHORA, E. C., JANUEL, J. M., ARISPE, I., KIM, S. M. & KLAZINGA, N. S. O. B. O. T. O. H. C. Q. I. E. G. 2015. Towards actionable international comparisons of health system performance: expert revision of the OECD framework and quality indicators. *International Journal for Quality in Health Care,* 27**,** 137-146.
CARR-HILL, R., HARDMAN, G., MARTIN, S., PEACOCK, S., SHELDON, T. & SMITH, P. 1994. *A formula for distributing NHS revenues based on small area use of hospital beds,* York, Centre for Health Economics, University of York.
CARTER, N. 1989. Performance Indicators - Backseat Driving or Hands Off Control. *Policy and Politics,* 17**,** 131-138.
CARTWRIGHT, N. & HARDIE, J. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better,* New York, Oxford University Press.
CASALINO, L. 1999. The Unintended Consequences of Measuring Quality on the Quality of Medical Care. *New England Journal of Medicine,* 341**,** 1147-1150.
CASTELLI, A., STREET, A., VERZULLI, R. & WARD, P. 2015. Examining variations in hospital productivity in the English NHS. *European Journal of Health Economics,* 16**,** 243-254.
CHASSIN, M. R., KOSECOFF, J., PARK, R. E., WINSLOW, C. M., KAHN, K. L., MERRICK, N. J., KEESEY, J., FINK, A., SOLOMON, D. H. & BROOK, R. H. 1987. Does Inappropriate Use Explain Geographic Variations in the Use of Health-Care Services - a Study of 3 Procedures. *Journal of the American Medical Association,* 258**,** 2533-2537.
CHERCHYE, L., MOESEN, W., ROGGE, N. & VAN PUYENBROECK, T. 2007. An introduction to 'benefit of the doubt' composite indicators. *Social Indicators Research,* 82**,** 111-145.
CMS 2009. Centers for Medicare & Medicaid Services. Premier Hospital Quality Incentive Demonstration: Fact sheet.

http://www.cms.hhs.gov/HospitalQualityInits/downloads/HospitalPremierFactSheet200907.pdf [9 May 2014].
COCHRANE, A. 1972. *Effectiveness and efficiency: Random reflections on health services,* London, The Nuffield Provincial Hospitals Trust.
COMMISSION ON THE FUTURE DELIVERY OF PUBLIC SERVICES 2011. *Report on the Future Delivery of Public Services by the Commission chaired by Dr Campbell Christie,* Cheadle Heath, APS Group Scotland.
CONGDON, P. & BEST, N. 2000. Small area variation in hospital admission rates: Bayesian adjustment for primary care and hospital factors. *Journal of the Royal Statistical Society Series C* 49**,** 207-226.
CORALLO, A. N., CROXFORD, R., GOODMAN, D. C., BRYAN, E. L., SRIVASTAVA, D. & STUKEL, T. A. 2014. A systematic review of medical practice variation in OECD countries. *Health Policy,* 114**,** 5-14.
COUCHOUD, C., GUIHENNEUC, C., BAYER, F., LEMAITRE, V., BRUNET, P. & STENGEL, B. 2012. Medical practice patterns and socio-economic factors may explain geographical variation of end-stage renal disease incidence. *Nephrology Dialysis Transplantation,* 27**,** 2312-22.
COX, L. A. 2007. Does concern-driven risk management provide a viable alternative to QRA? *Risk Analysis,* 27**,** 27-43.
COX, L. A. 2012. Confronting Deep Uncertainties in Risk Analysis. *Risk Analysis,* 32**,** 1607-1629.
CRESWELL, J. 2003. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* London, Sage.
CRILLY, T., JASHAPARA, A., TRENHOLM, S., PECKHAM, A., CURRIE, G. & FERLIE, E. 2013. *Knowledge mobilisation in healthcare organisations: Synthesising the evidence and theory using perspectives of organisational form, resource based view of the firm and critical theory,* London, HMSO, NIHR Service Delivery and Organisation programme.
CULYER, A. & WAGSTAFF, A. 1993. Equity and equality in health and health care. *Journal of Health Economics,* 12**,** 431-457.
CULYER, A. J. 1995. Need: the idea won't do-but we still need it. *Social Science & Medicine,* 40**,** 727-30.
CURTIS, S., CONGDON, P., ALMOG, M. & ELLERMANN, R. 2009. County variation in use of inpatient and ambulatory psychiatric care in New York State 1999-2001: Need and supply influences in a structural model. *Health & Place,* 15**,** 568-577.
DANIEL, M., KAMANI, T., EL-SHUNNAR, S., JABEROO, M. C., HARRISON, A., YALAMANCHILI, S., HARRISON, L., CHO, W. S., FERGIE, N., BAYSTON, R. & BIRCHALL, J. P. 2013. National Institute for Clinical Excellence guidelines on the surgical management of otitis media with effusion: are they being followed and have they changed practice? *International Journal of Pediatric Otorhinolaryngology,* 77**,** 54-8.
DAVIES, H. 2005. *Measuring and reporting the quality of health care: issues and evidence from the international research literature,* Edinburgh, NHS Quality Improvement Scotland.
DECANCQ, K. & LUGO, M. A. 2012. Weights in multidimensional indices of wellbeing: An overview. *Econometric Reviews,* 32**,** 7-34.
DELANEY, G., JACOB, S., FEATHERSTONE, C. & BARTON, M. 2005. The role of radiotherapy in cancer treatment - Estimating optimal utilization from a review of evidence-based clinical guidelines. *Cancer,* 104**,** 1129-1137.

DEPARTMENT OF HEALTH 1991. *Care management and assessment: A practitioners' guide,* London, Department of Health.
DEPARTMENT OF HEALTH 2006. *PCT and SHA roles and functions,* London, Department of Health.
DEPARTMENT OF HEALTH 2010a. *Equity and excellence: Liberating the NHS. White Paper,* London, Department of Health.
DEPARTMENT OF HEALTH 2010b. *The Handbook to the NHS Constitution,* London, Department of Health.
DEPARTMENT OF HEALTH 2010c. *The NHS Performance Framework: implementation guidance 2010/11,* http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_115035 [13 May 2012].
DEPARTMENT OF HEALTH 2010d. *The NHS Quality, Innovation, Productivity and Prevention Challenge: an introduction for clinicians,* London, Department of Health.
DEVERS, K. J. 1999. How will we know "good" qualitative research when we see it? Beginning the dialogue in health services research. *Health Services Research,* 34**,** 1153-88.
DIEHR, P., CAIN, K., CONNELL, F. & VOLINN, E. 1990. What is too much variation? The null hypothesis in small-area analysis. *Health Services Research,* 24**,** 741-71.
DIEHR, P. & GREMBOWSKI, D. 1990. A small area simulation approach to determining excess variation in dental procedure rates. *American Journal of Public Health,* 80**,** 1343-8.
DITTO, P. H., JACOBSON, J. A., SMUCKER, W. D., DANKS, J. H. & FAGERLIN, A. 2006. Context Changes Choices: A Prospective Study of the Effects of Hospitalization on Life-Sustaining Treatment Preferences. *Medical Decision Making,* 26**,** 313-322.
DIXON-WOODS, M., MCNICOL, S. & MARTIN, G. 2012. *Overcoming challenges to improving quality,* London, Health Foundation.
DOLAN, P. 1997. Valuing health states: A comparison of methods. *Journal of Health Economics,* 16**,** 617-617.
DONABEDIAN, A. 1978. The quality of medical care. *Science,* 200**,** 856-64.
DONABEDIAN, A. 1981. Criteria, norms and standards of quality: what do they mean? *American Journal of Public Health,* 71**,** 409-12.
DONABEDIAN, A. 1988. The quality of care. How can it be assessed? *Journal of the American Medical Association,* 260**,** 1743-8.
DOWD, B., SWENSON, T., KANE, R., PARASHURAM, S. & COULAM, R. 2014. CAN DATA ENVELOPMENT ANALYSIS PROVIDE A SCALAR INDEX OF 'VALUE'? *Health Economics,* 23**,** 1465-1480.
DRUCKER, P. 1954. *Practice of management,* New York, Harper.
DRUMMOND, M., BARBIERI, M., COOK, J., GLICK, H. A., LIS, J., MALIK, F., REED, S. D., RUTTEN, F., SCULPHER, M. & SEVERENS, J. 2009. Transferability of Economic Evaluations Across Jurisdictions: ISPOR Good Research Practices Task Force Report. *Value in Health,* 12**,** 409-418.
EDDY, D. M., HOLLINGWORTH, W., CARO, J. J., TSEVAT, J., MCDONALD, K. M. & WONG, J. B. 2012. Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. *Medical Decision Making,* 32**,** 733-743.

EDWARDS, A., ELWYN, G., COVEY, J., MATTHEWS, E. & PILL, R. 2001. Presenting Risk Information: A Review of the Effects of Framing and other Manipulations on Patient Outcomes. *Journal of Health Communication,* 6**,** 61-82.
ELWYN, G., LAITNER, S., COULTER, A., WALKER, E., WATSON, P. & THOMSON, R. 2010. Implementing shared decision making in the NHS. *British Medical Journal,* 341: c5146. doi: http://dx.doi.org/10.1136/bmj.c5146.
EPSTEIN, A. M., JHA, A. K., ORAV, E. J., LIEBMAN, D. L., AUDET, A.-M. J., ZEZZA, M. A. & GUTERMAN, S. 2014. Analysis Of Early Accountable Care Organizations Defines Patient, Structural, Cost, And Quality-Of-Care Characteristics. *Health Affairs,* 33**,** 95-102.
ERNST, K., WISMAR, M., BUSSE, R. & MCKEE, M. 2008. Improving the Effectiveness of Health Targets. *In:* WISMAR, M., MCKEE, M., ERNST, K., SRIVASTAVA, D. & BUSSE, R. (eds.) *Health Targets in Europe: Learning from experience.* Copenhagen: World Health Organization, on behalf of the European Observatory on Health Systems and Policies.
ETTELT, S., FAZEKAS, M., MAYS, N. & NOLTE, E. 2012. Assessing health care planning – A framework-led comparison of Germany and New Zealand. *Health Policy,* 106**,** 50-59.
EUROSTAT 2014. *Morbidity Statistics in the EU - Report on pilot studies,* Luxembourg, Publications Office of the European Union.
EVANS, R. 1990. The Dog in the Night-Time. *In:* ANDERSEN TV & MOONEY G (eds.) *The Challenges of Medical Practice Variation* London: MacMillan.
EZZATI, M., LOPEZ, A., RODGERS, A. & MURRAY, C. 2004. *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors,* Geneva, World Health Organization.
FARRAR, S., HARRIS, F., SCOTT, T. & MCKEE, L. 2004. *The Performance Assessment Framework: experiences and perceptions of NHSScotland*, A Report to the Analytical Service Division, Directorate of Performance Management and Finance, Scottish Executive Health Department.
FEAR, J., HILLMAN, M., CHAMBERLAIN, M. & TENNANT, A. 1997. Prevalence of hip problems in the population aged 55 years and over: access to specialist care and future demand for hip arthroplasty. *Rheumatology,* 36**,** 74-76.
FERGUSON, B., GRAVELLE, H., DUSHEIKO, M., SUTTON, M. & JOHNS, R. 2002. Variations in practice admission rates: the policy relevance of regression standardisation. *Journal of Health Services Research & Policy,* 7**,** 170-6.
FERRIS, G., RODERICK, P., SMITHIES, A., GEORGE, S., GABBAY, J., COUPER, N. & CHANT, A. 1998. An epidemiological needs assessment of carotid endarterectomy in an English health region. Is the need being met? *British Medical Journal,* 317**,** 447-51.
FIELLAU-NIKOLAJSEN, M. 1983. Epidemiology of secretory otitis media. A descriptive cohort study. *Annals of Otology, Rhinology & Laryngology,* 92**,** 172-7.
FIERLBECK, K. 2014. The changing contours of experimental governance in European health care. *Social Science & Medicine,* 108**,** 89-96.
FISCHER, M. & FERLIE, E. 2013. Resisting hybridisation between modes of clinical risk management: Contradiction, contest, and the production of intractable conflict. *Accounting, Organizations and Society,* 38**,** 30-49.
FOLLAND, S. & STANO, M. 1990. Small area variations: a critical review of propositions, methods, and evidence. *Medical Care Review,* 47**,** 419-65.

FONG, A., SHAFIQ, J., SAUNDERS, C., THOMPSON, A. M., TYLDESLEY, S., OLIVOTTO, I. A., BARTON, M. B., DEWAR, J. A., JACOB, S., NG, W., SPEERS, C. & DELANEY, G. P. 2012. A comparison of surgical and radiotherapy breast cancer therapy utilization in Canada (British Columbia), Scotland (Dundee), and Australia (Western Australia) with models of "optimal" therapy. *Breast,* 21**,** 570-577.

FOROUDI, F., TYLDESLEY, S., BARBERA, L., HUANG, J. & MACKILLOP, W. J. 2003. An evidence-based estimate of the appropriate radiotherapy utilization rate for colorectal cancer. *International Journal of Radiation Oncology Biology Physics,* 56**,** 1295-1307.

FOSSUM, J. E. 2012. Reflections on experimentalist governance. *Regulation & Governance,* 6**,** 394-400.

FOSTER, J. & SEN, A. 1997. *On Economic Inequality,* Oxford, Oxford University Press.

FOSTER, J. E., MCGILLIVRAY, M. & SETH, S. 2012. Composite Indices: Rank Robustness, Statistical Association, and Redundancy. *Econometric Reviews,* 32**,** 35-56.

FRANKEL, S. 1991. The epidemiology of indications. *Journal of Epidemiology and Community Health,* 45**,** 257-9.

FRANKEL, S., EACHUS, J., PEARSON, N., GREENWOOD, R., CHAN, P., PETERS, T. J., DONOVAN, J., SMITH, G. D. & DIEPPE, P. 1999. Population requirement for primary hip-replacement surgery: a cross-sectional study. *Lancet,* 353**,** 1304-9.

FREEMAN, T. 2002. Using performance indicators to improve health care quality in the public sector: a review of the literature. *Health Services Management Research,* 15**,** 126-37.

FREY, H. & PATIL, S. 2002. Identification and Review of Sensitivity Analysis Methods. *Risk Analysis,* 22**,** 553-578.

FRIA, T. J., CANTEKIN, E. I. & EICHLER, J. A. 1985. Hearing acuity of children with otitis media with effusion. *Archives of Otolaryngology,* 111**,** 10-6.

FRIED, T., BYERS, A., GALLO, W., VAN NESS, P., TOWLE, V., O'LEARY, J. & DUBIN, J. 2006. Prospective study of health status preferences and changes in preferences over time in older adults. *Archives of Internal Medicine,* 166**,** 890-895.

FROST, A., HOPPER, C., FRANKEL, S., PETERS, T. J., DURANT, J. & SPARROW, J. 2001. The population requirement for cataract extraction: A cross-sectional study. *Eye,* 15**,** 745-752.

GAULD, R., HORWITT, J., WILLIAMS, S. & COHEN, A. 2011. What Strategies Do US Hospitals Employ to Reduce Unwarranted Clinical Practice Variations? *American Journal of Medical Quality* 26 120-126.

GIBSON, A., ASTHANA, S., BRIGHAM, P., MOON, G. & DICKER, J. 2002. Geographies of need and the new NHS: methodological issues in the definition and measurement of the health needs of local populations. *Health Place,* 8**,** 47-60.

GLASZIOU, P. & HAYNES, B. 2005. The paths from research to improved health outcomes. *Evidence Based Nursing,* 8**,** 36-38.

GLOVER, J. A. 1938. The Incidence of Tonsillectomy in Schoolchildren. *Proceedings of the Royal Society of Medicine,* 31**,** 1219–36.

GODDARD, M. & JACOBS, R. 2009. Using composite indicators to measure performance in health care. *In:* SMITH, P., MOSSIALOS, E., PAPANICOLAS, I. & LEATHERMAN, S. (eds.) *Performance measurement for health system improvement: experiences, challenges and prospects.* Cambridge: Cambridge University Press, pp. 339-368.

GODDARD, M., MANNION, R. & SMITH, P. 2000. Enhancing performance in health care: a theoretical perspective on agency and the role of information. *Health Economics,* 9**,** 95-107.
GOLDSTEIN, H. & SPIEGELHALTER, D. J. 1996. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society Series A -Statistics in Society,* 159**,** 385-409.
GOODMAN, D. C. 2009. Unwarranted variation in pediatric medical care. *Pediatric Clinics of North America,* 56**,** 745-55.
GÖPFFARTH, D., KOPETSCH, T. & SCHMITZ, H. 2015. Determinants of Regional Variation in Health Expenditures in Germany. *Health Economics,* Article first published online: 12 May 2015.
GOUVEIA, M., DIAS, L., ANTUNES, C., MOTA, M., DUARATE, E. & TENREIRO, E. 2015. An application of value-based DEA to identify the best practices in primary health care. *OR Spectrum. DOI 10.1007/s00291-015-0407-x*.
GRAVELLE, H., JACOBS, R., JONES, A. M. & STREET, A. 2003. Comparing the efficiency of national health systems: A sensitivity analysis of the WHO approach. *Applied Health Economics and Health Policy,* 2**,** 141-7.
GREENHALGH, T., PLSEK, P., WILSON, T., FRASER, S. & HOLT, T. 2010. Response to 'The appropriation of complexity theory in health care'. *Journal of Health Services Research & Policy,* 15**,** 115-117.
GUEST, G., BUNCE, A. & JOHNSON, L. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field Methods,* 18**,** 59-82.
GUILLEN, Ú., DEMAURO, S., MA, L., ZUPANCIC, J., WANG, E., GAFNI, A. & KIRPALANI, H. 2011. Survival rates in extremely low birthweight infants depend on the denominator: avoiding potential for bias by specifying denominators. *American Journal of Obstetrics and Gynecology,* 205**,** 329.e1-329.e7.
GUINEY, H., FELICIA, P., WHELTON, H. & WOODS, N. 2012. Comparing epidemiologically estimated treatment need with treatment provided in two dental schemes in Ireland. *BMC Oral Health,* 12**,** doi:10.1186/1472-6831-12-31.
GUTACKER, N., BOJKE, C., DAIDONE, S., DEVLIN, N. & STREET, A. 2013. Hospital variation in patient-reported outcomes at the level of EQ-5D dimensions: evidence from England. *Medical Decision Making,* 33**,** 804 - 818.
HAGGARD, M. 2009. Air-conduction estimated from tympanometry (ACET): 2. The use of hearing level-ACET discrepancy (HAD) to determine appropriate use of bone-conduction tests in identifying permanent and mixed impairments. *International Journal of Pediatric Otorhinolaryngology,* 73**,** 43-55.
HAI TASKFORCE 2008. *Healthcare Associated Infection Taskforce Delivery Plan April 2008 to March 2011,* http://www.scotland.gov.uk/Publications/2008/03/07110818/0 [5 April 2014].
HALEY, R. W., CULVER, D. H., WHITE, J. W., MORGAN, W. M., EMORI, T. G., MUNN, V. P. & HOOTON, T. M. 1985. The Efficacy of Infection Surveillance and Control Programs in Preventing Nosocomial Infections in United-States Hospitals. *American Journal of Epidemiology,* 121**,** 182-205.
HAM, C. 2013. Doctors must lead efforts to reduce waste and variation in practice. *British Medical Journal,* 346**,** doi: http://dx.doi.org/10.1136/bmj.f3668
HASMAN, A., HOPE, T. & OSTERDAL, L. P. 2006. Health care need: three interpretations. *J Appl Philos,* 23**,** 145-56.

HAUCK, K. & STREET, A. 2006. Performance assessment in the context of multiple objectives: a multivariate multilevel analysis. *Journal of Health Economics,* 25**,** 1029-48.
HAUCK, K., ZHAO, X. & JACKSON, T. 2012. Adverse event rates as measures of hospital performance. *Health Policy,* 104**,** 146-54.
HAWKER, G. A., WRIGHT, J. G., COYTE, P. C., WILLIAMS, J. I., HARVEY, B., GLAZIER, R., WILKINS, A. & BADLEY, E. M. 2001. Determining the need for hip and knee arthroplasty: the role of clinical severity and patients' preferences. *Medical Care,* 39**,** 206-16.
HEALTH PROTECTION SCOTLAND 2012. *National Infection Prevention and Control Manual Version 2.3., updated March 2014,* Glasgow, Health Protection Scotland.
HEALTH PROTECTION SCOTLAND 2014. *Healthcare Associated Infection Annual Report 2013,* Glasgow, Health Protection Scotland.
HEALTHCARE COMMISSION 2005. *2005 performance ratings,* London, Healthcare Commission.
HOLLINGSWORTH, B. & STREET, A. 2006. The market for efficiency analysis of health care organisations. *Health Economics,* 15**,** 1055-1059.
HOLLINGWORTH, W., ROOSHENAS, L., BUSBY, J., HINE, C., BADRINATH, P., WHITING, P., MOORE, T., OWEN-SMITH, A., STERNE, J., JONES, H., BEYNON, C. & DONOVAN, J. 2015. Using clinical practice variations as a method for commissioners and clinicians to identify and prioritise opportunities for disinvestment in health care: a cross-sectional study, systematic reviews and qualitative study. *Health Services and Delivery Research,* 3**,** a-169.
HOOD, C. 1991. A Public Management for All Seasons. *Public Administration,* 69**,** 3-19.
HOOD, C. 2007. Public service management by numbers: Why does it vary? Where has it come from? What are the gaps and the puzzles? *Public Money & Management,* 27**,** 95-102.
HOOD, C. 2012. Public Management by Numbers as a Performance-Enhancing Drug: Two Hypotheses. *Public Administration Review,* 72**,** S85-S92.
HOOD, C. & JACKSON, M. 1994. Keys for Locks in Administrative Argument. *Administration & Society,* 25**,** 467-488.
HUESCH, M., ONG, M. & GOLDMAN, D. 2013. *Policy approaches to addressing geographic variation in spending, utilization, and high value care and the implications of those approaches,* Washington, Institute of Medicine.
HUNTER, D. J. W., GRANT, H. J., PURDUE, M. P. H., SPASOFF, R. A., DORLAND, J. L. & BAINS, N. 2004. An epidemiologically-based needs assessment for stroke services. *Chronic Diseases in Canada,* 25**,** 138-146.
HURST, J. 2002. Performance measurement and improvement in OECD health systems: Overview of issues and challenges. *In:* SMITH, P. (ed.) *Measuring Up: Improving Health System Performance in OECD Countries.* Paris: OECD, pp. 35-54.
HURST, J. & JEE-HUGHES, M. 2001. *Performance measurement and performance management in OECD health systems. Labour market and social policy - Occasional papers 47,* Paris, OECD.
HURST, J. & WILLIAMS, S. 2012. *Can NHS hospitals do more with less?,* London, Nuffield Trust.
HUSSEY, P. S., DE VRIES, H., ROMLEY, J., WANG, M. C., CHEN, S. S., SHEKELLE, P. G. & MCGLYNN, E. A. 2009. A systematic review of health care efficiency measures. *Health Services Research,* 44**,** 784-805.

INSTITUTE OF MEDICINE 2001. *Crossing the quality chasm; a new health system for the 21st century,* Washington DC, National Academy Press.
ISD SCOTLAND 2010a. The Resource Allocation Formula in Scotland. http://www.isdscotland.org/Health-Topics/Finance/Resource-Allocation-Formula/ [11 November 2014].
ISD SCOTLAND 2010b. Resource allocation. http://www.isdscotland.org/Health-Topics/Finance/Resource-Allocation-Formula/information.asp [11 November 2014].
ISD SCOTLAND 2014. *HEAT TARGET: Emergency Admissions for Patients Aged 75+ (Numbers, Bed Days & Rates per 1,000 population),* Edinburgh, Information Services Division.
JACOB, S., WONG, K., DELANEY, G. P., ADAMS, P. & BARTON, M. B. 2010. Estimation of an optimal utilisation rate for palliative radiotherapy in newly diagnosed cancer patients. *Clinical Oncology,* 22**,** 56-64.
JACOBS, R., GODDARD, M. & SMITH, P. 2005. How robust are hospital ranks based on composite performance measures? *Medical Care,* 43**,** 1177-84.
JACOBS, R., MARTIN, S., GODDARD, M., GRAVELLE, H. & SMITH, P. 2006a. Exploring the determinants of NHS performance ratings: lessons for performance assessment systems. *Journal of Health Services Research & Policy,* 11**,** 211-217.
JACOBS, R., SMITH, P. & STREET, A. 2006b. *Measuring efficiency in health care: analytic techniques and health policy*, Cambridge University Press.
JEFFREYS, B. 2010. Variation in amputation rate 'shocking', 25 November 2010. *BBC News*.
JIT (JOINT IMPROVEMENT TEAM) 2014. *Annual Report 2013/14,* Edinburgh, Joint Improvement Team.
JIT (JOINT IMPROVEMENT TEAM) 2015. Immediate Discharge Service. http://www.jitscotland.org.uk/example-of-practice/immediate-discharge-service/ [20 February 2015].
JOHNSTON, L., LARDNER, C. & JEPSON, R. 2008. *Overview of evidence relating to shifting the balance of care: a contribution to the knowledge base,* Edinburgh, Scottish Government Social Research.
JORDAN, K., CLARKE, A. M., SYMMONS, D. P., FLEMING, D., PORCHERET, M., KADAM, U. T. & CROFT, P. 2007. Measuring disease prevalence: a comparison of musculoskeletal disease using four general practice consultation databases. *British Journal of General Practice,* 57**,** 7-14.
JUDGE, A., WELTON, N. J., SANDHU, J. & BEN-SHLOMO, Y. 2009. Modeling the need for hip and knee replacement surgery. Part 2. Incorporating census data to provide small-area predictions for need with uncertainty bounds. *Arthritis Care and Research,* 61**,** 1667-1673.
JÜNI, P., DIEPPE, P., DONOVAN, J., PETERS, T., EACHUS, J., PEARSON, N., GREENWOOD, R. & FRANKEL, S. 2003. Population requirement for primary knee replacement surgery: A cross-sectional study. *Rheumatology,* 42**,** 516-521.
KANG, H. C. & HONG, J. S. 2011. Do differences in profiling criteria bias performance measurements? Economic profiling of medical clinics under the Korea National Health Insurance program: an observational study using claims data. *BMC Health Services Research,* 11**,** doi:10.1186/1472-6963-11-189.
KEYHANI, S., FALK, R., BISHOP, T., HOWELL, E. & KORENSTEIN, D. 2012. The relationship between geographic variations and overuse of healthcare services: a systematic review. *Medical Care,* 50**,** 257-61.

KEYHANI, S., KLEINMAN, L. C., ROTHSCHILD, M., BERNSTEIN, J. M., ANDERSON, R. & CHASSIN, M. 2008a. Overuse of tympanostomy tubes in New York metropolitan area: evidence from five hospital cohort. *British Medical Journal,* 337**,** doi: 10.1136/bmj.a1607.

KEYHANI, S., KLEINMAN, L. C., ROTHSCHILD, M., BERNSTEIN, J. M., ANDERSON, R. & CHASSIN, M. 2008b. Overuse of tympanostomy tubes in New York metropolitan area: evidence from five hospital cohort. *British Medical Journal,* 337**,** a1607.

LAFFONT, J. & MARTIMORT, D. 2002. *The Theory of Incentives: The Principal-Agent Model,* Princeton, Princeton University Press.

LAWTON, A., MCKEVITT, D. & MILLAR, M. 2000. Coping with ambiguity: Reconciling external legitimacy and organizational implementation in performance measurement. *Public Money & Management,* 20**,** 13-19.

LE GRAND, J. 2007. The Other Invisible Hand: Delivering Public Services Through Choice and Competition. Woodstock: Princeton University Press.

LE GRAND, J. 2010. Knights and Knaves Return: Public Service Motivation and the Delivery of Public Services. *International Public Management Journal,* 13**,** 56-71.

LEAPE, L. L., PARK, R. E., SOLOMON, D. H., CHASSIN, M. R., KOSECOFF, J. & BROOK, R. H. 1990. Does Inappropriate Use Explain Small-Area Variations in the Use of Health-Care Services? *Journal of the American Medical Association,* 263**,** 669-672.

LEGIDO-QUIGLEY, H., PANTELI, D., BRUSAMENTO, S., KNAI, C., SALIBA, V., TURK, E., SOLE, M., AUGUSTIN, U., CAR, J., MCKEE, M. & BUSSE, R. 2012. Clinical guidelines in the European Union: mapping the regulatory basis, development, quality control, implementation and evaluation across member states. *Health Policy,* 107**,** 146-56.

LESKE, M. C., EDERER, F. & PODGOR, M. 1981. Estimating incidence from age-specific prevalence in glaucoma. *American Journal of Epidemiology,* 113**,** 606-13.

LILFORD, R. 2009. Should the NHS strive to eradicate all unexplained variation? No. *British Medical Journal,* 339**,** doi: 10.1136/bmj.b4811.

LUFT, H. S. 2012. From Small Area Variations to Accountable Care Organizations: How Health Services Research Can Inform Policy. *Annual Review of Public Health,* 33**,** 377-392.

MACKENBACH, J. P. 2012. The persistence of health inequalities in modern welfare states: The explanation of a paradox. *Social Science & Medicine,* 75**,** 761-769.

MAJEED, A., ELIAHOO, J., BARDSLEY, M., MORGAN, D. & BINDMAN, A. B. 2002. Variation in coronary artery bypass grafting, angioplasty, cataract surgery, and hip replacement rates among primary care groups in London: association with population and practice characteristics. *Journal of Public Health Medicine,* 24**,** 21-26.

MANNION, R. & GODDARD, M. 2001. Impact of published clinical outcomes data: case study in NHS hospital trusts. *British Medical Journal,* 323**,** 260-3.

MANNION, R. & GODDARD, M. 2003. Public disclosure of comparative clinical performance data: lessons from the Scottish experience. *Journal of Evaluation in Clinical Practice,* 9**,** 277-286.

MARLOW, A. 1995. Potential years of life lost: what is the denominator? *Journal of Epidemiology & Community Health,* 49**,** 320-2.

MARMOR, T. 2012. The unwritten rules of cross-national policy analysis. *Health Economics, Policy and Law,* 7**,** 19-20.
MARTIN, R. M., HEMINGWAY, H., GUNNELL, D., KARSCH, K. R., BAUMBACH, A. & FRANKEL, S. 2002. Population need for coronary revascularisation: Are national targets for England credible? *Heart,* 88**,** 627-633.
MASON, T., SUTTON, M., WHITTAKER, W. & BIRCH, S. 2015. Exploring the limitations of age-based models for health care planning. *Social Science & Medicine,* 132**,** 11-19.
MAYNARD, A. 2012. Deal with productivity variation, or risk the long term future of the NHS. *Health Service Journal,* Article first published online: 12 January 2012.
MAYNARD, A. 2013. Health Care Rationing: Doing It Better in Public and Private Health Care Systems. *Journal of Health Politics Policy and Law,* 38**,** 1103-1127.
MAYS, N. 2006. *Use of Targets to Improve Health System Performance: English NHS Experience and Implications for New Zealand,* Wellington, New Zealand Treasury.
MAYS, N. 2011. Reducing unwarranted variations in healthcare in the English NHS. *British Medical Journal,* 342**,** doi: 10.1136/bmj.d1849.
MCDERMOTT, A., HAMEL, L., STEEL, D., FLOOD, P. & MCKEE, L. 2015. Hybrid healthcare governance for improvement? Combining top-down and bottom-up approaches to public sector regulation. *Public Administration,* Article first published online: 19 January 2015.
MCGLYNN, E. 1998. *Assessing the Appropriateness of Care: How Much is Too Much?,* Santa Monica, CA, Rand Corporation.
MCKEE, M. 1996. Health Needs Assessment. *In:* JANOVSKY, K. (ed.) *Health policy and systems development.* Geneva: World Health Organization, pp. 61-78.
MCKEE, M. & CLARKE, A. 1995. Guidelines, enthusiasms, uncertainty, and the limits to purchasing. *British Medical Journal,* 310**,** 101-104.
MCKIBBEN, L., HORAN, T., TOKARS, J. I., FOWLER, G., CARDO, D. M., PEARSON, M. L., BRENNAN, P. J. & THE HEALTHCARE INFECTION CONTROL PRACTICES ADVISORY, C. 2005. Guidance on Public Reporting of Healthcare-Associated Infections: Recommendations of the Healthcare Infection Control Practices Advisory Committee. *American Journal of Infection Control,* 33**,** 217-226.
MCPHERSON, K., WENNBERG, J. E., HOVIND, O. B. & CLIFFORD, P. 1982. Small-area variations in the use of common surgical procedures: an international comparison of New England, England, and Norway. *New England Journal of Medicine,* 307**,** 1310-4.
MCQUEEN, D., WISMAR, M., LIN, V., JONES, C. & DAVIES, M. 2012. *Intersectoral Governance for Health in All Policies. Structures, actions and experiences. Observatory Studies Series No.26,* Copenhagen, World Health Organization, on behalf of the European Observatory on Health Systems and Policies.
MELTZER, D. O. & CHUNG, J. W. 2014. The population value of quality indicator reporting: a framework for prioritizing health care performance measures. *Health Affairs,* 33**,** 132-9.
MERCURI, M., BIRCH, S. & GAFNI, A. 2013. Using small-area variations to inform health care service planning: what do we 'need' to know? *Journal of Evaluation in Clinical Practice,* 19**,** 1054-9.
MERCURI, M. & GAFNI, A. 2011. Medical practice variations: what the literature tells us (or does not) about what are warranted and unwarranted variations. *Journal of Evaluation in Clinical Practice,* 17**,** 671-677.

MILNER, P. C., PAYNE, J. N., STANFIELD, R. C., LEWIS, P. A., JENNISON, C. & SAUL, C. 2004. Inequalities in accessing hip joint replacement for people in need. *European Journal of Public Health,* 14**,** 58-62.
MOHER, D., LIBERATI, A., TETZLAFF, J. & ALTMAN, D. G. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *British Medical Journal,* 339**,** doi: 10.1136/bmj.g7647.
MOONEY, G. & HOUSTON, S. 2004. An alternative approach to resource allocation. *Applied Health Economics and Health Policy,* 3**,** 29-33.
MOONEY, G. H. & BLACKWELL, S. H. 2004. Whose health service is it anyway? Community values in healthcare. *Medical Journal of Australia,* 180**,** 76-78.
MORGAN, M. & HENRION, M. 1990. *Uncertainty. A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis,* Cambridge, Cambridge University Press.
MORGENSTERN, H., KLEINBAUM, D. & KUPPER, L. 1980. Measures of disease incidence used in epidemiologic research. *International Journal of Epidemiology,* 9**,** 97-104.
MORRIS, S., CARR-HILL, R., DIXON, P., LAW, M., RICE, N., SUTTON, M. & VALLEJO-TORRES, L. 2007. Combining Age Related and Additional Needs (CARAN) Report. The 2007 Review of the Needs Formulae for Hospital Services and Prescribing Activity in England. ACRA(2007)22.
NAESSENS, J. M. & HUSCHKA, T. R. 2004. Distinguishing hospital complications of care from pre-existing conditions. *International Journal for Quality in Health Care,* 16**,** I27-I35.
NATIONAL AUDIT OFFICE 2011. Assessing value for money. *In:* NATIONAL AUDIT OFFICE (ed.) *Successful Commissioning Toolkit.* http://www.nao.org.uk/successful-commissioning/general-principles/value-for-money/assessing-value-for-money/ [12 December 2012].
NAVARRO, V. 2000. Assessment of the world health report 2000. *Lancet,* 356**,** 1598-1601.
NEWHOUSE, J., GARBER, A., GRAHAM, R., MCCOY, M., MANCHER, M. & KIBRIA, A. 2013. *Variation in Health Care Spending: Target Decision Making, Not Geography,* Washington, Institute of Medicine.
NHS ENGLAND 2015. *Revised Never Events Policy and Framework,* London, NHS England.
NHS INFORMATION CENTRE 2011. *Main procedures and interventions: 4 character. Hospital Episode Statistics for England. Inpatient statistics, 2010-11. Procedure D15.1 Myringotomy with insertion of ventilation tube through tympanic membrane,* The Health and Social Care Information Centre, www.hesonline.nhs.uk [15 December 2012].
NHS RIGHT CARE 2010. *The NHS Atlas of Variation in Healthcare 2010,* http://www.rightcare.nhs.uk/index.php/atlas/atlas-of-variation-2010/ [4 March 2012].
NHS RIGHT CARE 2011. *The NHS Atlas of Variation in Healthcare 2011,* London, NHS Right Care.
NHS RIGHT CARE 2012a. *International Atlases,* London, NHS Right Care.
NHS RIGHT CARE 2012b. *NHS Atlas of Variation in Healthcare for Children and Young People,* London, NHS Right Care.
NHS TAYSIDE 2011. *LDP Risk Management Plan 2011/12,* Dundee, NHS Tayside.
NICE 2011. *Services for people with chronic obstructive pulmonary disease,* London, National Institute for Health and Clinical Excellence.

NICE 2013. *Cardiac rehabilitation services,* London, National Institute for Health and Care Excellence.
NICE GUIDANCE 2008. *Surgical management of children with otitis media with effusion (OME)* Clinical guidelines, CG60. http://guidance.nice.org.uk/CG60 [23 January 2013].
NICHOLL, J., JACQUES, R. M. & CAMPBELL, M. J. 2013. Direct risk standardisation: a new method for comparing casemix adjusted event rates using complex models. *BMC Medical Research Methodology,* 13**,** doi 10.1186/1471-2288-13-133.
NOLTE, E. & MCKEE, M. 2008. Integration and chronic care: a review. *In:* NOLTE E & MCKEE M (eds.) *Caring for people with chronic conditions. A health system perspective.* Copenhagen: World Health Organization on behalf of the European Observatory on Health Systems and Policies, pp. 64-91.
NOLTING, H., ZICH, K., DECKENBACH B, GOTTBERG A, LOTTMANN K, KLEMPERER D, GROTE WESTRICK M & SCHWENK U 2011. *Faktencheck Gesundheit. Regionale Unterschiede in der Gesundheitsversorgung,* Gütersloh, Bertelsmann Stiftung.
NOONAN, K. G., SABEL, C. F. & SIMON, W. H. 2009. Legal Accountability in the Service-Based Welfare State: Lessons from Child Welfare Reform. *Law and Social Inquiry-Journal of the American Bar Foundation,* 34**,** 523-568.
NUTLEY, S., WALTER, I. & DAVIES, H. 2007. *Using evidence: How research can inform public services,* Bristol, The Policy Press.
O'CONNOR A, LLEWELLYN-THOMAS HA & FLOOD AB 2004. Modifying Unwarranted Variations In Health Care: Shared Decision Making Using Patient Decision Aids. *Health Affairs* Web exclusive**,** doi: 10.1377/hlthaff.var.63.
O'HAGAN, A., BUCK, C., DANESHKHAH, A., EISER, J., GARTHWAITE, P., JENKINSON, D., OAKLEY, J. & RAKOW, T. 2006. *Uncertain Judgements. Eliciting Experts' Probabilities,* Chichester, England, Wiley.
OECD 2008. *Handbook on Constructing Composite Indicators,* Paris, OECD.
OECD (ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT) 2014. *Geographic Variations in Health Care: What Do We Know and What Can Be Done to Improve Health System Performance?,* Paris, OECD publishing.
OED 2015. Standard, n. and adj. *Oxford English Dictionary Online version March 2015.* http://www.oed.com/view/Entry/188962?rskey=ocI9mb&result=1&isAdvanced=false#eid [2 April 2015].
OFFICE FOR NATIONAL STATISTICS 2011. *Primary Care Organisations Population Estimates (experimental) - mid-2010,* http://www.ons.gov.uk/ons/rel/sape/pco-pop-est-exp/mid-2010-release/pco-mid-2010.html [25 May 2012]. Break-down by age groups obtained by personal communication [ONS SAPE, 28 February 2012].
OFFICE FOR NATIONAL STATISTICS 2012. Expenditure on healthcare in the UK: 2011. www.ons.gov.uk [12 August 2013].
OLIVER, A. & MOSSIALOS, E. 2004. Equity of access to health care: outlining the foundations for action. *Journal of Epidemiology and Community Health,* 58**,** 655-658.
OSBERG, L. & SHARPE, A. 2002. An index of economic well–being for selected OECD countries. *Review of Income and Wealth,* 48**,** 291-316.
PAPANICOLAS, I. & SMITH, P. 2014. Theory of system level efficiency in health care. *In:* CULYER, A. (ed.) *Encyclopedia of Health Economics, Volume 3.* Philadelphia, USA: Elsevier, pp. 386-394.

PARENTE, S. T., PHELPS, C. E. & O'CONNOR, P. J. 2008. Economic analysis of medical practice variation between 1991 and 2000: the impact of patient outcomes research teams (PORTs). *International Journal of Technology Assessment in Health Care,* 24**,** 282-93.
PATTON, M. 1990. *Qualitative evaluation and research methods,* Newbury Park, Sage Publications.
PATTON, M. 1999. Enhancing the quality and credibility of qualitative analysis. *Health Services Research,* 34**,** 1189-1208.
PATTON, M. 2002. *Qualitative research & evaluation methods,* Thousand Oaks, Sage
PAUL-SHAHEEN, P., CLARK, J. & WILLIAMS, D. 1987. Small Area Analysis: A Review and Analysis of the North American Literature. *Journal of Health Politics, Policy and Law,* 12**,** 741-809.
PBRA TEAM 2009. *Developing a person-based resource allocation formula for allocations to general practices in England,* London, The Nuffield Trust.
PEDRAJA-CHAPARRO, F., SALINAS-JIMENEZ, J. & SMITH, P. 1997. On the Role of Weight Restrictions in Data Envelopment Analysis. *Journal of Productivity Analysis,* 8**,** 215-230.
PIDD, M. 1996. *Tools for Thinking: Modelling in Management Science,* Chichester, John Wiley & Sons.
PIRES, R. 2011. Beyond the fear of discretion: Flexibility, performance, and accountability in the management of regulatory bureaucracies. *Regulation & Governance,* 5**,** 43–69.
POLANYI, M. 1966. *The Tacit Dimension,* London, Routledge.
POLLITT, C., HARRISON, S., DOWSWELL, G., JERAK-ZUIDERENT, S. & BAL, R. 2010. Performance Regimes in Health Care: Institutions, Critical Junctures and the Logic of Escalation in England and the Netherlands. *Evaluation,* 16**,** 13-29.
POPPER, K. 1948. What can logic do for philosophy? Logical positivism and ethics. *The Symposia Read at the Joint Session of the Aristotelian Society and the Mind Association at Durham, July 9-11.* London: Harrison and Sons.
PORTER, M. 2010. What Is Value in Health Care? *New England Journal of Medicine,* 363**,** doi 10.1056/NEJMp1011024.
PORTER, M. & TEISBERG, E. 2006. *Redefining health care: creating value-based competition on results,* Boston, Mass., Harvard Business School Press.
POWER, M. 1999. *The Audit Society: Rituals of Verification,* Oxford, Oxford University Press.
PRICE, C. E., PAUL, E. A., BEVAN, R. G. & HOLLAND, W. W. 1992. Equity and medical practice variation: relationships between standardised discharge ratios in total and for selected conditions in English districts. *Journal of Epidemiology and Community Health,* 46**,** 58-62.
PROPPER, C., SUTTON, M., WHITNALL, C. & WINDMEIJER, F. 2008. Did 'targets and terror' reduce waiting times in England for hospital care? *BE Journal of Economic Analysis & Policy,* 8**,** 1-27.
QUEST (QUALITY AND EFFICIENCY SUPPORT TEAM) 2014. *Annual Report 2013: Reporting on the Quality and Efficiency Support Team,* Edinburgh, The Scottish Government.
REAY, T. & HININGS, C. 2009. Managing the Rivalry of Competing Institutional Logics. *Organization Studies,* 30**,** 629-652.

REEVES, D., CAMPBELL, S. M., ADAMS, J., SHEKELLE, P. G., KONTOPANTELIS, E. & ROLAND, M. O. 2007. Combining multiple indicators of clinical quality: an evaluation of different analytic approaches. *Medical Care,* 45**,** 489-96.
REIN, M. 1976. *Social Science and Public Policy,* New York, Penguin.
RETTENMAIER, A. J. & WANG, Z. 2012. Regional variations in medical spending and utilization: a longitudinal analysis of US Medicare population. *Health Economics,* 21**,** 67-82.
RITCHIE, J. & SPENCER, L. 1994. Qualitative Data Analysis for Applied Policy Research. *In:* BRYMAN, A. & BURGESS, R. (eds.) *Analyzing Qualitative Data.* Taylor & Francis Books Ltd, pp. 173–194.
RITTEL, H. W. J. & WEBBER, M. M. 1973. Dilemmas in a General Theory of Planning. *Policy Sciences,* 4**,** 155-169.
ROBSON, C. 2002. *Real world research: a resource for social scientists and practitioner-researchers,* Oxford, U.K. Malden, Mass., Blackwell.
ROLAND, P., FINITZO, T., FRIEL-PATTY, S., CLINTON BROWN, K., STEPHENS, K., BROWN, O. & COLEMAN, M. 1989. Otitis Media. Incidence, Duration, and Hearing Status. *Archives of Otolaryngology—Head & Neck Surgery,* 115**,** 1049-1053.
ROMANO, P., HUSSEY, P. & RITLEY, D. 2010. *Selecting Quality and Resource Use Measures: A Decision Guide for Community Quality Collaboratives,* Rockville, Agency for Healthcare Research and Quality.
ROSENFELD, R. M., GOLDSMITH, A. J. & MADELL, J. R. 1998. How accurate is parent rating of hearing for children with otitis media? *Archives of Otolaryngology-Head & Neck Surgery,* 124**,** 989-992.
ROSENKÖTTER, N. & VAN BON-MARTENS, M. 2015. Public health monitoring and reporting: Maintaining and improving the evidence base. *Eurohealth,* 21**,** 17-20.
ROVERS, M. M., SCHILDER, A. G., ZIELHUIS, G. A. & ROSENFELD, R. M. 2004. Otitis media. *Lancet,* 363**,** 465-73.
RUSSELL, J., GREENHALGH, T., LEWIS, H., MACKENZIE, I., MASKREY, N., MONTGOMERY, J. & O'DONNELL, C. 2013. Addressing the 'postcode lottery' in local resource allocation decisions: a framework for clinical commissioning groups. *Journal of the Royal Society of Medicine,* 106**,** 120-123.
SABEL, C. 2004. Beyond principal-agent governance: experimentalist organizations, learning and accountability. *In:* ENGELEN, E. & SIE DHIAN HO, M. (eds.) *De staat van de democratie. Democratie voorbij de staat.* Amsterdam: Amsterdam University Press, pp. 173-196.
SABEL, C. & ZEITLIN, J. 2012. Experimentalist Governance. *In:* LEVI-FAUR, D. (ed.) *The Oxford Handbook of Governance.* Oxford: Oxford University Press, pp. 169-183.
SABEL, C. F. & ZEITLIN, J. 2008. Learning from difference: The new architecture of experimentalist governance in the EU. *European Law Journal,* 14**,** 271-327.
SABO, D. L., PARADISE, J. L., KURS-LASKY, M. & SMITH, C. G. 2003. Hearing levels in infants and young children in relation to testing technique, age group, and the presence or absence of middle-ear effusion. *Ear and Hearing* 24**,** 38-47.
SALO, A. & PUNKKA, A. 2011. Ranking intervals and dominance relations for ratio-based efficiency analysis. *Management Science,* 57**,** 200-214.
SALTELLI, A., RATTO, M., ANDRES, T., CAMPOLONGO, F., CARIBONI, J., GATELLI, D., SAISANA, M. & TARANTOLA 2008. *Global Sensitivity Analysis: The Primer*, Wiley E-book.

SANDERSON, C. F., HUNTER, D. J., MCKEE, C. M. & BLACK, N. A. 1997. Limitations of epidemiologically based needs assessment. The case of prostatectomy. *Medical Care,* 35**,** 669-85.
SCHANG, L., DE POLI, C., AIROLDI, M., MORTON, A., BOHM, N., LAKHANPAUL, M., SCHILDER, A. & BEVAN, G. 2014a. Using an epidemiological model to investigate unwarranted variation: the case of ventilation tubes for otitis media with effusion in England. *Journal of Health Services Research & Policy,* 19**,** 236-44.
SCHANG, L. & MORTON, A. 2012. *LSE/ Right Care project on NHS Commissioners' use of the NHS Atlas of Variation in Healthcare. Case studies of local uptake,* London, NHS Right Care and Department of Management, LSE.
SCHANG, L., MORTON, A., DASILVA, P. & BEVAN, G. 2014b. From data to decisions? Exploring how healthcare payers respond to the NHS Atlas of Variation in Healthcare in England. *Health Policy,* 114**,** 79-87.
SCHLAUD, M., BRENNER, M. H., HOOPMANN, M. & SCHWARTZ, F. W. 1998. Approaches to the denominator in practice-based epidemiology: a critical overview. *Journal of Epidemiology & Community Health,* 52**,** 13S-19S.
SCHÖN, D. 1987. *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions,* San Francisco, Jossey-Bass.
SCOTT, W., RUEF, M., MENDEL, P. & CARONNA, C. 2000. *Institutional change and healthcare organizations: From professional dominance to managed care,* Chicago, The University of Chicago Press.
SCOTTISH EXECUTIVE 2005. *The Unscheduled Care Collaborative Programme,* Edinburgh, Scottish Executive.
SCOTTISH GOVERNMENT 2014. NHS Boards. http://www.scotland.gov.uk/Topics/Health/NHS-Workforce/NHS-Boards [14 May 2014].
SCOTTISH PUBLIC HEALTH OBSERVATORY 2014. *Population: 2013 mid-year population estimates by NHS board and age categories,* http://www.scotpho.org.uk/population-dynamics/population-estimates-and-projections/data/nhs-board-population-estimates [2 May 2015].
SHLEIFER, A. 1985. A Theory of Yardstick Competition. *Rand Journal of Economics,* 16**,** 319-327.
SIMPSON, S. A., THOMAS, C. L., VAN DER LINDEN, M. K., MACMILLAN, H., VAN DER WOUDEN, J. C. & BUTLER, C. 2007. Identification of children in the first four years of life for early treatment for otitis media with effusion. *Cochrane Database of Systematic Reviews,* 24**,** CD004163.
SMITH, D. M., PEARCE, J. R. & HARLAND, K. 2011. Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? An example of smoking prevalence in New Zealand. *Health & Place,* 17**,** 618-624.
SMITH, P. 1995. On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration,* 18**,** 277-310.
SMITH, P. 1997. Model misspecification in data envelopment analysis. *Annals of Operations Research,* 73**,** 233-252.
SMITH, P. 2002. Developing composite indicators for assessing health system efficiency. *In:* SMITH, P. (ed.) *Measuring up: improving health system performance in OECD countries.* Paris: OECD, pp. 295-318.

SMITH, P. 2005. Performance measurement in health care: History, challenges and prospects. *Public Money & Management,* 25**,** 213-220.
SMITH, P., MOSSIALOS, E., PAPANICOLAS, I. & LEATHERMAN, S. (eds.) 2009. *Performance measurement for health system improvement: experiences, challenges and prospects,* Cambridge: Cambridge University Press.
SMITH, P. & PAPANICOLAS, I. 2012. *Health system performance comparison: an agenda for policy, information and research. Policy Summary 4,* Copenhagen, World Health Organization.
SMITH, P. C. 2003. Formula funding of public services: An economic analysis. *Oxford Review of Economic Policy,* 19**,** 301-322.
SMITH, P. C. & STREET, A. 2005. Measuring the efficiency of public services: the limits of analysis. *Journal of the Royal Statistical Society Series A-Statistics in Society,* 168**,** 401-417.
SOLBERG, L. I., MOSSER, G. & MCDONALD, S. 1997. The three faces of performance measurement: Improvement, accountability and research. *Joint Commission Journal on Quality Improvement,* 23**,** 135-147.
SORENSON, C., DRUMMOND, M. & KANAVOS, P. 2008. *Ensuring value for money in health care: The role of health technology assessment in the European Union,* Copenhagen, World Health Organization, on behalf of the European Observatory on Health Systems and Policies.
SPIEGELHALTER, D. J. 1999. Surgical audit: statistical lessons from Nightingale and Codman. *Journal of the Royal Statistical Society Series A-Statistics in Society,* 162**,** 45-58.
STEEL, D. 2013. Scotland. *In:* HAM, C., HEENAN, D., LONGLEY, M. & STEEL, D. (eds.) *Integrated care in Northern Ireland, Scotland and Wales. Lessons for England.* London: The King's Fund, pp. 25-56.
STEEL, D. & CYLUS, J. 2012. United Kingdom (Scotland): Health system review. *Health Systems in Transition,* 14**,** xv-xxii, 1-150.
STEVENS, A. & GILLAM, S. 1998. Needs assessment: from theory to practice. *British Medical Journal,* 316**,** 1448-52.
SUTTON, M., GRAVELLE, H., MORRIS, S., LEYLAND, A., WINDMEIJER, F., DIBBIN, C. & MUIRHEAD, M. 2002. Allocation of Resources to English Areas: Individual and Small Area Determinants of Morbidity and Use of Health Care. Report for Department of Health (AREA Report).
TANENBAUM, S. J. 2012. Reducing Variation in Health Care: The Rhetorical Politics of a Policy Idea. *Journal of Health Politics, Policy and Law,* 38**,** 5-26.
TENNANT, A., FEAR, J., PICKERING, A., HILLMAN, M., CUTTS, A. & CHAMBERLAIN, M. A. 1995. Prevalence of knee problems in the population aged 55 years and over: identifying the need for knee arthroplasty. *British Medical Journal,* 310**,** 1291-3.
TEUTSCH, S. M., BERGER, M. L. & WEINSTEIN, M. C. 2005. Comparative Effectiveness: Asking The Right Questions, Choosing The Right Method. *Health Affairs,* 24**,** 128-132.
THE DARTMOUTH INSTITUTE. 2012. *The Dartmouth Atlas of Health Care* [Online]. The Dartmouth Institute for Health Policy and Clinical Practice. http://www.dartmouthatlas.org/ [3 January 2012].
THE GUARDIAN 2011. NHS postcode lottery survey reveals wide UK disparities. The Guardian 9 December 2011.

http://www.theguardian.com/society/2011/dec/09/nhs-lottery-survey-uk-disparities [12 December 2012].
THE SCOTTISH GOVERNMENT 2009a. *Improving Outcomes by Shifting the Balance of Care,* Edinburgh, The Scottish Government.
THE SCOTTISH GOVERNMENT 2009b. *Long Term Conditions Collaborative. Programme 2008-2011,* Edinburgh, The Scottish Government.
THE SCOTTISH GOVERNMENT 2010. *Reshaping Care for Older People. A Programme of Change 2011 -2021,* Edinburgh, The Scottish Government.
THE SCOTTISH GOVERNMENT 2013. *A Route Map to the 2020 Vision for Health and Social Care,* Edinburgh, The Scottish Government.
THOMSEN, J. & TOS, M. 1981. Spontaneous improvement of secretory otitis. A long-term study. *Acta Otolaryngologica,* 92**,** 493-9.
THOMSON, S., FIGUERAS, J., EVETOVITS, T., JOWETT, M., MLADOVSKY, P., MARESSO, A., CYLUS, J., KARANIKOLOS, M. & KLUGE, H. 2014. *Economic crisis, health systems and health in Europe: impact and implications for policy, Policy summary 12,* Copenhagen, WHO Regional Office for Europe.
TOS, M. 1984. Epidemiology and natural history of secretory otitis. *American Journal of Otolaryngology,* 5**,** 459-62.
TREAGUST, J., MORKANE, T. & SPEAKMAN, M. 2001. Estimating a population's needs for the treatment of lower urinary tract symptoms in men: what is the extent of unmet need? *Journal of Public Health,* 23**,** 141-147.
TVERSKY, A. & KAHNEMAN, D. 1991. Loss Aversion in Riskless Choice - a Reference-Dependent Model. *Quarterly Journal of Economics,* 106**,** 1039-1061.
TYLDESLEY, S., BOYD, C., SCHULZE, K., WALKER, H. & MACKILLOP, W. 2001. Estimating the need for radiotherapy for lung cancer: an evidence-based, epidemiologic approach. *International Journal of Radiation Oncology, Biology, Physics,* 49**,** 973-985.
UNGKANONT, K., CHARULUXANANAN, S. & KOMOLTRI, C. 2010. Association of otoscopic findings and hearing level in pediatric patients with otitis media with effusion. *International Journal of Pediatric Otorhinolaryngology,* 74**,** 1063-6.
USMANI, N., FOROUDI, F., DU, J., ZAKOS, C., CAMPBELL, H., BRYSON, P. & MACKILLOP, W. J. 2005. An evidence-based estimate of the appropriate rate of utilization of radiotherapy for cancer of the cervix. *International Journal of Radiation Oncology Biology Physics,* 63**,** 812-827.
VAN DER WEES, P. J., NIJHUIS-VAN DER SANDEN, M. W., VAN GINNEKEN, E., AYANIAN, J. Z., SCHNEIDER, E. C. & WESTERT, G. P. 2014. Governing healthcare through performance measurement in Massachusetts and the Netherlands. *Health Policy,* 116**,** 18-26.
VANDENBROUCKE, J. P., VON ELM, E., ALTMAN, D. G., GOTZSCHE, P. C., MULROW, C. D., POCOCK, S. J., POOLE, C., SCHLESSELMAN, J. J., EGGER, M. & INITIATIVE, S. 2014. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *International Journal of Surgery,* 12**,** 1500-24.
WEINSTEIN, J. N., CLAY, K. & MORGAN, T. S. 2007. Informed Patient Choice: Patient-Centered Valuing Of Surgical Risks And Benefits. *Health Affairs,* 26**,** 726-730.
WENNBERG, J. 2010. *Tracking Medicine: A Researcher's Quest to Understand Health Care,* New York, Oxford University Press.

WENNBERG, J. & GITTELSOHN, A. 1973. Small area variations in health care delivery. *Science,* 182**,** 1102-8.
WHITE, J. 2011. Prices, Volume, and the Perverse Effects of the Variations Crusade. *Journal of Health Politics Policy and Law,* 36**,** 775-790.
WHITEHURST, D. G. T., BRYAN, S., LEWIS, M., HAY, E. M., MULLIS, R. & FOSTER, N. E. 2015. Implementing Stratified Primary Care Management for Low Back Pain: Cost-Utility Analysis Alongside a Prospective, Population-Based, Sequential Comparison Study. *Spine,* 40**,** 405-414.
WHITWORTH, A. 2013. *Evaluations and improvements in small area estimation methodologies,* Sheffield, National Centre for Research Methods.
WHO 2000. *The world health report 2000 - Health systems: improving performance,* Geneva, World Health Organization.
WILLIAMSON, I. G., DUNLEAVEY, J., BAIN, J. & ROBINSON, D. 1994. The natural history of otitis media with effusion: A three-year study of the incidence and prevalence of abnormal tympanograms in four South West Hampshire infant and first schools. *The Journal of Laryngology & Otology,* 108**,** 930-4.
WISDOM, J. P., CAVALERI, M. A., ONWUEGBUZIE, A. J. & GREEN, C. A. 2012. Methodological Reporting in Qualitative, Quantitative, and Mixed Methods Health Services Research Articles. *Health Services Research,* 47**,** 721-745.
WRIGHT, J., DUGDALE, B., HAMMOND, I., JARMAN, B., NEARY, M., NEWTON, D., PATTERSON, C., RUSSON, L., STANLEY, P., STEPHENS, R. & WARREN, E. 2006. Learning from death: a hospital mortality reduction programme. *Journal of the Royal Society of Medicine,* 99**,** 303–308.
WRIGHT, J., WILLIAMS, R. & WILKINSON, J. R. 1998. Health needs assessment - Development and importance of health needs assessment. *British Medical Journal,* 316**,** 1310-1313.
YIN, R. 2003. *Case study research: design and methods,* Thousand Oaks, Sage.
YIN, R. K. 1999. Enhancing the quality of case studies in health services research. *Health Services Research,* 34**,** 1209-1224.
YONG, P., MILNER, P., PAYNE, J., LEWIS, P. & JENNISON, C. 2004. Inequalities in access to knee joint replacements for people in need. *Annals of the Rheumatic Diseases,* 63**,** 1483-1489.
ZHAN, C. L., ELIXHAUSER, A., FRIEDMAN, B., HOUCHENS, R. & CHIANG, Y. P. 2007. Modifying DRG-PPS to include only diagnoses present on admission - Financial implications and challenges. *Medical Care,* 45**,** 288-291.
ZHANG, X., HAUCK, K. & ZHAO, X. 2013. Patient safety in hospitals - a Bayesian analysis of unobservable hospital and specialty level risk factors. *Health Economics,* 22**,** 1158-74.
ZHOU, P., ANG, B. W. & ZHOU, D. Q. 2010. Weighting and aggregation in composite indicator construction: A multiplicative optimization approach. *Social Indicators Research,* 96**,** 169-181.
ZIELHUIS, G., STRAATMAN, H., RACH, G. & VAN DEN BROEK, P. 1990a. Analysis and presentation of data on the natural course of otitis media with effusion in children. *International Journal of Epidemiology,* 19**,** 1037-44.
ZIELHUIS, G. A., RACH, G. H., VAN DEN BOSCH, A. & VAN DEN BROEK, P. 1990b. The prevalence of otitis media with effusion: a critical review of the literature. *Clinical Otolaryngology & Allied Sciences,* 15**,** 283-8.

ZIELHUIS, G. A., RACH, G. H. & VAN DEN BROEK, P. 1990c. The natural course of otitis media with effusion in preschool children. *European Archives of Oto-Rhino-Laryngology,* 247**,** 215-21.
ZIELHUIS, G. A., RACH, G. H. & VAN DEN BROEK, P. 1990d. The occurrence of otitis media with effusion in Dutch pre-school children. *Clinical Otolaryngology & Allied Sciences,* 15**,** 147-53.