FOUR ESSAYS ON PERFORMANCE MEASURES BASED ON PATIENT-REPORTED OUTCOMES

Nils Gutacker

Doctor of Philosophy

University of York
Economics

August 2015

Abstract

Agency relationships, and associated information asymmetries, exist in many areas of economic activity including healthcare. Information on healthcare providers' relative performance can be used to reduce information asymmetries and hold providers to account. This collection of essays focuses on the appropriate derivation and use of performance measures to incentivise healthcare providers in the English National Health Service (NHS). It gives special consideration to the role of patient self-reported health status measures to assess the differential effect of healthcare providers' care on their patients' health.

The thesis explores three themes: the relationship between variation in resource use and quality, the appropriate assessment and reporting of multidimensional hospital performance, and the use of performance information to motivate hospitals in a public reporting context.

Chapter 2 examines cost variation between hospitals for the four surgical procedures covered by the national patient-reported outcome measures (PROM) programme. It explores the empirical relationship between costs and patient health outcomes to assess the claim of hospital providers that their higher costs are justified by better quality of care.

Chapter 3 sets out an empirical methodology to conduct provider performance comparisons when there are multiple dimensions of health-related quality of life affected by treatment. It discusses the advantages and disadvantages of analysing disaggregate PROM data for the purpose of informing prospective patients, clinicians and managers.

Chapter 4 extends the previous chapter by providing a methodology for assessing and summarising multidimensional provider performance using dominance criteria. This methodology is then applied to study the performance of providers of hip replacement surgery with respect to length of stay, emergency readmissions, waiting time and improvements in PROMs.

Chapter 5 estimates the demand elasticity of providers with respect to quality. It makes use of choice models to assess the usefulness of disseminating hospital PROM scores to prospective patients as a market-based incentive for providers to compete on quality.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my gratitude to a number of people for their help and support throughout the six years of my part-time PhD research. Most importantly, I would like to thank my supervisor, Andrew Street, who provided guidance, expertise, encouragement and support. He helped me to stay focused when needed but also allowed me to experiment and explore and, ultimately, become an independent researcher. I am also grateful to the members of my advisory panel, Maria Goddard and Luigi Siciliani, for their guidance and advice. I thank my examiners, Alastair Gray, Rowena Jacobs and Alistair McGuire, for an interesting discussion and many helpful comments.

The Centre for Health Economics at the University of York provided a friendly and intellectually stimulating environment to carry out my research. I would like to thank all my colleagues, fellow PhD students and visiting researchers - they all made invaluable contributions.

My greatest thanks go to my family, especially my partner Berenice Villanueva and our beautiful daughter Jana. Without you, I could not have done it. This thesis is dedicated to you.

Author's declaration

I confirm that the work presented in this thesis is my own, except where co-authorship is explicitly acknowledged. Funding for my studies was provided by the Centre for Health Economics at the University of York, the Department of Health in England through its policy research programmes, and the National Institute for Health Research (NIHR).

Chapters 1 and 6 are sole-authored.

Chapter 2 is written in co-authorship with Dr Chris Bojke, Dr Silvio Daidone, Professor Nancy Devlin, Professor David Parkin and Professor Andrew Street. I am the main author of this essay, having defined the theoretical and empirical model, assembled the data, constructed the variables, carried out the empirical analysis and written up the paper. All co-authors provided advice and comments during the development of the work and were involved in editing the paper. The work has been funded under a Department of Health Policy Research Programme (Reference: 027/0038) and has been published as a peer-reviewed research article under the title: *Truly inefficient or providing better quality of care? Analysing the relationship between risk-adjusted hospital costs and patients' health outcomes. Health Economics 2013; 22(8): pp. 931-947*. Some minor revisions have been made subsequent to publication following improvements in the data extraction algorithm.

Chapter 3 is written in co-authorship with Dr Chris Bojke, Dr Silvio Daidone, Professor Nancy Devlin and Professor Andrew Street. I am the main author of this essay, having defined the theoretical and empirical model, assembled the data, constructed the variables, carried out the empirical analysis and written up the paper. All co-authors provided advice and comments during the development of the work and were involved in editing the paper. The work has been funded under the NIHR Health Services Research (HSR) stream (Project number: 09/2000/47). It has been published as a peer-reviewed research article (*Hospital variation in patient-reported outcomes at the level of EQ-5D dimensions - Evidence from England. Medical Decision Making 2013; 33(6): pp. 804-818*) and as a peer-reviewed report to the funder (*Variations in outcome and costs among NHS providers for common surgical procedures: econometric analyses of routinely collected data. Health Services and Delivery Research 2014:2(1)*). The work was awarded the EuroQol prize for best scientific paper in 2014. Some minor revisions have been made subsequent to publication following improvements in the data extraction algorithm.

Chapter 4 is written in co-authorship with Professor Andrew Street. I am the main author of this essay, having conceived the original research idea, defined the theoretical and empirical model, assembled the data, constructed the variables, carried out the empirical analysis and written up the paper. Professor Street provided advice and comments during the development of the work and was involved in editing the paper. The work has been funded under the Policy Research Unit in the Economics of Health and Social Care Systems (Reference: 103/0001) and is currently under review at the *Journal of Health Economics* (since May 2015).

Chapter 5 is written in co-authorship with Professor Luigi Siciliani, Dr Giuseppe Moscelli and Professor Hugh Gravelle. It is available as a working paper under the title: *Do patients choose hospitals that improve their health? Centre for Health Economics Research Paper 111*. I am the main author of this essay, having defined the empirical model, assembled the data, constructed the variables, carried out the empirical analysis and written up the paper. All co-authors provided advice and comments during the development of the work and were involved in editing the paper. The work has been funded under the Policy Research Unit in the Economics of Health and Social Care Systems (Reference: 103/0001) and is currently under review at the *Journal of Health Economics* (since November 2015).

Chapters 2 to 5 have been presented at a number of national and international conferences, as listed in the acknowledgements to each chapter.

I affirm that this thesis has not previously been presented to any other university or educational institution for examination. In addition, any views expressed in this document are exclusive responsibility of the author.

1 Introduction

1.1 Economic framework

This collection of essays focuses on the appropriate derivation and use of performance measures to incentivise healthcare providers in the English National Health Service (NHS), giving special consideration to the role of patient self-reported health status measures to assess the differential effect of healthcare providers' care on their patients' health. The relevant economic framework for these analyses is a principal-agent framework.

Agency relationships, in which a principal delegates a task to an agent in return for a reward, exist in many areas of economic activity (Sappington 1991; Laffont and Tirole 1993). Examples include car owners paying mechanics to carry out repairs or homeowners hiring decorators to paint their living rooms. A common feature of these relationships is that the agent enjoys an information advantage over the principal with respect to the likely costs and outcomes of the task, and that the principal cannot directly verify the appropriateness of the actions taken by the agent. This information asymmetry derives from the nature of the agency relationship and the principal's lack of specialist knowledge about the production function and associated constraints, and the optimal mix of inputs.[1] Hence, agents are able to extract information rent, for example, by reducing the amount of costly effort they exert while maintaining their agreed reward; a form of *hidden action*.[2] This prevents

[1] As pointed out by Arrow (1968, p.538) the "*agent has been selected for his specialized knowledge and the principal can never hope to completely check the agent's performance*". If the principal was in possession of the required knowledge and abilities, she could perform the task herself without risking rent extraction. Of course, she may still decide against it for a number of other reasons such as a more constrained production environment.

[2] This behaviour may not always be conscious and opportunistic. For example, the agent may simply

the first-best allocation of resources that maximises the principal's utility.

The market for healthcare is especially prone to agency problems due to the complexity of healthcare and the pronounced information asymmetry between the principal (e.g. a patient, purchaser or regulator of care[3]) and the agent (a provider of care such as a doctor, nurse, or, more generally, a healthcare institution) (Arrow 1963; Evans 1974; Ryan 1992). For example, patients generally have little knowledge of the type of illness they suffer from or how urgently they require treatment, what treatment options are available and which treatment is most appropriate for their condition. Furthermore, because healthcare contributes to health but is not the only input in the health production function – health behaviour, other consumption and random variation also play into this; see Grossman (1972) – there is considerable uncertainty about the likely health effect of care. Similar issues arise in the agency relationships between purchasers or regulators of care and healthcare providers. For example, these principals can rarely verify whether costly diagnostic or therapeutic actions taken by the agent were appropriate given the patients' condition, or indeed whether the reported diagnosis reflects the health condition of the patient (Dafny 2005). That said, some mechanisms such as altruism, the high status that medical professionals enjoy in society, and peer review through other medical professionals, may help align the interests of principals and agents and reduce the incentive for rent extraction.

One key insight from the literature on incentive contracts is that information about agents' comparative performance can be incorporated into contracts to re-

not be aware of more effective ways of carrying out the task because he or she failed to invest sufficiently into keeping abreast of the evidence base. Nevertheless, because such investments are typically costly, i.e. they may require time, effort and expenditure on learning materials, and cannot be directly observed this may be considered an information rent.

[3] In many healthcare systems patients do not bear the full cost of care but share it with a public or private insurer. These insurers will often use their greater bargaining power and act as purchasers of care. Consequently, both patients and purchasers enter into agency agreements with providers of care. Similar agency problems are likely to arise at least in general, although both the principals' objectives and the degree of information asymmetry towards the agent may differ. For example, a purchaser will want to strike a balance between the amount of effort assigned to quality and cost containment. In contrast, since patients are protected from most costs they will put more emphasis on the agent's effort to provide high quality of care. For the purpose of this introduction I do not distinguish between different types of principals and may use one or the other to illustrate concepts.

duce information asymmetry and improve the principal's utility (Holmström 1979; Holmström 1982; Shleifer 1985; Arrow 1986). Such performance information can be derived from comparing the observed outcome of an agent against those of other agents or against a predetermined target; a practice known as performance assessment or benchmarking. If all agents were to face identical production constraints and experience common shocks, variation in outcome could be attributed to the effort that agents exert (Holmström 1982). Because circumstances are rarely identical comparisons are typically preformed within a multiple regression framework to isolate performance variation from observable exogenous influences (e.g. case-mix differences between providers) and random noise (Shleifer 1985; Ash et al. 2012). The resulting performance information can then be utilised in multiple ways to incentivise agents. For example, agents' rewards can be adjusted retrospectively in the context of a pay-for-performance (P4P) contract with rewards and penalties according to observed performance. Alternatively, performance information can be used prospectively to inform contracts, e.g. by influencing the choice of agent to contract with in the first place. Finally, the public dissemination of performance information can provide non-pecuniary incentives against shirking, which might be effective if the agent cares about her reputation (Hibbard et al. 2003; Hibbard et al. 2005).

Performance assessment, public or private reporting of comparative performance data and P4P schemes have now become common features of many healthcare systems (Smith 2002; Marshall et al. 2003; Maynard 2012).[4] Many of these schemes are explicitly concerned about variation in the quality of care that agents provide although some focus exclusively on costs or a combination of costs and quality. For example, since April 2004 the UK Quality and Outcome Framework has been rewarding general practitioners according to the proportion of their practice population with chronic illness which receives care that meets defined process standards (Roland

[4] The use of performance data to motivate agents is not restricted to healthcare; see Propper and Wilson (2003) and Prendergast (1999) for examples from the wider public sector and the private sector.

2004). The underlying data are also made available to the public and may be used by patients to choose practices (Santos et al. 2015). Similarly, the Advancing Quality programme in the North West of England rewarded[5] hospitals for their relative performance with regard to 28 quality measures covering five clinical areas (Sutton et al. 2012; Meacock et al. 2014). While the effectiveness and cost-effectiveness of such initiatives is contested (e.g. Petersen et al. 2006; Fung et al. 2008; Emmert et al. 2012; Maynard 2012), their popularity is growing.

1.2 Measuring quality of care through patient-reported outcomes

An important issue in the implementation of performance assessment regimes is how to define and measure the relative quality of care provided. In his seminal work on the definition of quality of care Donabedian (1966) delineated three broad dimensions: structural quality (i.e. the characteristics of the care environment), process quality (the manner in which care is provided) and outcome quality (the change in patients' health as a result of care). Many elements of structural quality, such as the availability of a computer tomography scanner, are observable and thus easily contractible. The choice between process and outcome quality remains a point of contention between medical professionals and economists. Medical professionals are often reluctant to be assessed on the basis of outcomes since the link between healthcare and health is not straightforward and the outcome of care is thus uncertain. Instead, there seems to be a preference for process measures that are felt to be more directly under the control of the professional (Lilford and Pronovost 2010). Economists, on the other hand, argue that healthcare is primarily a means to improve health or avoid future deteriorations of it (Evans 1974; Porter 2010). Patients derive utility from their health and the consumption that good health allows. The relevant concept is therefore the change in health trajectory (i.e. the difference in cumulative health and health-related quality of life over the life

[5] The programme ran from October 2008 to March 2010 before being subsumed into a national P4P programme.

course) that patients experience as the result of treatment[6] (Smith and Street 2013) - although process utility from e.g. reassurance may also be of some importance (McGuire et al. 1988; Mooney and Lange 1991). The change in health trajectory has also been described as '*the "value-added" [...] as a result of the contact with the health system*' (Jacobs et al. 2006, p.23). Of course, establishing the change in health trajectory is a formidable task, not least because patients cannot be observed simultaneously on their treated and untreated health trajectories and for a sufficiently long follow-up period.[7] However, in a comparative performance framework with agents providing the same type of treatment, patients' health trajectories if untreated do not necessarily need to be known since these are assumed to be the same for all patients, conditional on observed pre-operative patient characteristics (Smith and Street 2013).[8]

The reluctance of the medical profession to adopt changes in health trajectory as the primary measure of their quality performance may in part stem from the lack of sufficiently discriminating and routinely available measures of patient health. Many healthcare systems rely on administrative data for the assessment of provider performance. Historically, patient outcomes had therefore been confined to measures of mortality after treatment over short periods of time (e.g. 30 days post-operatively). More recently, rates of emergency readmission and severe adverse events have gained importance and are increasingly used to adjust payments (Rosenthal 2007; Department of Health 2012b).[9] However, all these measures have a number of limitations (Appleby and Devlin 2004). First, for many commonly performed healthcare

[6] This is not to say that structural and process quality measures do not play an important role in managing the quality of care within institutions, e.g. by hospitals managers seeking to identify problems in the care process. As pointed out by Donabedian (1988, p.1745), '*good structure increases the likelihood of good process, and good process increases the likelihood of good outcome*'. However, improving structural and process quality are a means to an end and '*outcomes, by and large, remain the ultimate validators of the effectiveness and quality of medical care*' (Donabedian 1966, p.169).

[7] Randomised controlled trials and extrapolation within a modelling framework can help overcome these challenges. However, at least randomisation is not a feasible approach to assess the performance of healthcare providers in routine care settings.

[8] If, however, the aim of the analysis is to establish the productivity of the agent or compare the cost-effectiveness of agents across different treatments, the absolute change in health would be required.

[9] These do not measure patients' health directly but indicate deteriorations of unknown magnitude.

interventions these outcomes occur rarely. For example, the 30-day mortality rate after elective hip replacement surgery in the English NHS is approximately 3 per 1,000 patients (Berstock et al. 2014). This complicates any statistical analysis of performance since noise and signal are more difficult to differentiate. Second, the ability to adjust these outcomes for patient heterogeneity (i.e. a provider's case-mix) and thus create a level playing field for comparisons is often limited by the available data recorded in routine administrative records. The case-mix adjustment in many performance assessment programmes is confined to a limited number of patient characteristics, such as age, gender and co-morbidity burden, whereas information on pre-treatment health status and severity are generally absent. Incomplete case-mix adjustment may give rise to adverse behaviour such as refusal to treat high risk patients in order to improve measured performance (Dranove et al. 2003). Third, the effect of an adverse event or emergency re-admission on patients' health may differ across patient groups, yet these measures are silent about their impact on patients' health. Fourth, all these measures focus on negative outcomes and are not informative about the size of the health improvement that the vast majority of patients experience. Finally, health is multi-dimensional and different providers may have a differential impact on these health dimensions. Indicators based on mortality, re-admissions or adverse events cannot reveal this.

The limitations of the existing measures of outcome have led to calls for routine collection of more detailed and comprehensive measures of patient health in the English NHS and elsewhere (Kind and Williams 2004; Atkinson 2005; Chauhan and Sussex 2008; McGrail et al. 2012). The term *patient-reported outcome measure* (PROM) has become synonymous with a large[10] number of measurement instruments that assess the health status and health-related quality of life (HRQoL) of patients from their own perspective (Fitzpatrick 2009). Examples include generic instruments such as the EuroQoL-5D (EQ-5D) (Brooks 1996) or the Short Form-36 (SF-36) (Ware

[10] Garratt et al. (2002) identified over different 1,200 instruments during a systematic review conducted in 2002. This number has been estimated to have increased to over 3,000 instruments by 2007 (Fitzpatrick 2009).

and Sherbourne 1992), which can be applied to different health conditions, and disease- or procedure-specific instruments such as the Oxford Hip and Knee Scores (OHS/OKS) (Dawson et al. 1996; Dawson et al. 1998). Many of these instruments are multi-dimensional and assess patients' HRQoL along physical, emotional and social domains. The resulting health profiles are not directly comparable across patients but can be transformed into interval scores through the use of aggregation functions i.e. sets of weights. These weights either express von Neumann-Morgenstern utilities, non-utility preferences or are not preference-based, and they can be obtained from different audiences (e.g. general population, patient groups) and using different elicitation techniques (e.g. standard gamble, time trade-off); see Drummond et al. (2005) and Walker et al. (2011) for more detail. Data from preference-based instruments can be combined with information on the duration of health states to calculate quality-adjusted life years (QALYs). These reflect the cumulative HRQoL as generated by the long-term health trajectory of the patient.

PROMs and QALYs have an established role in clinical research and in the economic evaluation of health technologies. Conversely, their application outside of clinical trials and for outcome assessment in routine care has been limited. However, this is now changing. Since April 2009, PROMs have been collected routinely for all NHS-funded patients undergoing four elective surgical procedures in English hospitals (Department of Health 2008a).[11] These are unilateral hip and knee replacement, groin hernia repair and varicose vein surgery. Each year, over 240,000 patients are invited to report their health before and three or six months after surgery using both generic the EQ-5D and EQ-VAS and disease-specific instruments; the exception being hernia repair for which only the generic instruments are available (see Table 1.1; details of these PROMs are provided in Chapter 2). This before and after measurement allows calculation of changes in health as perceived by the

[11] To my knowledge, the English NHS is the first healthcare system to make collection of PROM data mandatory for hospital providers. Other countries, most notably Sweden, collect PROMs as part of clinical registers and achieve nearly full coverage (Garellick et al. 2009). However, participation is optional for hospitals and these initiatives are not used by the regulator for routine performance measurement.

patient. While this falls short of the ambition to measure changes in health trajectory over the life span, data from the national PROM programme offer important insights into the short-term benefits that patients receive from treatment by different hospital providers. Furthermore, information on pre-treatment health status may help to overcome some of the challenges associated with case-mix adjustment, thereby making comparisons more viable.

Table 1.1: PROM instruments by procedure

Procedure	Condition-specific PROM	Generic PROM	Post-op data collection after
Knee replacement	Oxford Knee Score (OKS)	EQ-5D, EQ-VAS	6 months
Hip replacement	Oxford Hip Score (OHS)	EQ-5D, EQ-VAS	6 months
Varicose vein surgery	Aberdeen Varicose Vein Questionnaire (AVVQ)	EQ-5D, EQ-VAS	3 months
Groin hernia repair	-	EQ-5D, EQ-VAS	3 months

The Health and Social Care Information Centre (HSCIC) derives and publishes performance indicators based on PROM data for all English hospitals performing one of the four NHS-funded surgical procedures. To this end it has developed a case-mix adjustment methodology that takes into account a number of patient characteristics, including patients' pre-treatment PROM response (Coles 2010). Results are either presented as adjusted post-operative scores or, equivalently, as gain scores. Neuburger, Hutchings, Meulen et al. (2013) have shown that performance indicators based on EQ-5D utility scores and condition-specific PROMs are highly correlated, but there was substantial disagreement with respect to the EQ-VAS. Up until the financial year 2012/13, the case-mix adjustment models for joint replacement surgery did not differentiate between primary and revision surgery. Following clinical advice, this approach has since been revised and data for these two patient subgroups are now analysed separately and published as separate performance indicators.

One potential obstacle for the use of PROM data in performance assessment contexts is the potential for reporting bias. Patients may differ both in their underlying latent health and in the way they report and judge their health. As a result, two hospitals providing equally effective services to otherwise identical patients may be judged to perform differently. However, the availability of pre- and post-treatment PROM data may help mitigate the problem if a) rating scales are stable over time so that change is measured consistently, and b) reporting heterogeneity manifests only as scale shifts and not cut-point shifts (Lindeboom and Doorslaer 2004). Also, since PROM scores are averaged across providers' patient populations, individual differences in reporting may also even out. Ultimately, however, a satisfactory way to adjust for reporting heterogeneity would require the collection of anchoring vignettes as part of all PROM data collection (Murray et al. 2003; Rice et al. 2012).

A second issue is the substantial risk of missing data. In contrast to outcome measures like mortality or readmission, which are based on administrative data with nearly complete coverage, PROMs are collected as part of a survey. This opens up the possibility of non-response. Patients are invited to participate but may decline to do so. Also, providers, who are responsible for administering the questionnaire as part of the pre-operative assessment, may fail to collect data or pass them on to the HSCIC. Even where data has been collected it may not be possible to link these to inpatient records. As a result, these data do not contribute to the calculation of performance indicators since important information on case-mix factors is absent.

Gutacker, Street et al. (2015) have estimated that 76% of hip replacement patients treated in the financial year 2011/12 responded to the pre-operative questionnaire and 62% (82% of received questionnaires) could be linked to inpatient records. Non-response is related to observable characteristics of the patient, such as their age, gender and socio-economic background (Hutchings et al. 2014; Gomes et al. 2015). It also relates to the provider of care, with privately operated independent sector treatment centres achieving, on average, higher response rates (Gomes et al. 2015). However, there is only a weak and not statistically significant association between response rates and health outcome (Hutchings et al. 2014; Street et al. 2014). Furthermore, assuming that providers' ability to improve health is indeed uncorrelated with their patients' propensity to take part in the PROM survey, Gomes et al. (2015) have shown that non-response has only a relatively small impact on provider average scores, but a relatively large impact on the statistical uncertainty surrounding these scores. This suggests that non-response bias may lead to more conservative judgements about hospital performance since it reduces the risk of being detected as a positive or negative performer.

1.3 Econometric assessment of performance using multilevel modelling

Throughout this thesis I make extensive use of multilevel modelling techniques, also known as hierarchical modelling, to estimate the performance of healthcare providers. It is therefore useful to describe these techniques here in detail. Comprehensive reviews can be found in Snijders and Bosker (1999), Hox (2002) and Gelman and Hill (2007).

The objective of any performance assessment is to identify the contribution of providers' (unobservable)[12] actions to their patients' observed outcomes (Jacobs et al. 2006). To achieve this, the provider effect must be isolated from other determinants of outcomes, most notably patient *case-mix* and *random noise*; both of which are assumed to be outside the providers' control. Formally, let

$$Y = f(X, \theta, \epsilon) \tag{1.1}$$

where Y denotes the patients' outcome of interest (e.g. post-operative PROM), X denotes factors outside of the providers' control (e.g. case-mix), θ denotes the providers' actions, and ϵ denotes random variation.

Define performance as the *systematic* effect that providers have on all their patients' outcomes. Hence, θ varies across providers but not across patients within the same provider.[13] Performance variation thus implies that patients are clustered within

[12] Observable actions are rarely of concern. If the principal could observe the agent's actions, she could also contract them.

[13] Only systematic variation in outcomes can be identified. If a provider's efforts would vary randomly across its patient population, this variation in performance could not be distinguished from random chance variation. It would in principle be possible to allow providers' actions to have a differential

providers: two otherwise identical patients treated by the same provider experience more similar outcomes than the same two patients treated by different providers. The degree of clustering can be measured by the intraclass correlation coefficient (ICC) defined as

$$ICC = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{1.2}$$

where τ^2 denotes the variance in outcomes Y between providers and σ^2 denotes the variance within providers. A non-zero estimate of ICC indicates performance variation.

Provider performance cannot be directly observed. The performance assessment literature has taken two different approaches to estimating θ, both of which are based on the notion that, after adjusting for all other relevant factors, the remaining variation in outcomes can be reasonably assumed to derive from providers' actions. In the first approach the observed outcomes for a provider are compared against those predicted from a regression model conditioning on X. The ratio of [difference between] average observed and predicted outcomes gives an indication of the performance level with 1 [0] indicating expected performance and larger values indicating better than expected performance. This form of *indirect standardisation* has the advantages of being easily conducted using standard regression techniques and of allowing performance estimates to be expressed in the natural unit of the outcome (e.g. by multiplying the observed-over-expected ratio with the average outcome across all providers). However, it does not exploit the clustering of patients in providers and thus does not make use of all available information (Austin et al. 2003; Ash et al. 2012; Ding et al. 2013).

The second approach exploits the hierarchical nature of the data to estimate θ directly. These models are known to epidemiologists and statisticians as *multilevel* or *hierarchical* models, and are referred to as *panel data* models in econometrics. For

effect according to patients' observable characteristics using e.g. random coefficient models. I focus on the simpler case here for ease of exposition.

example, consider the following linear model

$$Y_{ij} = \alpha + X_{ij}^{'}\beta + \theta_j + \epsilon_{ij} \tag{1.3}$$

with patient $i = 1, \ldots, n_j$ treated in hospital $j = 1, \ldots, J$. The coefficient α denotes an overall intercept, whereas the coefficient θ_j captures provider-specific intercept shifts.[14] The random error term ϵ_{ij} is assumed to be distributed as $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

The econometric literature emphasises two ways in which the provider-specific intercepts can be modelled (Hsiao 1986; Wooldridge 2002). Fixed effect (FE) models treat θ_j as parameters to be estimated from the data. This is typically achieved by including a dummy variable for each provider. The associated coefficient θ_j captures the difference between the average level of Y_{ij} for provider j and the overall intercept α, conditional on other modelled covariates. Jones and Spiegelhalter (2009) point out that the FE model implicitly assumes an identically distributed random variable θ_j with infinite variance. Random effects (RE) models make the additional assumptions that all θ_j are identically distributed random variables with finite variance and are uncorrelated with the covariates. Provider effects are typically assumed to be distributed as $\theta_j \sim \mathcal{N}(0, \tau^2)$ although other distributions (e.g. T-distribution) would be possible. Crucially, only estimates of τ^2, defining the distribution of provider effects, are obtained from e.g. maximum likelihood estimation of (1.3) and the provider-specific parameters θ_j need to be *predicted* from this distribution (Searle et al. 1992).

A natural way to recover provider effects in a random effects framework is through Bayes' Theorem (Efron and Morris 1973; Skrondal and Rabe-Hesketh 2009): The posterior distribution of a parameter is obtained by combining prior beliefs about its distribution with the data (i.e. likelihood), or

$$\omega(\theta_j \mid Y_{ij}, X_{ij}; \tau, \sigma, \beta) = \frac{\varphi(\theta_j; \tau) f(Y_{ij} \mid \theta_j, X_{ij}; \sigma, \beta)}{g(Y_{ij} \mid X_{ij}; \tau, \sigma, \beta)} \tag{1.4}$$

[14] The model can be extended to allow for other coefficients to vary by provider. These types of models are known as *random coefficient* models, whereas the model above is typically called a *random intercept* model. Both are conceptually identical.

where $\omega(.)$ is the prior, $f(.)$ denotes the conditional density and $g(.)$ denotes the likelihood contribution of cluster j. In most empirical applications, the prior is taken to be $N(0, \hat{\tau}^2)$. This practice is known as *Empirical Bayes* (EB) prediction (Skrondal and Rabe-Hesketh 2009): In contrast to a fully Bayesian approach, the prior is not independent of the data; hence '*empirical*'.

Bayes' Theorem implies that the EB predictions of the provider effect θ_j are shrunken towards the mean of the prior distribution. This is illustrated in Figure 1.1. The amount of shrinkage is determined by the strength of information in the data and the degree of homogeneity in provider performance. When information is sparse, i.e. the number of patients n_j within a provider j is low, the posterior means resemble the mean of the prior more closely. Conversely, for units containing much information (i.e. large n_j), the results are primarily driven by the data and shrinkage is minimal. Shrinkage can therefore be seen as a form of '*borrowing strength*'. Since, by assumption, hospitals share some commonality in their production process, one can reasonably utilise information on all providers to inform estimates about individual providers. The more homogeneous providers are (i.e. smaller ICC), the larger the potential to borrow strength. Fixed effects estimation does not allow for such shrinkage since it ignores this commonality. This is best seen in the case of a linear random-intercept model, for which the expectation of (1.4) can be evaluated analytically. The EB predictor is

$$\hat{\theta}_j^{EB} = \hat{R}_j \left[\frac{1}{n_j} \sum_{i=1}^{n_j} (Y_{ij} - \hat{\alpha} - X_{ij}^{'} \hat{\beta}) \right] \tag{1.5}$$

with

$$\hat{R}_j = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2 / n_j} \tag{1.6}$$

and $0 < \hat{R}_j < 1$. Conversely,

$$\hat{\theta}_j^{FE} = \frac{1}{n_j} \sum_{i=1}^{n_j} (Y_{ij} - \hat{\alpha} - X_{ij}^{'} \hat{\beta}) \tag{1.7}$$

since, by assumption, $\tau = \infty$ and therefore $\hat{R}_j = 1$.

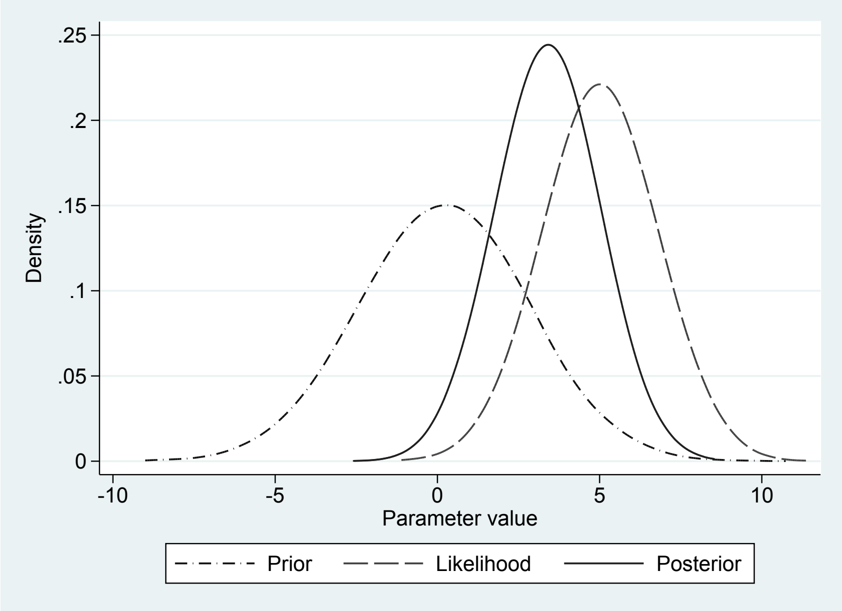


Figure 1.1: Combining prior and likelihood to compute posterior distribution

The advantages of EB estimation and shrunken provider effects have long been recognised in the literature on school effectiveness (Aitkin and Longford 1986; Goldstein 1997) and more recently in the performance assessment of healthcare providers (Bojke et al. 2011). Shrinkage is a form of precision-weighting and is therefore a valuable mechanism to account for uncertainty in estimates for hospitals treating a small number of patients. Indeed, shrunken estimates are shown to have lower mean squared prediction error than non-shrunken estimates obtained from FE estimation and are best linear unbiased predictors in linear models with random effects (Efron and Morris 1973). Shrinkage may also be desirable because it allows making inferences about *all* providers and does not require analysts to set arbitrary inclusion cut-offs with regard to cluster size. However, shrinkage also implies a bias towards zero. Performance estimates based on EB prediction are therefore likely to be conservative and have higher specificity but lower sensitivity than FE estimates (Austin et al. 2003; Kalbfleisch and Wolfe 2013).

The random effects approach and associated EB prediction techniques are commonly observed in the statistical and epidemiological literature. Conversely, when confronted with clustered data, economists tend to favour the FE approach. I believe this is largely due to the different foci of the analyses: Interest in many economic applications is confined to the unbiased estimation of β and unobserved heterogeneity through clustering is seen as a nuisance rather than of interest in itself.[15] These roles are reversed in the context of performance assessment: covariates used for case-mix adjustment and associated parameters are not of substantial interest, whereas provider effects are. An analyst therefore has to trade off improved estimation of provider effects against potential bias that enters through coefficients used for case-mix adjustment. Where this bias is substantial, more complex modelling strategies, such as those proposed by Mundlak (1978) and Chamberlain (1982), may be employed to obtain an unbiased shrinkage estimator of θ_j.

1.4 Structure of this thesis

In this thesis I make extensive use of pre- and post-operative PROM data from the national PROM programme to study issues of performance variation in quality and costs across English hospitals. The information gained this way can help reduce information assymetries and can be used by various principals to select, incentivise and hold to account the providers of healthcare with which they contract. The thesis explores three broad themes: the relationship between variation in resource use and quality, the appropriate assessment and reporting of multidimensional hospital performance, and the use of performance information to motivate hospitals in a public reporting context.

In Chapter 2 we explore whether observed outcome quality, as measured by average patient-reported health gains, can explain some of the variation in treatment costs reported by English NHS hospitals. Variations in costs across providers of

[15] RE estimation may lead to biased estimates of β and, by extension, θ_j if $cov(X_{ij}, \theta_j) \neq 0$, i.e. the covariates correlate with the provider effect. Following Mundlak (1978) this bias arises because the relationship between X_{ij} and Y_{ij} differs from that of $\bar{X}_j$ and Y_{ij} and $\bar{X}_j$ varies across j. The usual RE estimator uses between and within cluster variation in X_{ij} to estimate β, which therefore neither reflects the within nor between relationship appropriately. Also, if $\bar{X}_j$ varies across j but this is not modelled then the assumption of an identically distributed random intercept no longer holds. The FE estimator circumvents this problem by subsuming all between-provider variation into the provider effect so that coefficients are only estimated from within variation.

the same care are a common finding in many healthcare systems (e.g. Busse et al. 2008; Laudicella et al. 2010) and a source of concern for price-setting regulators. Prospective reimbursement systems based on yardstick competition, such as the English Payment by Results (PbR), should give strong incentives for hospitals to reduce resource utilisation and compete for elective patients on the basis of quality (Rogerson 1994; Ma 1994). If large variations in costs exist and persist over time, this indicates that the incentives created by the reimbursement system fail to change provider behaviour. Also, in many publicly funded health systems hospitals face soft budgets and are protected from the threat of market exit through public guardianship. Therefore, any costs that exceed the reimbursement fall ultimately onto the health budget and displace other valuable healthcare.

When challenged about their costs, hospitals may argue that they i) treat an unfavourable case-mix, ii) operate within a difficult production environment, and/or iii) invest in higher quality of care. We use data on hospitals' reference cost returns[16] and PROM change scores for the financial year 2009/10 to address two questions: First, are larger improvements in patient health associated with higher average costs, i.e. do hospitals have grounds to claim that their high costs are due to superior quality? The economic literature on this question is inconclusive; see Hussey et al. (2013) for a review of the US literature. Second, how much of the observed variation in hospitals' average costs can be attributed to quality as measured by PROM change scores? All our analyses are conducted within a multilevel modelling framework to account for case-mix differences, including the average pre-operative health status of the hospital's patient population, and observable production constraints (e.g. scale and scope of operation). We find some empirical support for a U-shaped association between outcome quality and costs, i.e. costs initially fall as quality increases but then start to rise again, but this finding is sensitive to the condition under study and the choice of PROM measure. There is no evidence that costs increase monotonically

[16] All NHS hospitals are required to provide information on their cost structure to the Department of Health for the calculation of current spending and future reimbursement schedules. These data are derived following the same accounting standards.

in quality. We show that an adjustment for hospitals' relative ability to improve their patients' health has only a modest effect on observed hospital cost variation.

In Chapter 3 we discuss appropriate means to analyse and present variation in hospitals' relative effect on patients' multidimensional HRQoL. We focus on performance information generated from EQ-5D data, where these information are made publicly available for the purpose of informing patients' choice of hospital. The EQ-5D measures patients' HRQoL along five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. On each dimension, patients can indicate whether they have no, some or extreme problems. Together, these responses form the patient's EQ-5D health profile.

Current practice in the national PROM programme is to aggregate health profiles into interval scores using preference estimates obtained from the English general public (NHS Information Centre 2010a; Dolan 1997). We argue that the use of public preferences is inappropriate - and potentially misleading - if the resulting performance estimates are intended to be used by patients choosing a hospital for treatment. Instead aggregate performance estimates should take into account each patient's individual preferences since they are the relevant decision makers in this setting and their preferences may differ substantially from those of the general public (e.g. Mann et al. 2009).[17] Because eliciting each patient's individual preferences and providing individualised performance reports is infeasible in practice, we suggest generating performance estimates for each of the five EQ-5D dimensions separately and reporting the expected probability of a patient to report, for example, no problems post-operatively. This approach to analysis and presentation of EQ-5D provider performance estimates is consistent with economic welfare theory since it allows patients to exercise choices based on performance information that is consistent with their own preferences. It may also help overcome some of the problems in interpreting the data by patients (Hildon et al. 2012) and help hospitals

[17] In contrast, there is a strong argument for using the general public's preferences when making decisions about the adoption of new medical technologies into a tax-funded healthcare system such as the English NHS. See Brazier et al. (2005) for a discussion.

understand where their performance falls short or where they excel (Smith 2015).

We illustrate our approach using data on all patients undergoing hip replacement surgery in the financial year 2009/10. We find that performance heterogeneity is most pronounced on those dimensions that receive a low weighting in the UK time trade-off EQ-5D tariff, i.e. the mobility and usual activities dimensions. Conversely, performance estimates based on aggregate scores correlate well with estimates for the anxiety/depression and pain/discomfort dimensions, which receive a high weighting. Hence, the currently reported performance estimates based on aggregate scores are driven by the preferences of the general public and may hide performance variation that patients may value.

Chapter 4 builds on the theme of the previous chapter by suggesting a general methodology to assess provider performance across multiple dimensions without imposing strong normative judgements about the preferences of the relevant principals. We propose the use of dominance criteria to identify providers that perform well or poorly under only weak assumptions about the principals' utility functions.[18] To this end, we apply multivariate multilevel regression models with correction for patients selecting into hospital to isolate performance variation from observed and unobserved case-mix differences (e.g. demographics and pre-operative health) and random noise (Hauck and Street 2006; Terza et al. 2008). We also propose a methodology to construct appropriate uncertainty statements around dominance classifications.

We apply this methodology to NHS hospitals and independent sector treatment centres (ISTCs) providing hip replacement surgery during the financial years 2009/10 to 2011/12. Provider performance is assessed in terms of patients' health gain, probability of emergency readmission, waiting time and length of stay. A number of interesting findings emerge from this analysis. First, all hospitals that dominate the benchmark are ISTCs and all hospitals that fall short of the benchmark

[18] Our approach is similar to the Pareto Classification System proposed by Parkin et al. (2010). However, whereas they compare changes in health profiles for individual patients over time, we compare cross-sectional performance estimates across multiple dimensions of performance.

are NHS hospitals. This may be due to a more streamlined production process, since ISTCs do not provide emergency care and focus on a small number of surgical procedures. Another important finding is a negative association between length of stay and health gain. This is consistent with some of the health service research literature on enhanced recovery pathways that show that care processes can be optimised to achieve both higher quality and lower resource use (Husted et al. 2008; Larsen et al. 2008). It also corroborates the finding of a lack of monotonically increasing cost-quality relationship reported in Chapter 2. Finally, patient selection into hospital has a negligible effect on outcomes, waiting time and length of stay once patients' pre-operative health is taken into account. Previous studies of the US market have shown that patient selection into hospital based on unobservable characteristics can severely bias performance estimates (Gowrisankaran and Town 1999; Geweke et al. 2003). However, these studies were limited in the availability of good pre-treatment health measures.

The contribution of Chapter 5 is to test whether demand for hip replacement surgery at a hospital is a function of published performance information on its ability to improve its patients' health. As noted before, the current prospective reimbursement system in the English NHS gives hospitals an incentive to compete for patients on the basis of quality. However, this requires patients to be sensitive to publicly available information on hospital quality and hospitals to operate in sufficiently competitive markets. The empirical work in this chapter contributes to the debate about the effectiveness of using public reporting of performance data as a means to incentivise healthcare providers (Cutler et al. 2004; Hibbard et al. 2003; Hibbard et al. 2005; Ketelaar et al. 2011).

We estimate hospital demand models using data on the observed choices of over 180,000 patients undergoing primary hip replacement surgery in the English NHS in the financial years 2010/11 to 2012/13. We find that patients respond to the published PROM quality measures and that hospitals can increase their demand by approximately 9% if they find ways to improve their patients' health outcomes by one standard deviation. However, patients' choices are driven primarily by the distance to hospital. This means that a hospital's ability to attract patients away from competitors through increased quality (i.e. the cross-elasticity of demand) reduces rapidly as the distance between hospitals increases. Overall this suggests that publishing PROM quality metrics may be an effective instrument to incentivise hospitals to provide high quality care, but its effectiveness depends on the local circumstances.

The last chapter reviews and discusses the salient points of the previous chapters and provides policy recommendations and suggestions for future research.

2 Truly inefficient or providing better quality of care? Analysing the relationship between risk-adjusted hospital costs and patients' health outcomes

2.1 Introduction

Any health system that aims to make the best use of its scarce resources will be concerned about variation in costs between different providers of the same health-care. If providers can reduce costs to the level of best practice, resources might be released to provide benefits elsewhere. But in analysing variations in provision, it is important to ensure that an assessment of best practice includes not only costs but also patient outcomes. High costs are not always simply due to inefficiency but may be associated with better outcomes. Low costs may sometimes be a symptom of low quality care leading to poor outcomes.

A better understanding of the complex relationship between costs and quality is required to address the important policy question of 'which variation in cost is justifiable' (Keeler 1990). The health economic literature contains several studies that explore empirically the effect of better health outcomes on costs. However their findings remain inconclusive. While some studies report costs to be positively related to health outcomes (Morey et al. 1992; Mukamel et al. 2001; Schreyögg and Stargardt 2010), others suggest that cost reductions and quality improvements may be achieved simultaneously (Fleming 1991; Carey and Burgess 1999; Deily and McKay 2006; Clement et al. 2008; McKay and Deily 2008). In an attempt to accommodate both sets of empirical results within a unified framework, some authors have reviewed the idea of a U-shaped cost-quality curve and found supporting evidence (e.g. Weech-Maldonado et al. 2006; Hvenegaard et al. 2011). This framework explicitly acknowledges that efforts to improve quality will sometimes contribute to lower resource use and better health outcomes, whereas in other cases additional resources are required to achieve better results.

An important limitation of the existing literature is its focus on negative health outcomes resulting from inadequate quality. With few exceptions (e.g. Picone et al. 2003), health outcomes are measured as rates of mortality, re-admission or adverse events.[19] While important, these 'failure' measures cannot reflect the full spectrum of patient health and are frequently deemed too noisy and insensitive to be useful for provider comparison (Thomas 1996; Lilford and Pronovost 2010).

Reliance on rates of failure as primary measures of health outcome stems from the lack of comprehensive data of patients' health. This is being addressed. Since April 2009, all providers of publicly-funded care in the English National Health Service (NHS) are required to collect patient-reported outcome measures (PROMs) for four procedures: unilateral hip and knee replacements, varicose vein surgery, and groin hernia repairs (Department of Health 2008a). Standardised questionnaires, including both generic (the EQ-5D) and condition-specific instruments, are collected from all eligible inpatients before and 3 or 6 months after surgery. The changes in patients' health status can be analysed to measure the hospitals' systematic contribution to health with finer granularity than previously possible (Appleby and Devlin 2004).

Building on this initiative, the work presented in this chapter has two aims. First, we explore to what extent variation in risk-adjusted costs is associated with variation in patient-reported health outcomes. Second, we investigate whether the new information on health outcomes changes our judgement of relative provider cost performance. We perform sensitivity analysis to assess the degree to which our findings depend on the choice of PROM instrument and to the way in which the

[19] Instead of observed measures of quality, Romley and Goldman (2011) use observed choices of hospitals to measure variation in unobserved, patient-perceived hospital quality and relate this to costs. They find a positive relationship but note that revealed quality is not strongly correlated with observed clinical quality.

cost-outcome relationship is modelled.

Our empirical approach is to estimate multilevel models that recognise the clustering of patients within providers. We treat patients as repeated observations of the hospital's production process. This allows us to distinguish random noise from systematic cost variation attributable to cost containment effort (e.g. Dormont and Milcent 2004; Olsen and Street 2008; Laudicella et al. 2010), without having to specify the production possibility frontier; a task that has been criticised in the past for its distributional assumptions and sensitivity to modelling choices (Newhouse 1994; Skinner 1994).

We estimate separate models for each of the four procedures, each of which is considered as akin to a production line. This modelling approach has two important advantages over consideration of hospital production in its entirety. First, hospitals are multi-product organisations, consisting of multiple units such as individual medical teams, departments and specialities, and the management (Harris 1977; Pauly 1980). Efforts to contain costs exerted in one part of the organisation are unlikely to affect health outcomes in other parts. Consequently, any attempt to disentangle the complex relationship between costs and outcomes using data from multiple units may lead to the identification of spurious relationships. By focusing on single production lines, we can relate variation in health outcomes more directly to variation in relevant resource use and ensure more thorough risk-adjustment (Bradford et al. 2001). Second, we can assume a common underlying production function that is shared by all providers of the procedure in question. This ensures 'like-for-like' comparisons across providers.

2.2 Conceptual Framework

Social systems are often sufficiently complex to require a less-informed principal to delegate a task to a specialised agent in return for some reward.[20] The principal's

[20] Such agency relationships exist not only between institutions (e.g. regulators and hospitals) but within institutions (e.g. management and medical staff) (Harris 1977). A better understanding of variation in effort across and within healthcare institutions is therefore crucial for policy makers

objective is to ensure that publicly-funded services are of adequate quality and delivered in a technically efficient manner. The potential agency problems arising in such situations are well known (Laffont and Tirole 1993) and occur when principal and agent have different objectives or value them differently and the agent's effort is unobserved. These information asymmetries allow agents to misreport effort and pursue their own objectives.

One way of mitigating the problem of misreporting is to improve the information base by undertaking comparative cost analysis. The problem is that when agents are heterogeneous with respect to their products and production processes, simple comparison does not suffice. In these instances, Shleifer (1985, p.324) proposes multiple regression of costs on legitimate *"characteristics that make firms differ, and correct[...] for this heterogeneity"*. The natural framework for this analysis is the industry cost function that underlies all agents' production processes. Following Bradford et al. (2001), we specify this cost function at the level of the individual patients. This formulation recognises that hospital care is tailored to the specific characteristics and requirements of each individual. The agent's cost function is then

$$C = C(X, q, r, w, Z, e) \tag{2.1}$$

where C is the unit cost for the specific episode of care, X is a vector of variables representing medical need, initial health and other case-mix factors of the patient, q is a measure of quality of care provided, r and w are price vectors for capital and labour, Z is a vector of environmental factors that constrain the production process and e is the level of effort exerted by the agent.

A major source of variation in production costs is heterogeneity with respect to patient case-mix. Even within production lines, some patients will require more attention and resources than others because they suffer from more severe conditions, present with initially lower health or differ with respect to other factors that determine treatment costs, e.g. age, gender or number and type of comorbidities.

and local managers alike.

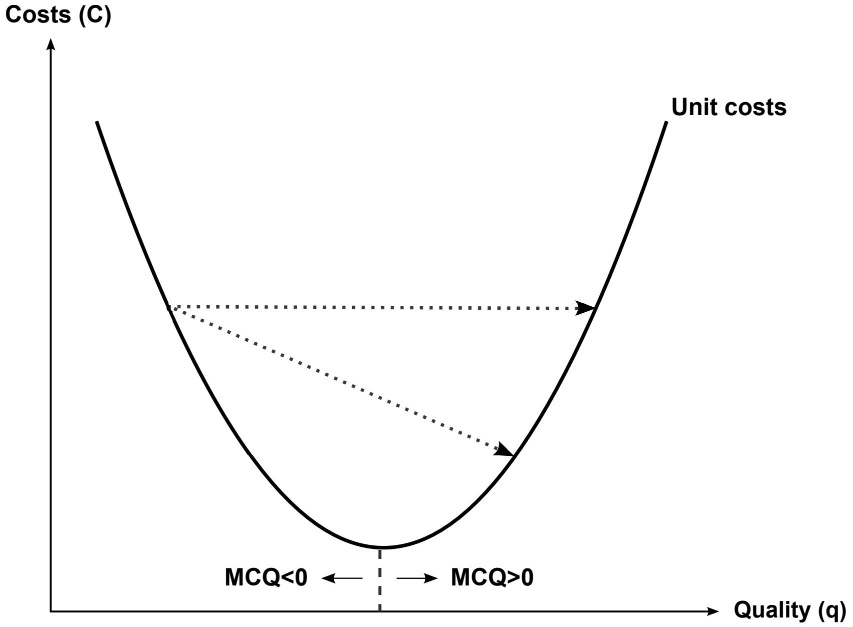
Unless patients are randomly allocated to hospitals, some providers may attract a more favourable case-mix than others and achieve similar costs while exerting less effort. It is therefore crucial to correct for patient heterogeneity in order to allow for fair comparison.

A second reason why production costs may differ across hospitals is because some providers face a more adverse production environment than others. For example, hospitals may differ in their access to factor markets and pay different prices for capital and labour inputs. Costs may also be determined by location or the existing infrastructure, both of which are, at least in the short run, not within the provider's control.

Finally, production costs may differ because of variations in the quality of care provided. Patients seek healthcare to improve their health and health-related quality of life or avert imminent deteriorations (Jacobs et al. 2006). Hospitals can, at least partially, control the outcome of care through their production decisions, for example, by investing in more effective medical technology or employing more experienced surgeons. Any principal seeking to maximise patients' health within a constrained budget will therefore request increasingly higher levels of quality as long as production remains cost-effective, i.e. the additional costs of better quality do not outweigh its benefits. As a consequence, differences in observed costs brought about by variation in quality should be taken into account when comparing hospital costs.

In order to determine the quality-adjusted level of cost performance, the principal must establish the production costs of quality. The marginal costs of quality (MCQ) are not necessarily constant over the observed range of quality. Indeed, if there are diminishing marginal returns to factor input hospitals that provide high quality care may find it more expensive to achieve further improvements than their low quality peers. Moreover, MCQ may not be positive for all levels of observed quality. The literature describes several organisational and medical interventions that lead to better health outcomes while reducing costs, for example by mobilising patients sooner after joint operations and discharging them earlier (Siggeirsdottir et al. 2005; Larsen et al. 2008), or preventing costly adverse events (Carey and Stefos

2011). When some hospitals have not (yet) implemented such cost-saving quality improvements[21], they may face negative MCQ, whereas other providers cannot improve quality without further resource use (Hvenegaard et al. 2011). This idea is depicted in Figure 2.1.



Notes: Upper dotted line illustrates improvement in quality at same costs. Lower dotted line illustrates improvement in quality at reduced costs.

Figure 2.1: The cost-quality relationship with non-constant marginal costs of quality (MCQ)

In summary, therefore, the possibility of negative and non-constant marginal costs of quality requires a careful approach to modelling and interpreting provider cost performance. The principal's judgement will differ according to where providers are deemed to lie on the cost-quality curve. If marginal costs are positive, then better quality justifies higher costs and cost performance estimates will need to be adjusted for quality. If marginal costs are negative at this point of the curve, the provider can reach higher levels of quality at equal or lower costs. This is depicted in Figure 2.1, where the dotted lines indicate such movements. In this situation, the principal should not consider quality information for these particular providers in

[21] Even when providers have utility functions that increase in quality and decrease in cost containment effort, one may still observe such a relationship because of implementation costs, or imperfect knowledge of best practice or of their own cost structure.

the benchmark because cost containment and quality efforts are complements. Any adjustment for quality would otherwise result in overstated cost performance that cannot be justified on economic grounds.

2.3 Methods

2.3.1 Statistical analysis

We estimate multilevel models with provider-specific intercepts, separately for each of the four procedures (Rice and Jones 1997; Snijders and Bosker 1999). Patients form the micro observations and hospitals constitute macro units. This approach acknowledges that some factors vary only between macro units (for example production constraints or cost effort), whereas others vary by micro unit.

We specify our empirical model as follows:

$$C_{ij} = \alpha_0 + X_{ij}^{'}\beta + Z_j^{'}\delta + M_j^{'}\vartheta + H_j^{'}\gamma + \theta_j + \epsilon_{ij} \tag{2.2}$$

where C_{ij} is the cost of care[22] for patient $i = 1, \ldots, n_j$ in hospital $j = 1, \ldots, J$. The vector X_{ij} contains case-mix controls that vary at micro level and Z_j is a vector of production constraints at macro level. Because we do not observe individual patients' PROM responses (see data section), we include the average initial health status of the provider's patients M_j to control for observed differences in average medical need. Similarly, the average change in health enters as a macro level covariate and is denoted as H_j. The coefficient α_0 gives the expected cost of a patient when all variables are set to zero.[23]

Unexplained variation is decomposed into two components: i) a random error

[22] We use the natural unit of costs instead of the logarithmic transformation or more flexible generalised linear models (GLMs). Results for models without provider random effects are similar to those obtained from GLMs with log link and gamma / poisson distribution (see Appendix Table A2.1). This reflects previous findings that linear models with identity link perform well in large samples (Deb and Burgess 2003; Montez-Rath et al. 2006; Daidone and Street 2011). Furthermore, GLMs with provider random effects are difficult to estimate in large samples due to the need to integrate over the random effects distribution.

[23] One can recover the expected costs for a specific patient with $X_{ij} = \tilde{X}$ treated in an average hospital as $\alpha_0 + \tilde{X}^{'}\widehat{\beta} + \bar{Z}^{'}\widehat{\delta}$.

term ϵ_{ij} that varies at micro level and is assumed to be distributed as $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and ii) a provider effect θ_j that captures unobserved heterogeneity at macro level. The latter is interpreted as variation in cost containment effort between hospitals. These provider effects can be interpreted directly, representing the amount of cost deviation from the risk-adjusted benchmark, α_0. Accordingly, if $\theta_j < 0$ the provider has lower average costs than would be expected given the characteristics of its patients and the constraints it faces, and vice versa.

In order to assess the sensitivity of estimates of provider effects to the addition of outcome information, we estimate an alternative model where the effect of health outcome on costs, θ, is restricted to be zero. We then calculate the difference between estimates of $\widehat{\theta_j}$ obtained from the 'full' and 'restricted' models to identify providers for which a naïve benchmark without quality controls provides misleading assessments of cost performance.

2.3.2 Modelling the provider effect

The econometric literature posits two classes of models that can be applied in the case of unobserved provider heterogeneity: fixed (FE) and random effects (RE) (Wooldridge 2002). We choose a RE approach, where provider effects are assumed to be distributed as $\theta_j \sim \mathcal{N}(0, \tau^2)$ and uncorrelated with the micro level covariates. We justify this decision on the basis of three observations: First, in our specific application, both FE and RE estimators yield estimates of β that are virtually identical. We conclude that any bias arising from a potential correlation between the provider effects θ_j and the vector X_{ij} is a trivial concern.[24] Second, the FE estimator does not permit the inclusion of macro level covariates because they would be perfectly collinear with the provider effects.[25] This well-known limitation

[24] We have conducted Hausman tests to verify the assumption of exogeneity. The null hypothesis of unbiasedness has been rejected for hip and knee replacement and groin hernia repair. However, the coefficients differ in the magnitude of 1-2 GBP; a difference that is statistically but not economically significant (see Appendix Tables A2.2 - A2.5). We believe that statistical significance is an artefact of the large sample size at micro level.

[25] While it is possible to circumvent this problem by using Estimated Dependent Variable (EDV) models (e.g. Lewis and Linzer 2005; Laudicella et al. 2010), this comes at the expense of additional complexity and inefficiency, and requires the analyst to correct the resulting standard errors (Beck

is problematic for our study because one of our key variables, health outcome, is only observed at provider level. Third, the RE estimator is more efficient than the FE estimator because it exploits both within- and between-hospital variation.

All models are estimated via maximum likelihood using the command `xtmixed` in Stata 12.

2.4 Data

2.4.1 Hospital Episode Statistics

Our study uses patient level data extracted from the Hospital Episode Statistics database (HES) for the period April 2009 to March 2010. This database allows us to analyse the characteristics and care received by each NHS-funded patient from admission to discharge (Lakhani et al. 2005). All patients are allocated to a Healthcare Resource Group (HRG v.4), the English equivalent of Diagnosis-Related Groups (DRGs).[26] We construct indicator variables for the ten most common HRGs for each procedure, with the most common HRG set as the omitted base category in the regressions.

The construction of any classification system necessarily requires a trade-off between parsimony and homogeneity of the resulting groups. As a consequence, HRGs are unlikely to capture all variation across providers. Hence, we include a set of variables that are based on diagnostic codes (ICD-10) and procedure codes (OPCS-4.5). These include the main reason and type of surgery[27], whether it was a primary or revision surgery, and the weighted Charlson index as a measure of co-morbidity (Charlson et al. 1987). Further, following Laudicella et al. (2010) we

2005).

[26] Patients may be assigned to more than one HRG during their hospital stay if they are transferred between departments. We focus on the HRG of the episode of care in which the PROM procedure took place.

[27] We follow the classification of procedures as set out in the policy guidance document (Department of Health 2008a) and distinguish, where appropriate, between primary and revision surgery. For hip and knee replacement we differentiate between a main diagnosis of osteoarthritis (ICD-10: M15-19), rheumatoid arthritis (M05-M06) or other. For varicose vein surgery we differentiate between varicose ulcer (I83.0), varicose vein with inflammation (I83.1), without inflammation (I83.9), with inflammation and ulcer (I83.2) or other.

generate counts of non-duplicate, secondary diagnoses and procedure codes within a spell as further controls for co-morbidities and complications.

We account for patient demographics by sorting patients into age quintiles and create an indicator variable for male gender. We attribute to each patient the proportion of residents in the patients' neighbourhood[28] that claim means-tested benefits, which we interpret as a measure of income deprivation. This information is obtained from the Index of Multiple Deprivation 2004 (Noble et al. 2006). To characterise the inpatient stay itself, we construct variables for transfers in and out of hospital, transfers between departments, whether the patient is discharged home or not, and in-hospital mortality.[29]

We construct variables that reflect (short-term) production constraints. Larger providers may be able to realise economies of scale and we generate a measure of size based on the count of patients treated by the provider for each of the four procedures. To address economies of scope, we create an index of specialisation that reflects the dispersion of HRGs treated within the hospital (Daidone and D'Amico 2009). The index resembles a Gini index and is bounded between zero (no specialisation) and one (all patients of hospital j fall into one HRG). Finally, hospitals are categorised into teaching and non-teaching facilities based on the classification system adopted by the National Patient Safety Agency (2011). Since these constraints may not be binding, we explore the sensitivity of our results to the inclusion of these variables into our regression models.

2.4.2 Reference cost

Hospital Episode Statistics do not include information on the cost of care. However, NHS hospitals are required to provide information on their cost structure to the Department of Health for the annual compilation of the reference cost schedule and

[28] HES records patients' locations in terms of the Lower Super Output area (LSOA; 2001 census boundaries) in which they reside. Each LSOA contains approximately 1,500 inhabitants and is designed to be homogeneous with respect to tenure and accommodation type.

[29] In-hospital mortality is less than 0.5% for all conditions studied. We therefore consider the risk of survivorship bias to be negligible.

calculation of reimbursement prices (Department of Health 2010). Reference cost data have been collected since 1997. They are made available both at aggregate level (overall spending) and disaggregated, i.e for each individual provider. We utilise individual hospitals' 2009/10 returns to construct patient-level cost data (Department of Health 2011b). No disaggregate data are available for private providers of NHS-funded care, namely independent sector treatment centres. We therefore focus our analysis on NHS hospitals.

The reference cost report is implemented using a top-down costing methodology (Department of Health 2010). Costs are attributed to individual patients where possible. For those cost components where this is not feasible (e.g. overheads), total hospital costs are progressively cascaded down through a hierarchy of costing levels, starting at treatment services, to specialities and finally to individual HRGs. Costs at HRG-level are reported separately for departments and are further disaggregated according to admission type (day case, elective and emergency care) and length of stay, where HRG-specific trim points are used to differentiate long inpatient spells.[30] Hence, within each department and HRG there can be up to five (since day cases cannot be further differentiated by length of stay) different cost estimate groups (*'cost baskets'*). For two of those, namely those where length of stay exceeds the trim point, costs also vary within providers. This allocation of resources is intended to reflect variation in costs between patients and is governed by strict accounting rules that all NHS providers have to adhere to.

Trusts report the average costs for each cost basket and the cost per excess bed day above this trim point. We map these reference cost data to admission records as described in Table 2.1 (see also Laudicella et al. (2010)). To mitigate the effects of measurement and coding errors, we drop observations for which reported costs are below 1% or above 99% of the observed costs within the specific cost basket across all providers.[31] We further exclude observations for one provider for which

[30] Trimpoints are set to the 75th percentile + 1.5*(75th - 25th percentile) of the distribution of length of stay in this HRG in the previous financial year. They are calculated separately by admission type.

[31] Alternatively, one could top- and bottom-code these observations, i.e. replace the values of all observations $< 1\%$ [$> 99\%$] of the cost distribution with the value observed at the 1st [99th]

the reference cost data are considered to be of insufficient quality.[32]

Table 2.1: Reference cost data allocation

Admission type	Normal stay $(los - tp \leq 0)$	Long stay $(los - tp > 0)$
Elective	C^{El}	$C^{El} + (los - tp^{El})ebd^{El}$
Emergency	C^{Em}	$C^{Em} + (los - tp^{Em})ebd^{Em}$
Day-case	C^{Dc}	C^{Dc}

El = Elective; Em = Emergency; Dc = Day-case; los = Length of stay; tp = Trimpoint; ebd = excess bed-day cost
Notes: C denotes the average (fixed) cost of care in this cost basket, whereas ebd denotes variable costs. Hence, costs only vary between two patients in the same hospital department and HRG if their length of stay differs and at least one patient's stay exceeds the trimpoint.

We adjust patient costs by the Market Forces Factor (MFF) specific to the provider. The MFF is an index of relative prices for buildings, land and labour that is used by the English Department of Health to account for unavoidable variation in input prices (Department of Health 2008b).

2.4.3 Patient-reported outcomes

Data from the PROMs programme cover April 2009 - March 2010 and are published at hospital-level by the NHS Information Centre (IC)[33] for all providers of NHS-funded care (NHS Information Centre 2010b). The data are obtained by surveying patients before and after their operation. For each hospital, data are available about the average health status pre-surgery, post-surgery, and the average change in health after treatment.[34] Individual-level data, on which these average scores are based, were not available to us at the time of study.

The PROMs survey includes both generic and condition-specific instruments for which data are reported separately (see Chapter 1). The EQ-5D is a generic PROM comprising a set of questions asking patients to indicate whether they have no,

percentile.

[32] We exclude data for Southend University Hospital NHS Foundation Trust, which reports unrealistically low average unit costs of £517 for hip replacement surgery. To be consistent, we exclude this provider from all analyses. This has no significant effect on our results.

[33] The NHS IC has subsequently been renamed the *Health and Social Care Information Centre*.

[34] The NHS IC also provides these averages adjusted for case-mix. However, we used the unadjusted data because a) at the time of writing the case-mix adjustment methodology was not yet finalised, and b) we undertake our own case-mix adjustments.

some or extreme problems on five dimensions (mobility; self care; usual activities; pain/discomfort; anxiety/depression) (Brooks 1996; Kind et al. 2005). These responses are used to describe a patient's EQ-5D health profile. That health profile is summarised using utility[35] weights obtained from members of the general public (Dolan 1997), anchored at 1 (full health) and 0 (dead), with scores < 0 indicating states considered worse than being dead. We multiply the EQ-5D utility scores with 100 to align its scale with the other PROMs. The patient also provides their own assessment of their overall health state on a visual analogue scale — the EQ-VAS — from 0 to 100 (worst to best imaginable health, respectively).

The condition-specific Oxford Hip and Knee Scores (OHS/OKS) consist of 12 questions, each of which requires responses on a 5-point severity scale (Dawson et al. 1996; Dawson et al. 1998). Equal importance is given to all questions and summary scores range from 0 (worst) to 48 (best). The Aberdeen Varicose Vein Questionnaire (AVVQ) contains 13 questions and scores between 0 and 100 (Garratt et al. 1993). In contrast to the aforementioned instruments, higher scores on the AVVQ indicate worse health states. We invert the scale of the AVVQ (i.e. 0 = worst, 100 = best) to facilitate interpretation and comparison across instruments.

2.5 Results

2.5.1 Descriptive statistics

We present descriptive statistics in Table 2.2.

Each of the four conditions is sufficiently populated to allow for precise estimation of case-mix effects at patient level. In contrast, the number of providers is comparably low (124 to 146 hospitals), reinforcing the value of multilevel analysis as compared to traditional hospital-level analysis.

The cost of care varies considerably across providers for each of the four pro-

[35] The Dolan tariff is derived from preference data obtained through time trade-off exercises. The resulting weights therefore do not constitute von Neumann-Morgenstern utilities since the elicitation exercises did not involve decisions under risk. However, the UK TTO weights are commonly referred to in the literature as 'utility weights' and we shall therefore follow the same convention.

Table 2.2: Descriptive statistics

	Knees		Hips		Groin hernia		Varicose veins	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Patient characteristics								
Cost of care in GBP, adjusted for MFF	6137.47	2073.02	6344.12	2096.23	1518.75	727.95	1246.52	565.19
Patient age (in years)	69.37	9.64	68.71	11.47	59.08	17.26	50.36	14.72
Patient gender (1=male, 0=female)	0.43	0.49	0.39	0.49	0.91	0.28	0.38	0.49
Admission: Transferred in from another provider	0.00	0.04	0.01	0.08	0.00	0.03	0.00	0.02
Discharge: Other	0.02	0.15	0.03	0.17	0.00	0.06	0.00	0.04
Discharge: Died in hospital	0.00	0.04	0.00	0.05	0.00	0.02	0.00	0.00
Discharge: Transferred out to another provider	0.02	0.13	0.03	0.16	0.00	0.03	0.00	0.01
Under the care of multiple consultants	0.02	0.14	0.03	0.17	0.01	0.09	0.00	0.04
Number of secondary procedure codes	1.40	0.96	1.48	1.13	1.38	0.73	1.55	0.91
Number of secondary diagnosis codes	2.53	2.19	2.52	2.27	1.19	1.65	0.56	1.06
Weighted Charlson comorbidity index	0.41	0.72	0.38	0.79	0.22	0.60	0.10	0.34
Neighbourhood income deprivation	0.14	0.11	0.13	0.10	0.14	0.11	0.16	0.12
Number of indicators for main procedure	6		8		7		4	
Number of indicators for main diagnosis	5		5		0		5	
Provider characteristics								
Teaching hospital (1=yes, 0=no)	0.16	0.36	0.15	0.35	0.19	0.39	0.25	0.43
Procedure volume (scale)	560	259.62	520	265.77	478	191.77	311	208.55
Degree of specialisation (scope)	0.30	0.11	0.32	0.13	0.32	0.06	0.31	0.06
Oxford Knee Score - Initial health status	18.61	1.62	-	-	-	-	-	-
Oxford Knee Score - Health gain	14.56	1.65	-	-	-	-	-	-
Oxford Hip Score - Initial health status	-	-	17.90	1.71	-	-	-	-
Oxford Hip Score - Health gain	-	-	19.57	1.76	-	-	-	-
Aberdeen Varicose Vein Score - Initial health status	-	-	-	-	-	-	80.88	2.24
Aberdeen Varicose Vein Score - Health gain	-	-	-	-	-	-	8.11	2.27
EQ-5D utility score - Initial health status	40.05	5.81	34.52	5.81	78.88	3.10	76.45	4.63
EQ-5D utility score - Health gain	29.21	5.68	40.93	5.31	8.42	2.67	9.70	5.45
EQ-VAS - Initial health status	68.58	3.60	65.74	4.01	80.07	2.28	80.61	3.62
EQ-VAS - Health gain	2.52	3.39	8.70	4.33	-0.95	1.81	-0.89	2.73
Number of patients (N)	60,780		53,235		57,346		23,079	
Number of hospitals (J)	140		138		146		124	
Number of patients per hospital (n_j)								
Minimum	14		37		1		4	
Average	434		386		393		186	
Maximum	1,307		1,181		975		904	

SD = standard deviation; MFF = market forces factor cedures. For example, for knee replacement surgery we observe average provider costs that range from below £2,000 to more than £10,000. High cost cases are not confined to one or two providers. Rather, we observe that many hospitals report cost for patients in excess of two standard deviations above the national average.

The generic nature of EQ-5D and EQ-VAS allows for comparison of health outcomes across conditions. Patients undergoing hip or knee replacement experience on average larger increases in health status than those receiving groin hernia or varicose vein surgery. This is consistent with the less serious nature of the underlying conditions. We observe disagreement between EQ-5D and EQ-VAS on the direction of health change for the latter groups of patients. Most providers report improvements in average health when measured with the EQ-5D. In contrast, more than 50% of providers in our sample report negative average health outcomes when measured by the EQ-VAS. Whether this is a result of aggregation (from patient to provider level) or a genuine difference between instruments cannot be explored with our dataset.

2.5.2 Regression results

2.5.2.1 Baseline estimates

Table 2.3 presents regression results from a model with EQ-5D outcome information.

We find significant coefficients on the majority of HRG variables (not reported; see Appendix Tables A2.2 - A2.5 for full results). This indicates that the current reimbursement system is able to distinguish between different types of patients and their expected costs. That said, several other patient characteristics explain costs over and above the allocated HRG, for example the patient's age and certain types of main diagnoses and procedures (see Appendix). Costs are also higher for patients that undergo more procedures or suffer from a higher number of comorbidities as well as for patients that are transferred in or out of hospital or are not discharged to their usual place of residence.

The results at provider-level are less clear cut. The average cost of patients treated in teaching hospitals is generally higher than in non-teaching hospitals but the effect

Table 2.3: Regression results with EQ-5D outcome information

	Knees		Hips		Hernia		Veins	
Variable	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Patient-level								
Intercept	8319.7	1320.5 ***	7355.8	1911.7 ***	1642.6	915.7	814.8	830.6
Age group 2	24.7	11.9 *	17.8	20.6	13.4	6.5 *	20.8	7.5 **
Age group 3	-10.7	21.6	53.1	21.4 *	28.8	7.9 ***	10.5	7.1
Age group 4	13.8	17.9	42.2	24.8	64.5	9.8 ***	23.6	10.6 *
Age group 5	70.6	16.2 ***	82.2	27.4 **	147.0	15.3 ***	32.3	13.7 *
Male patient	-11.1	8.1	-7.5	15.1	-22.8	11.7	21.1	5.2 ***
Admission: Transferred in from another provider	1696.1	692.5 *	1490.4	442.1 ***	660.4	224.5 **	43.5	218.6
Discharge: Other	128.5	77.2	231.1	54.9 ***	376.6	99.4 ***	123.2	66.6
Discharge: Died in hospital	-334.0	354.6	-171.1	291.1	751.4	432.0	-	-
Discharge: Transferred out to another provider	323.1	94.5 ***	145.1	93.3	513.3	212.3 *	-364.0	199.8
Under the care of multiple consultants	-125.0	68.7	-353.6	116.8 **	206.5	53.4 ***	296.0	136.6 *
Number of secondary procedure codes	155.4	21.5 ***	175.5	28.6 ***	78.3	14.8 ***	10.8	6.4
Number of secondary diagnosis codes	24.5	5.5 ***	46.3	7.0 ***	34.0	5.2 ***	22.8	5.9 ***
Weighted Charlson comorbidity index	-4.7	9.7	-16.3	15.0	54.9	9.8 ***	-2.0	12.3
Income deprivation - 1st quintile (lowest)	4.5	13.6	-8.4	24.7	-16.4	9.5	-25.5	8.6 **
Income deprivation - 2nd quintile	8.0	13.7	-6.8	21.4	-9.0	8.3	-9.4	8.7
Income deprivation - 3rd quintile	30.3	12.4 *	16.4	17.4	-7.9	6.7	-11.7	10.5
Income deprivation - 4th quintile	3.3	12.2	17.9	23.5	-5.0	6.8	-14.4	8.3
Provider-level								
Teaching hospital	-63.4	419.8	418.2	394.7	188.0	81.7 *	95.7	85.6
Procedure volume (scale)	-0.5	0.6	0.3	0.6	-0.5	0.3	0.0	0.2
Degree of specialisation (scope)	257.0	800.0	2542.0	990.4 *	-225.9	674.4	317.2	435.8
Average pre-operative EQ-5D score	-50.5	25.0 *	-51.5	26.1 *	-0.6	12.9	3.3	10.4
Average change in EQ-5D scores	-6.5	22.7	-22.0	31.9	-2.2	14.0	-1.8	8.8
τ	1556.2	156.7 ***	1502.2	170.4 ***	400.4	31.5 ***	394.5	27.5 ***
σ	1222.6	273.0 n/a	1355.4	91.3 n/a	523.0	24.2 n/a	379.0	22.8 n/a
ρ	0.62	0.12 ***	0.55	0.06 ***	0.37	0.04 ***	0.52	0.04 ***
$LogL$	-518,760		-459,874		-440,692		-170,072	
N	60,780		53,235		57,343		23,077	
J	140		138		146		124	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Standard errors are robust to heteroscedasticity. Significance of σ_u is established through Likelihood Ratio tests. Standard error and significance of ρ is calculated using the delta method.

Notes: All models include indicators for ten most frequent HRGs and 'other HRG' category as well as for condition-specific main diagnosis and procedures groups. Age categories are: Knees ($< 61; 61 - 66; 67 - 72; 73 - 77; > 77$), Hips ($< 60; 60 - 66; 67 - 72; 73 - 77; > 77$), Hernia ($< 44; 44 - 56; 57 - 65; 66 - 74; > 74$), Veins ($< 37; 37 - 44; 45 - 53; 54 - 63; > 63$).

is statistically significant only for groin hernia repair. We do not find conclusive evidence that NHS hospitals realise positive economies of scale or scope. This is somewhat surprising given the substantial differences in volume and, to a lesser degree, specialisation observed across providers.

With respect to PROM scores, we find that the coefficient on initial health status shows the expected negative sign for three out of four conditions but is only statistically significant for the two orthopaedic procedures. Hospitals serving, on average, healthier patients have lower average patient costs than those admitting patients with lower health status; a result that seems intuitively correct. The size of the effect is, however, relatively small: a SD increase in average pre-operative EQ-5D score is estimated to decrease the cost of knee replacement surgery by merely -£293. The association between average health outcome and costs is negative in all four models. This would indicate that some providers are able to secure greater health gains and provide care at lower costs than other providers. However, no results are statistically significant at the 95% confidence level.

2.5.2.2 Alternative PROM instruments

While there are good reasons to prefer generic instruments over condition-specific instruments, for example because the former facilitates broader comparisons across disease areas, one should not *a priori* exclude the latter for comparative cost analysis. We re-estimate the various models using condition-specific PROMs and, as an alternative to utility weighted EQ-5D profile, the EQ-VAS, and present results in the first column of Table 2.4.

With few exceptions, coefficients on health gain are negative, indicating that, on average, larger changes in patient health are associated with lower costs. However, only the coefficient on health gain measured by the EQ-VAS for knee replacement surgery is statistically significant. All coefficient estimates can be interpreted as marginal effects, i.e. a one point increase in average EQ-VAS health outcome is expected to reduce unit costs after knee replacement by about £79.

Table 2.4: Relationship between health outcome and costs

	Constant MCQ		Non-constant MCQ				
Variable	H		H		H^2		Test of joint significance
	Estimate	SE	Estimate	SE	Estimate	SE	$\chi^2(2)$
Knee replacement							
EQ-5D	-6.5	22.7	32.6	31.7	-0.8	0.7	1.59
EQ-VAS	-78.6	37.0 *	-117.8	44.6 **	5.5	4.6	7.19 *
OKS	-53.9	66.1	382.3	228.2	-15.5	7.7 *	5.15
Hip replacement							
EQ-5D	-22.0	31.9	-313.1	164.2	3.7	2.1	4.27
EQ-VAS	-55.7	40.0	-105.9	58.5	3.3	3.4	3.66
OHS	36.0	86.0	-2,466.3	778.2 **	63.7	20.2 **	10.05 **
Groin hernia repair							
EQ-5D	-2.2	14.0	-42.9	33.4	2.3	1.8	1.72
EQ-VAS	-10.2	17.4	-10.0	18.2	0.3	4.3	0.36
Varicose vein surgery							
EQ-5D	-1.8	8.8	8.3	18.7	-0.5	0.7	0.59
EQ-VAS	17.5	12.0	26.1	11.1 *	4.2	1.9 *	9.86 **
AVVQ	14.8	20.1	147.9	71.1 *	-7.6	4.0	4.46

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$
Standard errors are robust to heteroscedasticity.
OHS/OKS = Oxford Hip/Knee Score; AVVQ = Aberdeen varicose vein questionnaire; MCQ = marginal costs of quality; H = health outcome, i.e. the change in health status after treatment

2.5.2.3 Non-constant marginal costs of quality

So far, we have assumed that the marginal costs of quality are constant across the entire range of observed health outcomes. This assumption may be too restrictive to accommodate previous empirical findings and theoretical considerations (see Section 2.2). Following Fleming (1991) and Weech-Maldonado et al. (2006), we allow for a non-linear association between costs and outcomes by including squared terms in our model. Coefficient estimates are presented in the third and fifth column of Table 2.4. We focus on those which are jointly significant at a critical value of α=0.05.

For hip replacement surgery, we observe a statistically significant, non-linear association between health outcomes and risk-adjusted costs when the former are measured by the OHS. The marginal effect of outcomes on costs is negative at low levels of health outcome but turns positive after passing a saddle point. At the 25 percentile of the OHS outcome distribution, a one point increase in average outcome is associated with a cost reduction of £70 per patient. In contrast, unit costs are expected to increase by £79 when starting at the mean value.

For knee replacement surgery, the coefficient on the linear EQ-VAS health outcome term is statistically significant and negative but the squared term is insignificant. For varicose vein surgery, the estimated relationship is positive and exponential for the EQ-VAS. One may interpret this as the right-hand side of the U-shaped relationship that we observe for hip replacement surgery since EQ-VAS scores for varicose vein surgery also take on negative values.

Results for the other models are statistically insignificant or similar to the linear models.

We tested the robustness of these results to the exclusion of variables representing provider-level constraints (i.e. measures of scale and scope, and teaching status) since these may not be binding. Results are very similar; see Appendix Table A2.6. We also re-estimated all models excluding two specialised orthopaedic hospitals[36] since these may, arguably, operate under different production constraints (e.g. case-mix) that are unobservable to us. Again, results are robust to this modelling choice (Appendix Table A2.7).

2.5.3 Impact on provider effects

We now turn to the assessment of providers' efforts to contain costs. We illustrate our results using the example of hip replacement surgery and the Oxford Hip Score.

Figure 2.2 shows the Empirical Bayes estimates of provider effects obtained from the restricted model where health outcomes are not taken into account. Hospitals to the left of the graph have lower average costs than hospitals to the right. Bayesian 95% credible intervals are formed from the posterior distribution of each provider effect (Skrondal and Rabe-Hesketh 2009).

We find substantial differences in provider effects after accounting for case-mix, production constraints and average initial health. 95% of providers are located within the range of -£2,700 to +£3,920 around their expected costs (here normalised to

[36] These are the Royal Orthopaedic Hospital in Birmingham (provider code: RRJ) and the Royal National Orthopaedic Hospital in London (RAN).

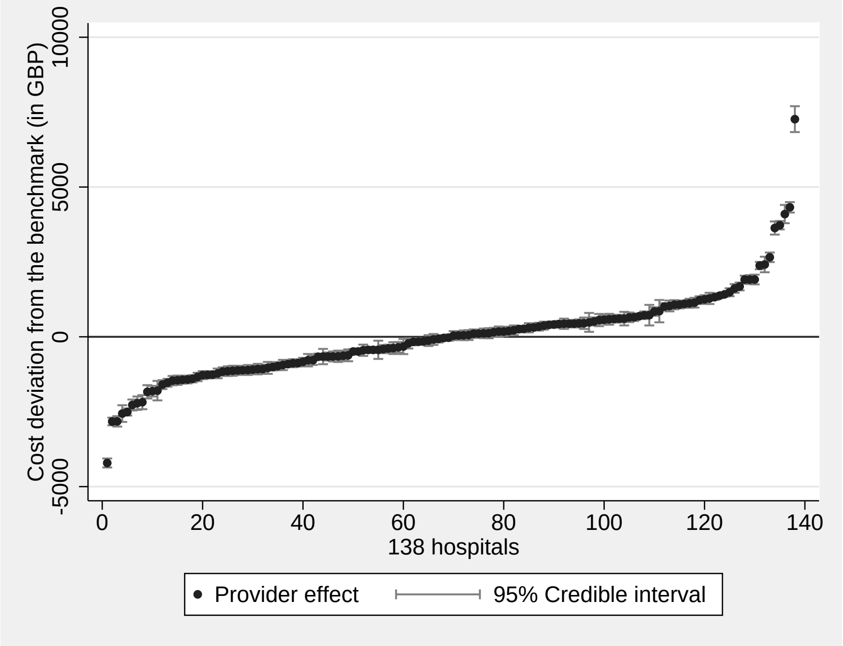
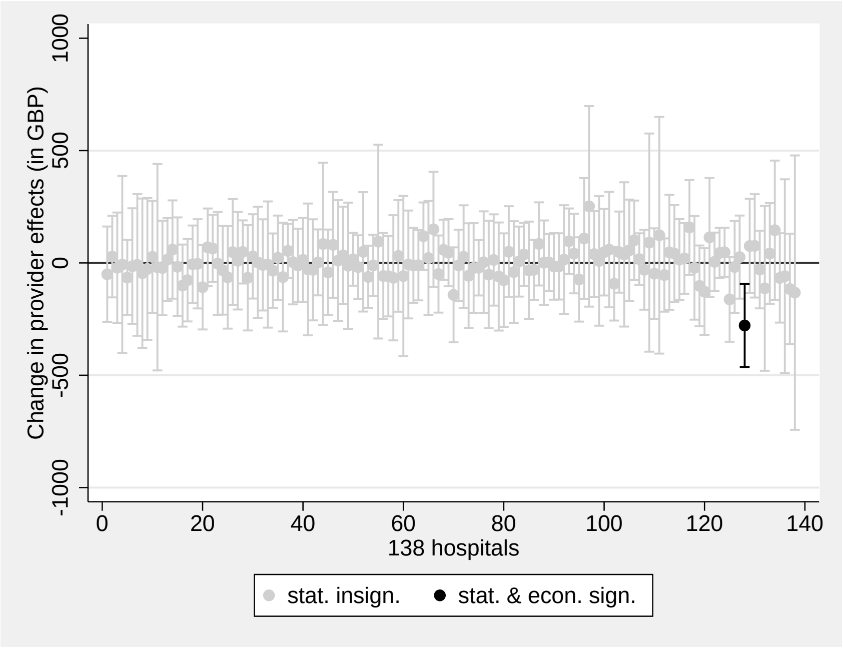


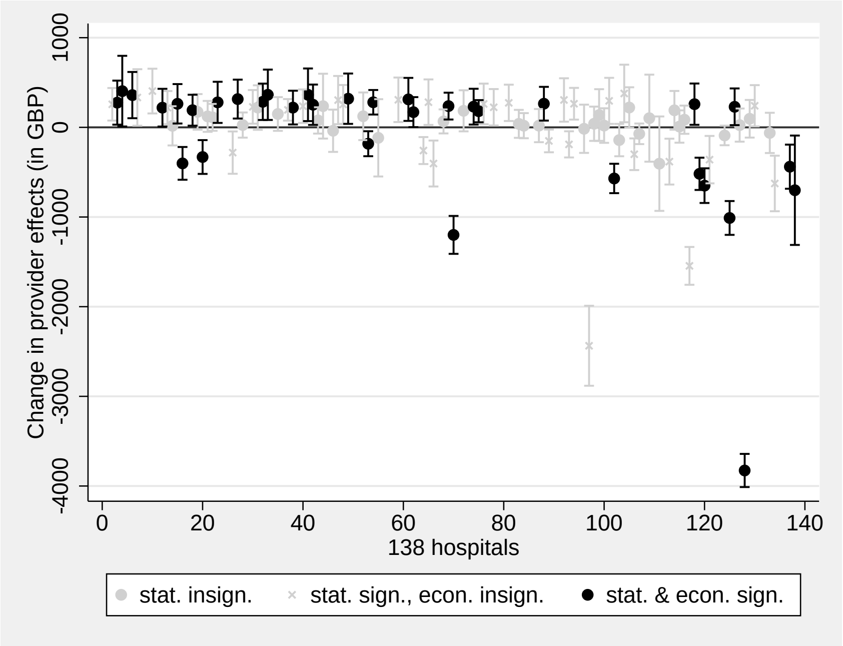
Figure 2.2: Hospital cost performance for hip replacement surgery, unadjusted for health outcomes

zero). The 'best' hospital has risk-adjusted production costs that lie about £4,210 below the benchmark, whereas the 'most expensive' hospital lies about £7,270 above. Neither of these two hospitals are specialised orthopaedic providers, nor are they otherwise unusual in their observable characteristics.

Differences in costs, while substantial, do not seem to be driven by variation in average health outcome when marginal costs of quality are modelled as constant. Comparing the estimates of a model with linear OHS health outcome term and the restricted model, we find that, for most hospitals, the adjustment does not result in different judgements with regard to their relative cost performance (Figure 2.3a; ordered as in Figure 2.2). Only one hospital experiences a change that is statistically significant. The magnitude of the adjustment is £279.



(a) OHS, constant marginal costs of quality



(b) OHS, non-constant marginal costs of quality

Figure 2.3: Change in estimated provider cost performance after accounting for average health outcome

The picture changes when comparing provider effects from the restricted model and a model with non-constant marginal costs of quality (Figure 2.3b). For a large number of hospitals, quality-adjusted cost performance differs statistically significantly from their unadjusted estimate by less than £500 in absolute terms. This said, a small group of hospitals experience changes greater than £1,000. Unsurprisingly, the hospital with the largest adjustment (-£3,828) reports the largest health outcome and therefore profits most from allowing for a non-linear relationship between costs and quality. Note that, out of 74 hospitals that experience statistically significant changes only 34 are 'economically significant'. These hospitals are estimated to face positive marginal costs of quality. As we earlier argued, regulators should not amend judgement about the 30 hospitals that face negative marginal costs of quality ('economically insignificant').

Results for the other procedures and PROMs are reported in Appendix Table A2.8. Again, we do not find estimates of hospital cost performance to be greatly affected by the addition of health outcome information.

2.6 Discussion

The objective of this study is to measure cost variation in the provision of four surgical procedures and to account for differences in the quality of care provided. Our work builds on a new policy initiative by the English Department of Health to collect patient-reported health outcome data using generic and condition-specific instruments. This study is a first attempt to incorporate patient-reported health outcomes into comparative cost analysis and explore whether this new measure of quality changes judgements about the relative performance of NHS hospitals. We use multilevel modelling techniques to distinguish random cost variation at patient-level from systematic variation at provider-level. We obtain precision-weighted Empirical Bayes estimates of provider effects and interpret these as relative measures of cost containment effort.

Our results suggest that even after allowing for (exogenous) patient or production process characteristics there exists systematic cost differences across hospitals in the provision of surgical procedures. These differences are substantial and economically significant. For example, even after excluding very high/low cost providers (top and bottom 2.5% of the distribution), hospitals still report average risk-adjusted costs for hip replacement surgery that differ by up to £6,600. Some of this variation in costs is associated with provider differences with respect to their patients' pre-operative health, although the overall effect is small. Variation in costs may also relate to the average health outcomes and we find evidence of a non-linear relationship between costs and outcomes for hip replacement surgery. For some hospitals, such health outcome adjustment leads to significant improvement in their relative cost performance but the effect is generally small in magnitude. Furthermore, we have argued that the economic judgement should differ depending on whether the hospital faces positive or negative marginal costs of quality and can reduce costs without negative effects on health outcomes.

Several implications for policy makers and future research arise from our results. First, even a number of years after the introduction of PbR, there remains marked, and largely unexplained, variation in costs across providers of the same care. This suggests that the prospective payment system with yardstick competition has failed one of its purposes, namely to reduce variation and change provider behaviour. Since policy makers in the English NHS seem reluctant to let providers exit the market as a result of overspending, losses due to excessive production costs will ultimately fall onto the health budget and displace other valuable healthcare.

Second, as the impact of health outcome information on estimates of cost containment effort seems, at best, minimal, it casts doubt on claims that might be made by some hospitals that their substantially higher production costs are a consequence of investing in better care that produces better health outcomes. Similarly, we find that only a small part of the observed variation in costs is explained by differences in patients' pre-operative health after adjusting for patient case-mix. Taken together, this suggests that, for the condition studied, the link between health and expenditure is weak.

Third, if the relationship between cost and quality is indeed non-linear, pay-for-performance and quality bonus programs have to acknowledge non-constant marginal costs and set different prices for different levels of health outcome. If the association between outcomes and cost is negative or non-existent (see e.g. groin hernia repair) then quality bonus payments of any form should be understood as incentive payments in excess of production costs. The way in which quality incentive schemes are designed might therefore differ by procedure.

Fourth, at this early stage of the PROM initiative and on the basis of our preliminary analysis, we cannot single out a preferred PROM instrument that should be applied exclusively in future analyses of hospital cost performance. Users of PROM information may prefer different instruments for a number of valid reasons. For example, one may argue that the vague definition of endpoints on the EQ-VAS ('best/worse *imaginable*') make it difficult to compare across patients or even across repeated responses by the same patient, and hence prefer to base inferences about provider performance on the EQ-5D descriptive system. However, the present study was not designed to explore differences in responsiveness and construct validity across PROMs, nor should any normative statements be made on the basis of descriptive statistics of the data. We therefore recommend using both generic and condition-specific instruments and conducting sensitivity analysis with regard to the choice of instrument as we have done here.

Our study has a number of relevant limitations, many of which are data related. First, while PROM data are collected at patient-level, these were not available to us at the time of study. Instead, our analysis utilises publicly available data averaged at provider-level. This may be problematic for two reasons: i) the association between costs and health outcomes is estimated from less variable data and statistical power is reduced accordingly, and ii) within-hospital heterogeneity with respect to health outcomes cannot be taken into account. Depending on the degree of heterogeneity, one may observe an association at provider-level that differs from (or even contradicts) the true association at patient-level (Robinson 1950).

Second, our measure of costs is derived from a top-down costing system, where overall costs are assigned to HRGs on the basis of activity within cost centres. Currently, this forms the best estimate of patient-level costs routinely calculated in the England NHS and is used in applied research (e.g. Laudicella et al. 2010) as well as for price setting purposes (Department of Health 2002; Daidone and Street 2011). However, there is a risk that overhead and indirect costs are assigned incorrectly or that patients are assigned equal costs when consuming different amounts of resources (limited within-product variation). Other healthcare systems, most notably the US Veterans Health Administration system, operate bottom-up (i.e. activity-based) costing systems, where resources devoted to individual patients are measured at the level of intermediate products (e.g. day on ward, unit of medication) and then summed across the inpatient stay (Carey and Burgess 2000). Bottom-up costing is generally regarded as preferable because it reflects true resource consumption more accurately.[37] Carey and Stefos (2011) have compared cost data derived from both systems and found patient costs to be higher and more variable under bottom-up costing, leading to different estimates of the costs of adverse events. While we cannot assess whether their findings hold for English hospitals, we have to acknowledge that our results may understate the impact of quality on costs. This may also explain why we do not observe evidence of economies of scale.[38]

Third, because we use observational data our analysis may suffer from various forms of endogeneity bias. We are especially concerned about the potential endogeneity of health outcomes and costs (Braeutigam and Pauly 1986; Gertler and Waldman 1992). If providers choose their level of resource allocation and quality of care simultaneously, the health outcomes in our model would be endogenous, and coefficient estimates would be biased downwards. To address this problem, one requires suitable instrumental variables (IV) that are sufficiently correlated with outcomes, but not with costs. Carey and Burgess (1999) use measures of past quality

[37] However, as pointed out by Jackson (2001) all costing systems make assumptions about cost allocation, so that a perfect representation of true resource use is unlikely to be achieved.

[38] An alternative to cost analysis would be to study variation in length of stay; a proxy for resource utilisation. See Chapter 4 for an analysis of the potential trade-off between length of stay and health outcome.

to instrument current quality. Focussing on the effect of resource use on outcomes Hauck and Zhao (2011) employ weekday and month of admission as IVs for length of stay, whereas Picone et al. (2003) and Schreyögg and Stargardt (2010) explore variation in regional price levels. Given our data limitations and the focus of our analysis, none of these IVs are applicable. Furthermore, as we model the relationship between patient-level costs and provider-level quality measures, we would expect endogeneity to be less accentuated. Still, we must conclude that our study has only explored the association between cost and health outcomes, but cannot ascertain causality.

Finally, while patient-reported outcome measures provide a more detailed picture of the health outcomes experienced by patients, they may not completely reflect the quality of hospital care for various reasons: i) PROMs are, by definition, subjective and may thus be affected by reporting bias, ii) health outcomes may be influenced by events taking place before admission or after discharge over which the hospital has no control, e.g. the care provided after discharge, and iii) PROM scores may be prone to selection bias in the form of non-random participation or drop-out if patients suffer poor outcomes (notably death).[39] We cannot explore these issues with our data. However, the use of hospital mean scores may prove helpful in this situation because some of the above effects are likely to be mitigated by averaging across patients. Clearly, further research is required to explore the validity and limitations of patient-reported health outcomes for provider performance assessments.

Acknowledgements

We would like to thank David Nuttall, Steve Morris, two anonymous referees and participants of the Health Econometric Data Group seminar series (York) and the HESG Summer 2011 conference (Bangor) for their inputs and comments. The project was funded by the Department of Health in England as part of a programme

[39] We have subsequently explored selection bias in hospital PROM scores and found them robust to non-random participation (Gomes et al. 2015).

of policy research. The views expressed are those of the authors and may not reflect those of the funder.

Hospital Episode Statistics are copyright ©2015, re-used with the permission of The Health & Social Care Information Centre. All rights reserved.

3 Hospital variation in patient-reported outcomes at the level of EQ-5D dimensions

3.1 Introduction

Recent years have seen a growing trend to measure and publish hospital data on health outcomes in order to facilitate patient choice and increase provider accountability (Marshall et al. 2003; Cutler et al. 2004). The focus of these activities has been on measures of mortality, re-admission or adverse events, which are easily derived from clinical records but reveal little about the health of the vast majority of patients. In order to allow for a more sensitive assessment of hospital performance it is necessary to move away from a focus on relatively rare 'failure' outcomes towards more comprehensive and sensitive measures of patients' health outcomes (Kind and Williams 2004; Appleby et al. 2010; McGrail et al. 2012).

Since April 2009, all providers of publicly-funded inpatient care in the English National Health Service (NHS) have been required to collect both EQ-5D (Brooks 1996) and condition-specific data for four elective procedures: unilateral hip and knee replacements, varicose vein surgery, and groin hernia repairs (Department of Health 2008a). Eligible patients are invited to report their health status before and three or six months after surgery. The changes in patients' health status are expected to *'provide an indication of the outcomes or quality of care delivered to NHS patients'* (Department of Health 2008a, p.5) and can be analysed to identify systematic variation across hospital providers with finer granularity than previously possible.

Traditionally, patient-reported outcome measures (PROMs) have been collected and analysed primarily within clinical trials to assess the treatment effect on patients' health. Their application in the context of routine performance assessment on a national scale breaks new ground and requires an appropriate methodology which takes into account the characteristics of the data and their intended use as measures of the relative quality of hospital treatment (Smith and Street 2013).

The NHS Information Centre has developed a preliminary risk-adjustment methodology that is currently being applied to the PROMs data (Coles 2010). For the EQ-5D, this involves transforming the patients' EQ-5D health profiles into utility-weighted index scores and estimating linear regression models to relate post-treatment utility scores to the pre-treatment scores and case-mix controls. The advantage of this approach is that patient health is expressed in terms of a (quasi-)continuous score, which facilitates statistical analysis and allows for ranking of hospitals with respect to a single performance metric: their ability to influence post-treatment utilities or, equivalently, changes in scores over time. However, for the purposes of performance measurement, identifying best practice and informing patient choice, the costs of aggregation may outweigh the benefits. We build this argument around three points.

First, any form of aggregation causes loss of detail and information (Smith 2002). Once constructed, an index measure cannot reveal information about the underlying components and the degree to which hospitals affect these. Certain hospitals may perform well on one EQ-5D dimension but fall short on another. Detailed information on the performance on each dimension can help to identify the source of the problem and foster improvement through adoption of best practice (Smith 2002).

Second, the use of an aggregation function introduces exogenous variation that can bias statistical inference and raises normative concerns about whose preferences the weights should reflect (Smith 2002; Goddard and Jacobs 2009; Parkin et al. 2010). In some circumstances, one may be willing to accept the weights underpinning the aggregation function, for example, when conducting economic evaluations of health technologies from a societal perspective (Siegel et al. 1997). But this is not always justified. The use of aggregate outcome data to inform patients' choice of hospital raises normative concerns because it imposes a common valuation of health dimensions. In fact, reporting relative hospital performance with respect to risk-adjusted post-operative EQ-5D utility is only justified if all (prospective) patients share the same relative values. But patients may be heterogeneous with respect to their relative valuations of health dimension or their relative valuations may differ from those of the general public (De Wit et al. 2000; Mann et al. 2009). If so, analysing variation on the level of health dimensions is more appropriate as it allows patients to apply their own values when interpreting performance data.[40]

Third, the use of performance data derived from EQ-5D utility scores may be limited by patients' difficulties in interpreting these quantities. In a recent qualitative study, Hildon and colleagues 2012 interviewed patients and clinicians about their views on four different metrics of hospital PRO performance, including mean follow-up score, mean change in score, proportion reaching a specified threshold at follow-up, and proportion reaching a minimally important difference. Their results suggest that *'for patients [. . .], unlike measures of height or weight, PRO[..] scores are unfamiliar and their values have no immediate meaning. It's therefore necessary to transform them into interpretable forms, or indeed into experiences rather than metrics, to make them useful'*. Furthermore, patients *'could not distinguish between the four [metrics], but liked a percentage or what was for them intuitive scaling'* (Hildon et al. 2012, p.11). Analysing responses on EQ-5D dimensions rather than utility scores allows reporting performance in a similar form to the way that the data were originally collected. Hospitals could then be compared with respect to the risk-adjusted probability of a given patient to report, for example, no problems with mobility or pain/discomfort at follow-up.

To explore these claims, we assess hospital performance with respect to self-reported health outcomes for hip replacement patients. We focus on the EQ-5D and develop multilevel risk-adjustment models for each of the five functional dimensions. Our approach draws on the literature on longitudinal modelling (Bryk and Raudenbush 1988; Molenberghs and Verbeke 2005; Hedeker and Gibbons

[40] Devlin et al. (2010) propose using Pareto dominance criteria to compare changes in patients' EQ-5D health profiles across time without imposing value judgements. However, this approach leads to information loss since neither the magnitude of change nor the distribution of health effects across health dimensions is considered.

2006) and performance assessment (Raudenbush and Willms 1995; Goldstein and Spiegelhalter 1996) to analyse variation in treatment impact across hospitals. More specifically, we model the hospital-specific contribution to post-treatment EQ-5D response as a random coefficient that varies between hospitals. The Empirical Bayes (EB) estimates of this coefficient are then interpreted as capturing relative hospital quality. We assess the correlation between performance assessments on the level of EQ-5D dimensions and aggregated utility scores.

3.2 Methods

3.2.1 Data

Our study exploits EQ-5D data routinely collected from English patients having a hip replacement during April 2009 to March 2010. All providers of NHS-funded care are required to participate in the survey (Department of Health 2008a). This includes all NHS-operated hospitals and private treatment centres. Patients aged 15 or over that undergo elective, unilateral hip replacement surgery are invited to take part in the survey (Department of Health 2008a). We extract information on each patient's pre- and post-operative EQ-5D health profile and EQ-5D utility score, where the latter is calculated using the UK time trade-off (TTO) utility weights (Dolan 1997). The pre-treatment (baseline) survey is collected either during the initial outpatient appointment that precedes hospital admission or at the day of admission. Follow-up data are collected by the NHS Information Centre via postal survey approximately 6 month after surgery. To ensure consistency with respect to the timing of measurements while retaining as much information as possible, we exclude all observations for which the recorded time between baseline survey and admission exceeds 12 weeks or the follow-up period is either shorter than 20 weeks or longer than one year.

We link these data to the Hospital Episode Statistics (HES) inpatient database, which contains detailed information on all inpatient care provided in English hospitals. The depth of information contained in HES allows us to account for a wide range of clinical and demographic risk adjusters. These include the most frequent main diagnoses (e.g. osteoarthritis (ICD-10: M15-19), rheumatoid arthritis (ICD-10: M05-06))(Singh 2011), the weighted Charlson score of comorbidities (Charlson et al. 1987; Sundararajan et al. 2004; Bjorgul et al. 2010), the number of additionally coded comorbidities, whether it was a primary or revision surgery and whether the revision was due to problems with the existing implant (ICD-10: T84), patient age, gender and the deprivation profile of the patient's neighbourhood of residence (Noble et al. 2006; Clement et al. 2011; Neuburger, Hutchings, Black et al. 2013). We only retain patient records that can be matched to the PRO survey and for which we observe a full EQ-5D profile at baseline and follow-up.

3.2.2 Statistical modelling

The objective of the empirical analysis is to obtain estimates of the relative systematic impact of hospital providers on patients' post-treatment health outcomes. We estimate hierarchical ordered probit models (Breslow and Clayton 1993; Gibbons and Hedeker 1997; Greene and Hensher 2010), separately for each of the five EQ-5D dimensions. We then compare the results to those obtained from a linear regression on the EQ-5D utility scores to study the practical implications of using disaggregated health dimensions for assessment of hospital performance.

Let y^*_{ijt} denote the health status (with respect to e.g. anxiety/depression) of patient $i = 1, \dots, n_j$ in hospital $j = 1, \dots, J$ at time point $t \in [0, 1]$. Health status is assumed to be continuous but not directly observable. Instead, we observe patients' own assessment of their status on the three-point EQ-5D response scale ($m = 1, 2, 3$ with 1 = no problems, 2 = some problems, 3 = extreme problems). The mapping of latent, continuous status y^*_{ijt} to observed, discrete responses y_{ijt} is given by the standard threshold model (McKelvey and Zavoina 1975)

$$y_{ijt} = \begin{cases} 3, & if\ y^*_{ijt} \leq \kappa_1 \\ 2, & if\ \kappa_1 < y^*_{ijt} \leq \kappa_2 \\ 1, & if\ y^*_{ijt} > \kappa_2 \end{cases} \tag{3.1}$$

where the threshold parameters κ are unobserved and must be estimated from the data. The categories are ordered from worst to best. This facilitates the qualitative interpretation of regression coefficients, where a positive sign indicates improvements in latent health and, thus, the probability of reporting no problems.

Each patient provides measures of their health status pre- and post-treatment. Both responses are outcomes of the same measurement process as well as being (partly) determined by common factors, such as patient characteristics and baseline level of latent health. Our interest lies in the latent health gain that follows from hospital treatment and the degree to which variation in health gain can be systematically associated with the provider of care. We make the assumption that, conditional on baseline health and a set of risk-adjustment factors, patients do not select into hospitals based on unobservable characteristics and that the health of patients in different hospitals would follow the same trajectory if untreated. This allows us to interpret hospital variation in latent health gain as a measure of relative quality.

Our data are characterised by a hierarchical structure, with measurement points clustered in patients, which themselves are clustered in hospitals. Given the non-linear nature of our model, these data can be analysed in two ways. One can collapse the hierarchy into two levels and model post-treatment latent health as a function of lagged, observed (pre-treatment) response y_{ij0}, observed patient characteristics and a hospital effect (e.g. Contoyannis et al. 2004). Alternatively, one can treat both pre- and post-treatment latent health as left-hand side variables and estimate longitudinal models with unobserved patient heterogeneity (i.e. *growth curve* modelling) (Molenberghs and Verbeke 2005; Hedeker and Gibbons 2006; Greene and Hensher 2010). This is the model advocated by Bryk and Raudenbush (1988) to study variation in learner's trajectories across schools. We adopt this approach because it allows us a) to explicitly account for unobserved, time-invariant determinants of latent health, b) to utilise information contained in both observations to estimate threshold parameters, c) to acknowledge heterogeneity in latent health within a response group as well as random noise in reported pre-treatment health[41], and d) to extend the model in a natural way as more measurement points become available in the future.[42]

Latent health status at any time point t is then described by the outcome equation

$$y^*_{ijt} = \alpha_{ij} + \zeta_j + x^{'}_{ij}\beta + T * (\nu_j + x^{'}_{ij}\delta) + \epsilon_{ijt} \tag{3.2}$$

with

$$\nu_j = \mu + \theta_j \tag{3.3}$$

The vector x_{ij} is a set of patient-level risk adjustment variables which are, in this study, time-invariant and assumed to be strictly exogenous.[43] Treatment is modelled as a dummy variable T, which takes a value of 1 if $t = 1$ (post-treatment) and 0 otherwise. The direct effect of treatment on post-treatment health is given by the coefficient ν_j. We also interact T with x_{ij} to allow for differential effects of patient characteristics on health status at baseline and on the effect of treatment.

Unexplained variation is decomposed into four variance components: i) a patient-specific intercept $\alpha_{ij} \sim \mathcal{N}\left(0, \sigma^2_\alpha\right)$ that captures unobserved, time-invariant patient heterogeneity in latent health[44], ii) a hospital-specific, time-invariant intercept $\zeta_j \sim \mathcal{N}\left(0, \tau^2_\zeta\right)$ that addresses hospital clustering and differences in intake, iii) a random coefficient $\theta_j \sim \mathcal{N}\left(0, \tau^2_\theta\right)$ that varies between hospitals and describes the systematic hospital effect on post-treatment latent health, and iv) a serially uncorrelated error term $\epsilon_{ijt} \sim \mathcal{N}(0, 1)$ that leads to the well-known probit specification. We do not allow for treatment effects to vary by patient, as in e.g. Bryk and Raudenbush

[41] Conversely, in a two-level model we would implicitly assume that pre-operative health status, rather than the patients' self-classification of it, is observed and that patients with identical responses have therefore identical pre-operative health.

[42] The National Joint Registry has announced plans to '*extend[] the pre-operative and post-operative capture of PROMs undertaken through the [Department of Health] programme*' and '*capture further post-operative PROMs from patients having undergone joint replacement surgery*' at one, three and five years post-operatively (National Joint Registry 2011, p.35).

[43] There exists no formal test to verify the assumption of exogeneity in non-linear models of this kind (Greene and Hensher 2010, p.278). Note that patient fixed effects are ruled out by the low number of observations (T≤2) on this level and the resulting incidental parameter bias.

[44] This is equivalent to specifying a model with unobserved patient heterogeneity in threshold parameters.

(1988), since we only observe patients twice. Covariance terms between random effects on the same level of the hierarchy are freely estimated, whereas terms across levels are constrained to zero. The variance partition coefficient ρ describes the extent to which unexplained variation in post-treatment latent health occurs at the level of the hospital and is calculated as follows (Goldstein et al. 2002):

$$\rho = \frac{\tau_\theta^2 + 2 * cov(\theta, \zeta) + \tau_\zeta^2}{\sigma_\alpha^2 + \tau_\theta^2 + 2 * cov(\theta, \zeta) + \tau_\zeta^2 + \sigma_\epsilon^2} \tag{3.4}$$

Larger values of ρ indicate that more variation in post-treatment latent health is attributable to hospital heterogeneity as captured in the hospital-level intercept and the random coefficient on treatment.

For the EQ-5D utility model, we adapt (3.2) to a linear specification with an identity link function (i.e. $y_{ijt}^* = y_{ijt}$) and $\epsilon_{ijt} \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$.

All ordered probit models are estimated by maximum likelihood using `GLLAMM` in Stata 13, where the integrals for the random effects are approximated by adaptive quadrature (8 integration points per random effect) (Rabe-Hesketh et al. 2002). Threshold parameters and the scale of the coefficient are identified through constraints on the mean and variance of the error term and the mean of the intercept. The linear EQ-5D utility model is estimated by maximum likelihood using `xtmixed` in Stata 13.

3.2.3 Provider profiling

Our interest lies in estimates of the relative quality of each hospital, θ_j, captured by the hospital-specific deviation from the average effect of treatment, μ. This parameter is not directly estimated but can be recovered in post-estimation using Bayes Theorem with variance estimates plugged in for the unknown population parameters; a technique known as Empirical Bayes prediction (Skrondal and Rabe-Hesketh 2009).

For the ordered probit models, we describe hospital performance in two different ways. First, we rank hospitals according to their impact on latent health status

y^*_{ij1}. This can be directly inferred from $\widehat{\theta}_j$, where more positive values indicate better performance. Second, we compute the probability of reporting a specific post-treatment outcome $(m = 1, 2, 3)$, based on the estimated quality effort of the hospital. For the average patient treated in a hospital of average patient intake, this is given by

$$Pr\left(y_{j1} = m|\overline{x}, \widehat{\theta}_j, \widehat{\alpha_{ij}} = \widehat{\zeta}_j = 0\right) = \Phi\left(\kappa_m - S_{j1}\right) - \Phi\left(\kappa_{m-1} - S_{j1}\right) \tag{3.5}$$

where

$$S_{j1} = \widehat{\mu} + \overline{x}'\widehat{\beta} + \overline{x}'\widehat{\delta} + \widehat{\theta}_j \tag{3.6}$$

and $\kappa_0 = -\infty$, $\kappa_3 = +\infty$. We calculate 95% credible intervals around $\widehat{\theta}_j$ based on their posterior distribution. Because our interest is on profiling hospital performance with respect to treatment impact we do not consider uncertainty in other parameters estimates when calculating credible intervals for $Pr\left(y_{j1} = m\right)$.[45]

Both methods produce identical rankings of relative hospital performance. However, only the second method relates the result back to the original scale of the PRO survey instrument and allows differences across hospitals to be investigated in terms of the probability of achieving a specific health outcome.

3.3 Results

3.3.1 Descriptive statistics and transition matrices

Our sample consists of 22,528 patients treated in 230 NHS and private hospitals. The number of patients in each hospital ranges from 1 to 545 (median=70, interquartile range (IQR)=14-147). We present descriptive statistics of patient characteristics in Table 3.1.

Elective hip replacement surgery is performed predominantly on elderly patients

[45] Note that these credible intervals are only appropriate for single comparison against a given quantity, like the average, but are too wide for direct comparisons of specific hospitals (Goldstein and Healy 1995).

Table 3.1: Descriptive statistics of patient characteristics

Variable	Description	Mean / %	SD
male	=1, if patient is male	0.42	0.49
age	Patient's age in years	67.83	10.69
deprivation	Index of Multiple Deprivation, income domain	0.12	0.10
wcharlson	Weighted Charlson index of comorbidities	0.30	0.66
add_comorbidities	Number of additional non-Charlson comorbidities	1.97	1.96
osteoarthritis	=1, if main diagnosis is osteoarthritis (OA)	0.87	0.34
rheumatoid_arthritis	=1, if main diagnosis is rheumatoid arthritis (RA)	0.05	0.23
other_maindiag	=1, if main diagnosis is not OA or RA	0.00	0.07
revision_complications	=1, if revision surgery due to complications	0.01	0.09
revision_other	=1, if revision surgery due to other reasons	0.07	0.25
pretest	Time between pre-operative assessment and admission (in days)	18.26	23.70
posttest	Follow-up (in days)	207.54	29.82
N	Patients	22,528	
J	Hospitals	230	

(mean=67.8, SD=10.7), with osteoarthritis being the most common reason for surgical intervention. The majority of patients in our sample are female (58.3%) and admitted for primary replacement of the hip joint (92.6%). The median time elapsed between baseline survey and date of admission is 14 days (IQR=5-28). The median follow-up period is 197 days (IQR=192-211).

Table 3.2 presents the transition matrices for each of the five EQ-5D dimensions. Rows report the patients' own classification of their status at baseline and columns show self-reported status six months after surgery. Accordingly, patients in the lower triangle report improvements in health status, whereas those in the upper triangle report deteriorations.

For each of the five dimensions, a considerable number of patients report no problems at baseline. This is especially pronounced on the dimensions self-care and anxiety/depression where 44.5% and 57.8% of patients fall into this category. 6.6% of patients report no problems prior to treatment with respect to mobility, whereas nearly all patients report at least moderate problems with pain/discomfort (99.1%). 72 patients report to have no problems in any of the EQ-5D dimensions.[46]

The number of patients improving since treatment varies greatly by the health

[46] Of these, 21 (29.2%) underwent revision surgery. The remainder underwent primary surgery, typically for rheumatoid arthritis (n=46, 63.9%).

Table 3.2: Transition matrices for all EQ-5D dimensions

	post-treatment			
pre-treatment	no (=1)	some (=2)	extreme (=3)	Total
Mobility				
I have *no* problems in walking about (=1)	1,236	257	0	1,493
I have *some* problems in walking about (=2)	11,133	9,791	13	20,937
I am *confined to* bed (=3)	21	73	4	98
Total	12,390	10,121	17	22,528
Self-Care				
I have *no* problems with self-care (=1)	9,092	916	13	10,021
I have *some* problems with self-care (=2)	7,910	4,242	71	12,223
I am *unable to* wash or dress myself (=3)	79	155	50	284
Total	17,081	5,313	134	22,528
Usual Activities				
I have *no* problems with performing my usual activities (=1)	1,077	290	24	1,391
I have *some* problems with performing my usual activities (=2)	8,940	7,374	438	16,752
I am *unable to* perform my usual activities (=3)	1,413	2,427	545	4,385
Total	11,430	10,091	1,007	22,528
Pain / Discomfort				
I have *no* pain or discomfort (=1)	156	46	1	206
I have *some* pain or discomfort (=2)	7,609	5,123	254	12,986
I am *extreme* pain or discomfort (=3)	3,978	4,708	653	9,339
Total	11,746	9,877	908	22,528
Anxiety / Depression				
I am *not* anxious or depressed (=1)	12,017	949	57	13,023
I am *moderately* anxious or depressed (=2)	5,686	2,501	202	8,389
I am *extremely* anxious or depressed (=3)	489	465	162	1,116
Total	18,192	3,915	421	22,528

dimension under consideration. The dimension most improved since treatment is pain/discomfort, where 72.3% of the patients report improvements as indicated by a transition to a more favourable category. In contrast, only 29.5% of patients report improvements on the anxiety/depression dimension. This reflects the large proportion of patients reporting to be not anxious or depressed prior to treatment.

Figure 3.1 presents the empirical distribution of the EQ-5D utility scores pre- and post-intervention. The mean pre-intervention score is 0.353 and the mean post-operative score is 0.763. Both distributions exhibit typical characteristics of empirical EQ-5D distributions observed for a wide range of medical conditions, including multimodality, discontinuity, and clustering at 1 ('full health') (Basu and Manca 2012; Hernández Alava et al. 2012). 87.3% of patients report improvements in health as measured by the EQ-5D utility index, whereas 6.5% report deteriorations.

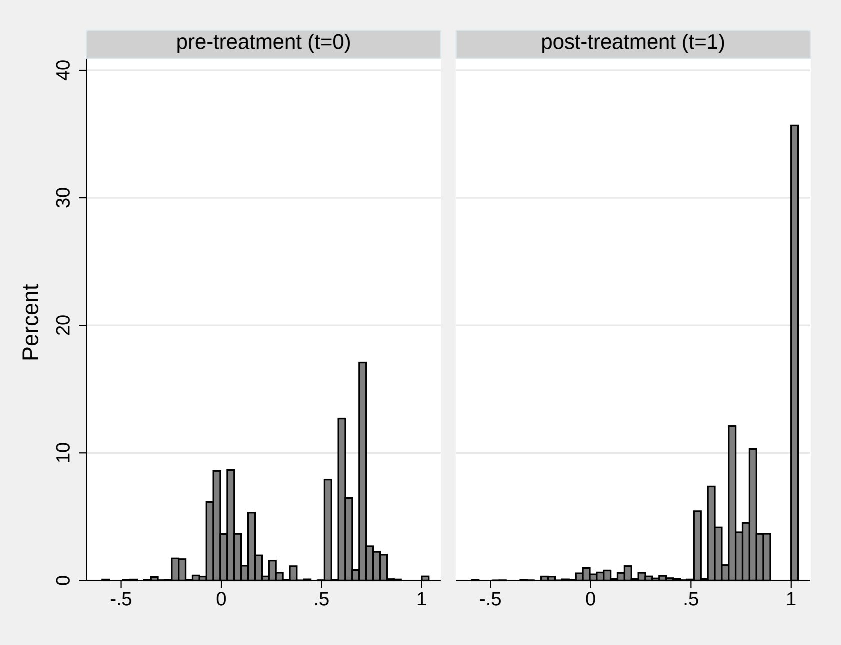


Figure 3.1: Distribution of EQ-5D utility scores pre- and post-treatment

3.3.2 Regression results

Table 3.3 presents parameter estimates and associated standard errors for each of the five dimension models and the EQ-5D utility index model.

Table 3.3: Regression results

	EQ-5D dimensions											
	Mobility		Self-Care		Usual Activities		Pain/Discomfort		Anxiety/Depression		EQ-5D utility index	
Variable	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
male	0.201	0.031 ***	0.013	0.024	0.101	0.020 ***	0.308	0.019 ***	0.469	0.025 ***	0.075	0.004 ***
age_15-60	0.079	0.040 *	-0.131	0.032 ***	-0.045	0.027	-0.080	0.025 **	-0.238	0.032 ***	-0.028	0.005 ***
age_71_80	-0.107	0.037 **	0.026	0.028	-0.039	0.023	0.016	0.022	0.030	0.029	0.001	0.005
age_80+	-0.266	0.060 ***	-0.317	0.043 ***	-0.311	0.036 ***	-0.077	0.034 *	-0.071	0.043	-0.043	0.007 ***
add_comorbidities	-0.081	0.009 ***	-0.073	0.006 ***	-0.048	0.005 ***	-0.058	0.005 ***	-0.066	0.006 ***	-0.018	0.001 ***
revision_complications	-0.053	0.171	-0.108	0.127	-0.106	0.105	0.054	0.100	-0.359	0.123 **	-0.021	0.021
revision_other	0.174	0.088 *	0.037	0.070	0.070	0.057	0.217	0.055 ***	-0.051	0.069	0.024	0.012 *
deprivation	-0.842	0.170 ***	-1.179	0.126 ***	-0.398	0.105 ***	-1.176	0.101 ***	-1.138	0.125 ***	-0.326	0.021 ***
wcharlson	-0.106	0.026 ***	-0.150	0.018 ***	-0.110	0.015 ***	-0.102	0.015 ***	-0.098	0.018 ***	-0.033	0.003 ***
rheumatoid_arthritis	-0.063	0.068	-0.072	0.054	-0.127	0.044 **	0.000	0.042	-0.056	0.053	-0.018	0.009
other_maindiag	-0.509	0.241 *	-0.780	0.166 ***	-0.396	0.137 **	-0.346	0.136 *	-0.268	0.163	-0.126	0.027 ***
pretest	0.000	0.000	0.002	0.000 ***	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000
treatment	2.697	0.093 ***	1.649	0.095 ***	2.239	0.077 ***	2.552	0.074 ***	1.654	0.097 ***	0.491	0.014 ***
treatment x male	0.075	0.035 *	0.039	0.030	0.182	0.025 ***	-0.068	0.024 **	-0.156	0.031 ***	-0.037	0.005 ***
treatment x age_15-60	-0.072	0.047	-0.047	0.041	-0.089	0.034 **	-0.006	0.033	-0.094	0.040 *	0.002	0.006
treatment x age_71_80	-0.017	0.042	-0.047	0.036	-0.144	0.030 ***	0.030	0.029	0.039	0.036	-0.003	0.006
treatment x age_80+	-0.240	0.067 ***	0.043	0.052	-0.210	0.045 ***	0.151	0.044 ***	0.093	0.053	0.009	0.008
treatment x add_comorbidities	-0.036	0.010 ***	-0.031	0.008 ***	-0.057	0.007 ***	-0.015	0.007 *	-0.035	0.008 ***	-0.000	0.001
treatment x revision_complicatio	-0.506	0.195 **	-0.736	0.147 ***	-0.509	0.132 ***	-0.573	0.129 ***	-0.407	0.138 **	-0.094	0.025 ***
treatment x revision_other	-0.797	0.101 ***	-0.587	0.082 ***	-0.607	0.072 ***	-0.594	0.070 ***	-0.312	0.082 ***	-0.113	0.013 ***
treatment x deprivation	-0.764	0.192 ***	-0.574	0.151 ***	-1.028	0.132 ***	-0.175	0.129	-0.653	0.149 ***	0.020	0.024
treatment x wcharlson	-0.111	0.029 ***	-0.077	0.022 ***	-0.063	0.019 ***	-0.023	0.019	-0.022	0.022	-0.000	0.004
treatment x rheumatoid_arthritis	0.034	0.078	0.019	0.064	0.048	0.056	-0.065	0.053	-0.005	0.064	0.002	0.010
treatment x other_maindiag	-0.521	0.274	-0.305	0.191	-0.341	0.172 *	-0.331	0.172	-0.072	0.194	-0.034	0.032
treatment x posttest	-0.002	0.000 ***	-0.001	0.000 **	-0.001	0.000 ***	-0.001	0.000 **	-0.002	0.000 ***	-0.000	0.000 ***
constant	constraint to 0		constraint to 0		constraint to 0		constraint to 0		constraint to 0		0.417	0.006 ***
κ_1	-3.492	0.071 ***	-3.469	0.052 ***	-1.141	0.029 ***	-0.351	0.027 ***	-2.540	0.041 ***	n/a	
κ_2	1.660	0.045 ***	-0.175	0.034 ***	1.494	0.030 ***	2.066	0.032 ***	-0.466	0.033 ***	n/a	
σ^2_ϵ	constraint to 1		constraint to 1		constraint to 1		constraint to 1		constraint to 1		0.057	0.001 ***
σ^2_α	0.515	0.036 ***	1.018	0.038 ***	0.442	0.019 ***	0.290	0.016 ***	1.129	0.039 ***	0.021	0.001 ***
τ^2_ζ	0.028	0.007 ***	0.033	0.007 ***	0.017	0.004 ***	0.020	0.004 ***	0.017	0.004 ***	0.002	0.000 ***
τ^2_θ	0.027	0.009 ***	0.007	0.004	0.027	0.006 ***	0.011	0.004 ***	0.010	0.005 **	0.001	0.000 ***
$cov(\zeta, \theta)$	-0.005	0.009	-0.002	0.004	0.002	0.003	-0.003	0.003	-0.004	0.003	-0.001	0.000 ***
ρ	0.029		0.018		0.031		0.019		0.010		0.010	
LogL	-20,329		-27,936		-33,627		-34,213		-29,062		-5,913	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Statistical significance of variance and covariance components is determined via Likelihood Ratio tests. SE of variance and covariance components provided for completeness.

$N = 22,528$; $J = 230$

We find several variables to be associated with self-reported health at baseline. These include male gender (+), higher weighted Charlson index score (-), number of additional comorbidities (-), and the deprivation profile of the patient's neighbourhood of residence (-). The mean effect of treatment on post-treatment latent health is positive and significant for all dimensions, resulting in substantial increases in the probability of reporting no problems after surgery (Table 3.4). The number of comorbidities and the indicators for revision surgery are negatively associated with the treatment effect, indicating that treatment is less beneficial for multimorbid or revision patients. Similarly, patients living in more deprived areas experience, on average, less improvement in latent health than those residing in less deprived areas. Longer follow-up is also associated with a smaller increase in post-operative latent health, albeit the effect being small. For example, for a patient of average characteristics, the probability of reporting no problems on anxiety/depression is estimated to reduce by 0.3% per additional week of follow-up. Post-operative EQ-5D utility scores are expected to reduce by 0.002 per additional week of follow-up.

Table 3.4: Predicted probabilities of reporting a given health status for a patient of average characteristics

	no (=1)			some (=2)			extreme (=3)		
	$t = 1$	$t = 0$	*change*	$t = 1$	$t = 0$	*change*	$t = 1$	$t = 0$	*change*
Mobility	0.543	0.026	0.517	0.457	0.974	-0.517	0.000	0.001	-0.001
Self-Care	0.838	0.412	0.426	0.162	0.587	-0.425	0.000	0.001	-0.001
Usual Activities	0.460	0.044	0.416	0.534	0.778	-0.244	0.006	0.178	-0.172
Pain/Discomfort	0.485	0.012	0.473	0.506	0.550	-0.044	0.009	0.438	-0.429
Anxiety/Depression	0.897	0.615	0.282	0.102	0.376	-0.274	0.000	0.009	-0.009

All variance components are statistically significant at the 95% confidence level as confirmed by likelihood ratio tests. The exception is the variance of the random effect on treatment for the dimension self-care ($p<0.054$). In contrast, only the covariance term in the EQ-5D utility model is statistically significant. About 1.0% (anxiety/depression) to 3.1% (usual activities) of the unexplained variation in latent health is estimated to be associated with the hospital itself.

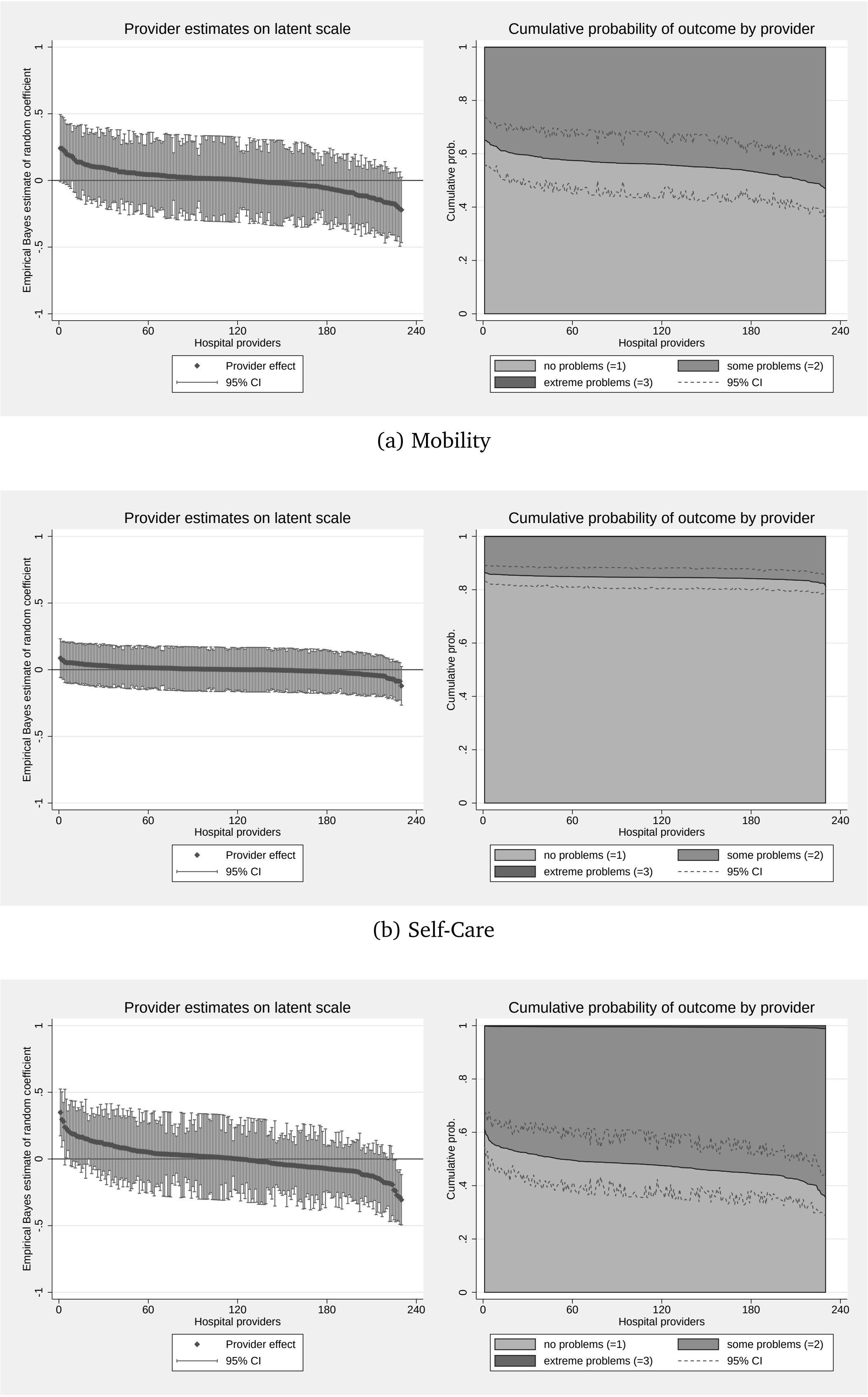
3.3.3 Assessment of hospital performance

3.3.3.1 Performance on individual EQ-5D dimensions and EQ-5D utility score

Figures 3.2a - 3.2e present estimates of hospital performance on the latent health scale (left graph) and the probability scale (right graph), where the latter is calculated for the average patient. Figure 3.2f presents the results of the EQ-5D utility model, where performance is measured directly on the utility scale. Hospitals located to the left side of each graph perform better than those to the right.

The random coefficient is standardised to zero which represents the expected outcome for a hospital with average case-mix. Hospital performance heterogeneity, as represented by the slope of the curve, is most pronounced on the mobility and usual activities dimensions. For the vast majority of hospitals, credible intervals contain zero but a small number of hospitals have a statistically significantly different treatment impact. Credible intervals on the mobility dimension are wider than on any other dimension. This reflects the lesser amount of information contained in the data, with only two outcome categories being reasonably well populated.

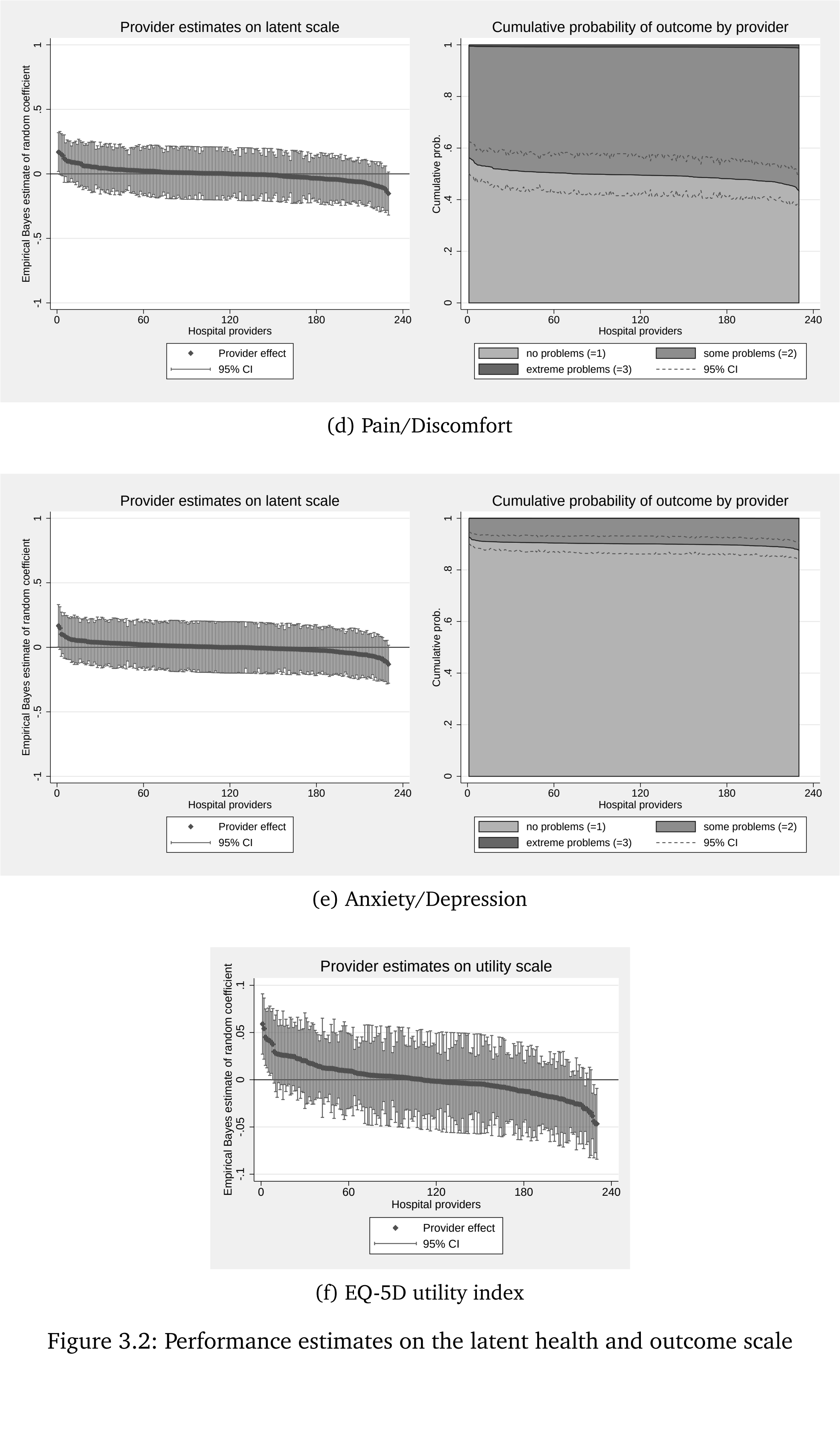
Hospital heterogeneity on the latent health scale translates into differences with respect to hospital-specific probabilities of reporting a given post-treatment health status (see also Table 3.5). The expected probabilities of reporting no problems on the usual activities dimension six month after surgery range from 35.6% to 61.3% (calculated for the average patient). In contrast, expected probabilities for the same outcome on the self-care dimension are significantly less dispersed and consistently above 80% for all hospitals. Performance variation is most pronounced on the dimensions mobility and usual activities, with gaps between best and worst performing hospital of 18.1% and 25.7%, respectively. The probability of reporting extreme problems after surgery is close to zero for all models. We refrain from reporting credible intervals around these predicted probabilities in Figures 3.2a - 3.2e to improve the readability of the graphs.



(a) Mobility

(b) Self-Care

(c) Usual Activities



(d) Pain/Discomfort

(e) Anxiety/Depression

(f) EQ-5D utility index

Figure 3.2: Performance estimates on the latent health and outcome scale

Table 3.5: Differences between providers in terms of the probability of reporting no problems post-operatively

EQ-5D dimension	Min	Max	Range	central 95%	IQR
Mobility	0.470	0.650	0.181	0.145	0.035
Self-Care	0.815	0.865	0.050	0.031	0.006
Usual Activities	0.356	0.613	0.257	0.178	0.047
Pain/Discomfort	0.435	0.562	0.128	0.086	0.020
Anxiety/Depression	0.876	0.927	0.051	0.030	0.007

Note: Calculated for a patient of average characteristics. The column 'central 95%' gives the differences between the 2.5th and 97.5th percentile of the distribution of hospitals.

3.3.3.2 Association of performance estimates on EQ-5D dimensions and the EQ-5D utility index

We explore the global agreement between estimates of hospital performance based on individual EQ-5D dimensions and the utility weighted EQ-5D index values by calculating Spearman's rank correlation coefficients (Spearman's ρ) and inspecting correlation patterns visually (Figure 3.3).[47]

The highest rank correlation is observed between performance estimates on the pain/discomfort dimension and EQ-5D utility index (ρ=0.333), followed by the anxiety/depression dimension (ρ=0.263). The rank correlation for all other dimensions and the EQ-5D utility index is smaller (ρ <0.2) and, indeed, not statistically significantly different from zero.

To explore whether judgement about individual providers would differ depending on which metric is used to assess performance, we identify providers with statistically significantly above/below average performance on each metric (Thomas et al. 1994; Laudicella et al. 2010; Racz and Sedransk 2010) and compare the overlap. In 16 out of 230 cases, performance classifications differ across metrics (Table 3.6).

[47] Correlations between performance estimates on individual EQ-5D dimensions are reported in Appendix Table A3.1.

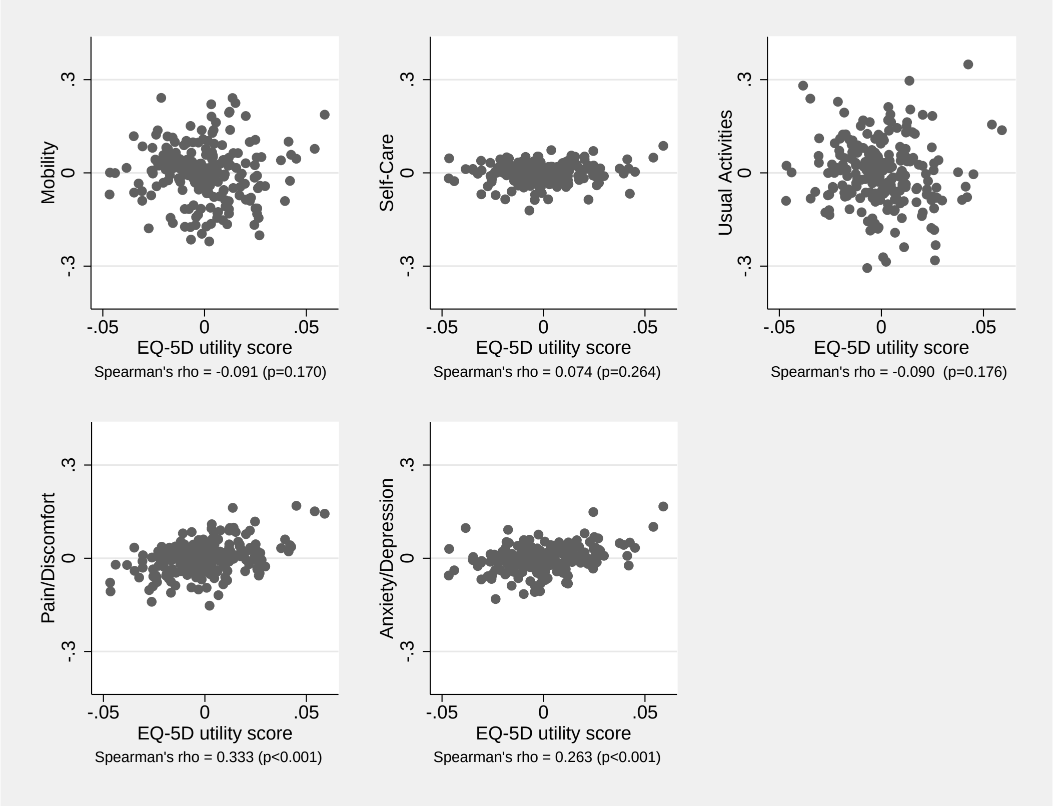


Figure 3.3: Hospital performance estimates on EQ-5D dimensions and EQ-5D utility scores

Table 3.6: Examples of hospitals for which performance assessments differ across EQ-5D dimensions and the EQ-5D utility model

Hospital	EQ-5D utilities	Mobility	Self Care	Usual activity	Pain / Discomfort	Anxiety / Depression
A	above	-	-	above	-	-
B	above	-	-	-	above	-
C	above	-	-	-	-	above
D	-	above	-	above	-	-
E	-	-	-	above	-	-
F	-	-	-	above	-	-
G	-	-	-	above	-	-
H	-	-	-	above	-	-
I	-	-	-	below	-	-
J	-	-	-	below	-	-
K	-	-	-	below	-	-
L	-	-	-	below	-	-
M	-	-	-	below	-	-
N	-	-	-	below	-	-
O	-	-	-	below	-	-
P	below	-	-	above	-	-

Note: Hospitals are either statistically *above* or *below* the average or not different from the average (-). Not adjusted for multiple testing.

Three hospitals (A-C) are identified as above average performers according to the EQ-5D utility model and one other EQ-5D dimension but do not stand out with respect to the other four dimensions. Five hospitals (D-H) achieve above average results with respect to at least one dimensions of the EQ-5D but this performance is not reflected in their performance estimate on aggregate utilities. Seven hospitals (I-O) fall short of the average benchmark on the dimension usual activities but would not be identified as underperformers in terms of their impact on utilities. The disagreement between performance in terms of EQ-5D utilities and individual dimensions is most apparent in the case of hospital P, where the hospital is classified as a low performer in terms of its impact on utilities but is a high performer with respect to restoring its patients' ability to carry out their usual activities.

3.4 Discussion

We set out an analytical strategy to explore patient-level and hospital-level variation in categorical responses within and across dimensions of the EQ-5D. This approach does not require assumptions about how to aggregate across health dimensions and offers insight about which dimensions are particularly affected by hospital heterogeneity. We find heterogeneity in performance to be more pronounced across the mobility and usual activities dimensions and less so for the pain/discomfort, anxiety/depression and self-care dimensions. Furthermore, we find that performance on the utility scale correlates well only with the dimensions anxiety/depression and pain/discomfort. Incidentally, these are the dimensions that receive the highest weighting in the UK TTO EQ-5D tariff (Dolan 1997). In contrast, the dimensions mobility, usual activities, and self-care have relatively low weights attached to them and performance heterogeneity remains undetected when analysing aggregated EQ-5D utility data. These findings re-emphasise the need to consider carefully the role that value sets play in hospital performance estimates based on aggregate utility scores. However, we note that our results are based on analysis of data for one specific condition and instrument and the influence of value sets may be more or less pronounced in other settings.

Policy-makers are interested in assessing the change in patient-reported outcomes as a result of treatment. There are various ways that this change can be measured and modelled. Our approach has been to model both pre- and post-treatment health status as outcomes of the same reporting process and to conduct multilevel analysis with measurement points clustered in patients, which themselves are nested in hospitals (see also Bryk and Raudenbush 1988). We argue that this is the appropriate modelling strategy because it acknowledges the features of the data generating process, allows for patient heterogeneity with respect to observed and unobserved factors and makes best use of the available information. The presented methodology is readily applicable to other conditions for which EQ-5D data are collected and, in principle, can be extended to other PRO instruments.

In recognition of the expectation that health outcome data are to be used by an audience unfamiliar with the interpretation of complex statistical results (e.g. patients and their relatives, family doctors, managers), we have suggested an intuitively appealing way of summarising the differential impact that hospitals have on treatment outcomes. Our graphical representation indicates the probability of reporting a given health outcome, and shows how these probabilities vary across health dimensions and hospitals. Prospective patients (or their agents) who place greater weight on a particular dimension may use this information to select a hospital that has a differentially greater impact on this than its peers do.

The primary limitation of our proposed approach is the increase in dimensionality of the decision problem for patients. Whereas aggregated scores result in one estimate of hospital performance, our approach generates five, potentially divergent, answers. In a recent study, Dijs-Elsinga et al. (2010) have shown that a large group of patients favour simple data presentation and prefer one overall measures of hospital quality.[48] But many patients intend to use more detailed quality information

[48] The phenomenon of ‘information overload’ is well-established in decision theory and refers to difficulties in collating, triangulating and interpreting a large amount of information (e.g. Keller and Staelin 1987). This may lead to a number of biases, including ‘status quo bias’ (Samuelson and Zeckhauser 1988) with patients ignoring information about poor performance and going to

when making decisions about where to seek care in the future (Dijs-Elsinga et al. 2010). The question then arises how much information should be provided for the different objectives for which performance information can be used (i.e. patient choice, accountability, identification of best practice) and who decides about the relative weighting of each component and objective (Parkin et al. 2010; Steyerberg and Lingsma 2010). Our study does not intend to resolve this debate. Rather, we present a means of making inferences about hospital quality and presenting results when health outcomes are assessed through the EQ-5D PRO instrument. How best to communicate such performance data requires careful consideration, to ensure it can be effectively understood and used.

Several issues remain that we have not addressed in this study. First, based on the full information contained in HES, we can identify those patients that have not participated or were not included in the follow-up. We find that, in our dataset, only about 50% of eligible hip replacement patients participate in the baseline survey, with a further 8% dropping out of the subsequent survey. These numbers should improve in time when data collection procedures become more established. However, falsely assuming that any substantial amount of missing values are generated at random could lead to biased inferences from a non-representative population (Little and Rubin 1987), raising questions about the validity of the assessment.

Second, in this study we have controlled for patient risk-factors that are deemed clinically relevant, assumed to be exogenous to the hospital, and can be derived from routine inpatient records. However we do not claim that this set of control variables is exhaustive: health outcomes may be affected by non-randomly distributed, unobserved patient characteristics such as severity of the medical condition or health-related behaviour. That said, a strength of our study is that we control for the initial health status with which the patient presents at admission. In many studies this is unobserved, and makes our analysis more robust than possible in the absence

their local hospital by default even though other providers would have been preferable given their preferences. In a study of switching behaviour in the mobile telecommunication market, Jilke (2015) found evidence that people with low educational attainment are especially prone to such biases.

of such information.

Third, we do not control for characteristics of the hospital in our analysis, our rationale being that these are within the hospital's control.[49] But they may not be. Hospitals may be constrained in their ability to choose and combine medical resources to their best effect by local regulation, access to factor markets or, in the short-run, the existing capital structure such as age and functionality and whether the hospital operates the service over multiple sites (Street et al. 2010). In this case, the assumption of exchangeability underlying the hierarchical modelling approach may not hold. Furthermore, procedures such as hip replacement are generally followed by extensive physiotherapy, which may be delivered outside the hospital. If constraints bind or if quality is not attributable solely to the hospital, our estimates of hospital performance may be biased.

Fourth, our study makes use of a large administrative dataset that contains rich information on patient characteristics and the type of care provided. The presented econometric approach is tailored to the data at hand. However, in other countries or disease areas, sample sizes may be smaller or information may be sparse. If patient characteristics are unobserved or cannot be included due to low degrees of freedom, then more of the time-invariant variation between patients would be captured by the patient random effect. Again, the assumption of exchangeability, i.e. that the unobserved patient heterogeneity is drawn from a random distribution, may become unrealistic and results may be biased (Hausman 1978). The same argument applies to the random coefficient and the interactions of covariates with the treatment effect. Researchers will need to consider this limitation case-by-case, based on their data and the available set of risk-adjustment variables.

Fifth, our econometric model could be extended in a number of ways. For example,

[49] Another implication of this is that our estimates of provider performance are '*Type A*' effects in the terminology of Raudenbush and Willms (1995); see also Goldstein and Spiegelhalter (1996). Such *Type A* effects are appropriate for patients selecting providers that are most likely to improve their health, independent of whether this is due to the providers' quality efforts or favourable production environments. In contrast, regulators seek estimates of providers' efforts net of the influence of binding constrains imposed by the production environment; so-called type *Type B* effects. See Chapter 2 for an example of such *Type B* performance estimates.

one limitation of the ordered probit model is the assumption of 'proportional odds', i.e. a change in covariate has a proportional effect on all outcome probabilities since the associated coefficient does not vary across outcome categories. In order to relax this assumption one could replace the ordered probit model with an unordered multinomial logit model; see Gray et al. (2006) for an application in the context of mapping between PROMs. Other extensions could seek to model cut points more flexibly as functions of observed parameters (Greene and Hensher 2010) or model the full joint multivariate distribution of outcomes. Also, while in linear models modelling changes in responses categories is equivalent to modelling post-operative responses conditional on pre-operative responses (Nuttall et al. 2015), this may not extend to the non-linear models employed here and one could explore the sensitivity of our results to this modelling choice. For example, one could model variation in the probability of improvement on individual dimensions, rather than derive this in post-estimation as we have done here.

Finally, further consideration should be given to the role that patient-reported health outcome performance information can play in existing quality assessment frameworks. While measures of risk-adjusted mortality, re-admission and adverse events have been criticised for their limited granularity and sensitivity (Lilford and Pronovost 2010), one should not a-priori dismiss their ability to identify high and low quality providers of care. Further research is required to establish the additional value of patient-reported outcome data for hospital quality assessments (see also Chapter 4 and 5) and contrast it to the costs of collection.

Acknowledgements

We are grateful for comments and suggestions from Stephen Barasi, Stephen Bloomer, David Nuttall, David Parkin, Aurore Pelissier, Wolfgang Greiner, three anonymous referees and those received during presentations at the University of Manchester, the Health Econometric Data Group seminar series (York), the EuroQoL plenary meeting 2012 (Rotterdam), the European Conference on Health Economics (Zurich) and the joint CES-HESG Winter conference 2012 (Marseille). The project was funded by the National Institute for Health Research (NIHR) in England under the Health Services Research (HSR) stream (project number 09/2000/47). The views expressed are those of the authors and may not reflect those of the NIHR HSR program or the Department of Health.

4 Multidimensional performance assessment of healthcare providers using dominance criteria

4.1 Introduction

Variation in healthcare quality and costs are well documented (Wennberg and Gittelsohn 1973; Keeler 1990; Busse et al. 2008; Bernal-Delgado et al. 2015) and may arise when providers enjoy discretion over how their services are organised and provided (Arrow 1963). Regulators, who are charged with overseeing the provision of care, are concerned about variation if it is not caused by differences in healthcare needs or patient preferences as it may signal inequity, inefficiency or unsafe care. To address this, many healthcare systems have implemented routine benchmarking (or 'profiling') of healthcare providers to identify comparative performance levels. This might help single out 'positive deviants' (Bradley et al. 2009; Berwick 2008; Lawton et al. 2014), or exemplars of best practice, that can be studied further or rewarded as part of a pay-for-performance scheme. At the other extreme, poor performers might be subject to penalties for falling short of their peers or to interventional actions by regulators.

Healthcare providers share two important features with other public sector organisations that complicate the assessment of their performance (Dixit 2002; Besley and Ghatak 2003; Propper and Wilson 2012). First, they lack a single overarching objective, such as profit, against which their performance can be assessed. Instead, they pursue multiple, sometimes conflicting, objectives and this requires the regulator to measure and incentivise achievements along a range of performance dimensions. These achievements are typically non-commensurate and include different aspects of performance reflecting resource use, clinical effectiveness, and other dimensions of quality such as accessibility (Smith 2002; Goddard and Jacobs 2009; Porter 2010; Devlin and Sussex 2011). Second, providers typically serve several stakeholders (e.g. patients, purchasers of care, and politicians) and the values these stakeholders attach to objectives are often not known to the regulator[50], but are unlikely to be identical (Smith 2002; Propper and Wilson 2012); see Devlin and Sussex (2011) for examples from healthcare and the wider public sector.

The lack of a set of common, explicit valuations for individual performance dimensions makes it difficult to construct a single, unidimensional performance measure. If valuations were known and common across stakeholders, it would be possible to aggregate multiple performance scores into unidimensional composite scores. Such measures are attractive as they allow a complete and transitive ranking of providers, facilitate the presentation and dissemination of performance information to stakeholders, and offer a simple means to adjust rewards in a pay-for-performance framework (Dowd et al. 2014). But without such knowledge, there is no guidance on how to aggregate achievements appropriately.

The empirical literature has addressed this problem in different ways: Some studies restricted their assessment of provider performance to those performance dimensions for which explicit valuations have been expressed. Examples include Timbie et al. (2008), Timbie and Normand (2008) and Karnon et al. (2013), all of which translate hospital mortality estimates into monetary units using the expressed valuation of a statistical life. The obvious shortcoming of this approach is that performance dimensions which lack explicit valuations (e.g. waiting times, patient satisfaction, or emergency re-admission rates[51]) are necessarily omitted from the

[50] It may be possible to estimate the preferences of individual stakeholders or groups thereof by means of elicitation or through the study of revealed preferences (Ryan et al. 2001). However, this would likely be a very difficult and costly undertaking and is therefore rarely done in practice.

[51] It may be possible to translate achievements on some objectives, e.g. emergency readmission rates or other measures of health outcomes, into quality-adjusted life years (QALYs) by means of modelling (Timbie et al. 2009; Appleby et al. 2013; Coronini-Cronberg et al. 2013). A monetary valuation of QALYs has been expressed in the English NHS and elsewhere. However, the data requirements are substantial and the statistical uncertainty introduced through modelling is likely to further compound the problem of differentiating between true performance signal and noise.

analysis. Their omission may lead to tunnel vision, whereby providers concentrate their efforts on performance dimensions with explicit valuations at the expense of other dimensions (Holmström and Milgrom 1991; Goddard et al. 2000).

Alternatively, analysts often either choose a set of weights, implement pre-defined scoring algorithms such as equal weighting, or derive weights from the data using approaches based on item response theory (Landrum et al. 2000; Landrum et al. 2003; Daniels and Normand 2006; Teixeira-Pinto and Normand 2008), data envelopment analysis (Dowd et al. 2014), and more ad-hoc econometric specifications (Chua et al. 2010). However, such practice conflicts with one of the key tenets of economic welfare theory, namely that the stakeholders are the only legitimate judges of their own preferences and that, ultimately, responsibility for specifying valuations for performance dimensions should rest with the relevant stakeholders (Smith and Street 2005). There is no guarantee that weights imposed by analysts, however these are arrived at, match the preferences of all stakeholders. Consequently, organisations being assessed might legitimately question the validity of the generated index.

There is an alternative way to address the problem of determining appropriate weights. Multidimensional performance assessment circumvents the issue by analysing performance against each achievement individually and then combining the results into an overall performance profile. In doing so, it makes explicit how healthcare providers perform on each performance dimension and how these dimensions correlate. The multidimensional approach has enjoyed increasing popularity in the health economic literature: Hall and Hamilton (2004) assess the performance of surgeons in terms of 30-day mortality and morbidity using a Bayesian hierarchical bivariate probit model. Hauck and Street (2006) use multivariate multilevel models to study the performance of health authorities across 13 performance indicators. Gutacker et al. (2013) study hospital performance with respect to five health dimensions and compare their results to those based on an composite measure. Portrait et al. (2015) compare Dutch Diabetes care groups in terms of costs and a broad range of quality indicators, whereas Häkkinen et al. (2014), Kruse and Christensen (2013) and Street et al. (2014) study the performance of hospitals in terms of costs and a single measure of patient health outcome for different conditions.

But multidimensional performance assessment is not a panacea for the problem of judging performance across multiple objectives. A multidimensional performance profile does not permit ranking of hospitals or comparison to some performance standard. Hence it remains unclear which providers excel or perform poorly across multiple performance dimensions. This constitutes a major limitation of the multidimensional approach for practical purposes, and one that we seek to overcome in this study. More specifically, we propose the use of dominance criteria to judge hospital performance against a multidimensional benchmark. The concept of dominance has the attractive feature that it allows comparison of multidimensional performance profiles against benchmarks under relatively weak assumptions about stakeholders' utility functions. Indeed, the only requirement is that the regulator can judge whether the marginal utility of an achievement is positive or negative and that this qualitative judgement applies to all stakeholders. We believe this to be a reasonable pre-requisite in most contexts.

We apply our approach to data on providers of hip replacement surgery in the English NHS during the period April 2009 to March 2012. Performance is assessed along four risk-adjusted performance metrics: inpatient length of stay ('efficiency'), waiting times ('access to care'), 28-day readmission rates and improvements in patient-reported health status after surgery (both 'clinical quality'). Each of these metrics has been the focus of recent health policy in England (Department of Health 2008a; Department of Health 2012b; Propper et al. 2008; Siciliani et al. 2014) and have been widely used in the academic literature to measure performance differences and changes therein over time (e.g. Jensen et al. 2009; Siciliani et al. 2013).[52] We estimate multivariate multilevel models to account for the clustering of patients in providers and exploit the correlation of provider achievements across

[52] Note that these metrics are not without criticism. For example, like mortality, emergency readmissions may not always be avoidable (Fischer et al. 2014). Hence, performance indicators based upon them may be noisy. Also, short length of stay, which would be interpreted as an indicator of efficient discharge management, may actually be harmful if patients are discharged prematurely (Qian et al. 2011). This may in turn increase emergency readmissions (Carey 2015). These limitations highlight the need to analyse and interpret performance estimates jointly.

dimensions (Zellner 1962; Hauck and Street 2006). Empirical Bayes estimates of the provider-specific posterior means and variance-covariance matrices are used to classify hospitals into three categories: dominant, dominated, and non-comparable. We quantify the uncertainty surrounding this classification in the form of Bayesian probability statements.

The study is the first to apply dominance criteria to multidimensional performance assessment of healthcare providers and derive appropriate confidence statements. Besides this, we make three further contributions to the empirical literature on hospital performance. First, we provide evidence about the correlations, and thus the potential for trade-offs, between a number of objectives that healthcare providers typically face. Previous research has focused predominantely on the association between hospital costs and mortality (see Hussey et al. (2013) for a review), largely ignoring other important dimensions such as waiting times or health-related quality of life. Second, in contrast to previous studies conducted at hospital level (e.g. Martin and Smith 2005), we focus on a single homogeneous patient population, thereby reducing the risk of ecological fallacy. Third, by exploiting novel data on pre-operative health status in addition to the co-morbidity markers that are usually available in administrative records, we are better able to isolate from case-mix differences the true impact that providers have on performance measures ('value added').

The remainder of this chapter is structured as follows: In section 4.2 we set out the assessment framework in conceptual terms. Section 4.3 presents the empirical methodology and section 4.4 describes our data. We report results in section 4.5 and offer concluding comments in section 4.6.

4.2 Multivariate performance assessment using dominance criteria

Assume that a regulator, acting on behalf of stakeholders, seeks to determine the overall performance of a number of hospital providers. Let there be $k = 1, \dots, K$ performance dimensions with observed achievement Y_k. Each achievement is determined by two factors, namely factors under the control of the provider θ_k and external production constraints X_k, so that

$$Y_k = f(X_k, \theta_k) \tag{4.1}$$

for each provider.

The parameter θ_k can be interpreted as the provider's contribution to achievement k over and above the circumstances in which they operate. This parameter is generally not directly observable and thus forms the target for inferences about performance. In order to isolate θ_k from X_k, the regulator must establish the contribution of production constraints to observed achievement by means of comparison with other providers, i.e. through risk-adjustment as applied in yardstick competition (Shleifer 1985).

Stakeholders derive utility from the providers' performance on each dimension, so that $U = U(\theta_1, \ldots, \theta_K)$, which is assumed to be monotonic in θ_k over the range of realistic values for all $k \in K$. The regulator has only limited knowledge about the characteristics of this utility function. This may be because there are multiple stakeholders with heterogeneous and/or unknown preferences. More specifically, the regulator has no information about the marginal utility $\partial U / \partial \theta_k$ that each stakeholder derives from achievements on each performance dimension, and hence the marginal rate of substitution (MRS) at which each stakeholder is willing to trade off performance on one dimension against that on another, i.e. $\partial \theta_k / \partial \theta_{k'}$ for $k \neq k^{'}$. However, the regulator has knowledge about the sign of $\partial U / \partial \theta_k$, i.e. whether achievements are expressed positively or negatively. To simplify the exposition, we assume from now on that achievements can be expressed so that utility increases in θ_k.

If only one performance dimension is assessed $(K = 1)$ or the MRS across multiple dimensions are known then achievements can be expressed as unidimensional (composite) scores. The regulator can then conduct either a *relative* or *absolute* assessment of performance. The first involves ranking the providers $j \in J$ according to their adjusted (composite) achievement θ_j, where $\theta_j > \theta_{j'}$ implies $U(\theta_j) > U(\theta_{j'})$ for $j \neq j'$. This will result in a complete and transitive ordering of providers, assuming no ties. One can then designate a specific number of providers as performing well or poorly based on their relative ranking, e.g. whether they fall within a given percentile of the distribution. Goldstein and Spiegelhalter (1996) provide a discussion of the statistical challenges associated with this approach. Alternatively, providers' performances can be classified based on $\theta_j - \theta^*$ being larger or smaller than zero, where θ^* denotes an absolute performance standard to which providers are compared.[53] The latter is often employed in the context of quality performance assessment, e.g. with respect to standardised mortality after surgery (Spiegelhalter 2005; National Clinical Audit Advisory Group 2011).

When multiple performance dimensions are assessed ($K \geq 2$) and the MRS is unknown, a complete and transitive ordering of providers is no longer guaranteed and relative assessments are unfeasible. As a result, it becomes impossible to identify providers that perform well or poorly in terms of stakeholders' aggregate utility. This is a well-known problem in the field of welfare economics and consumer theory (Boadway and Bruce 1984; McGuire 2001). However, some combinations of performance levels may be strictly preferable (dominant) or inferior (dominated) to other combinations, leading to a partial ordering of providers. As an analogue to the Pareto dominance criteria we can formalise the following general dominance classification rules[54]:

A provider either

1. *dominates* the comparator if $\theta_{jk} \geq \theta_{j'k}$ for all $k \in K$ and $\theta_{jk} > \theta_{j'k}$ for some $k \in K$, or
2. *is dominated* by the comparator if $\theta_{jk} \leq \theta_{j'k}$ for all $k \in K$ and $\theta_{jk} < \theta_{j'k}$ for

[53] Note that, when no external standards are specified, performance standards are typically based on the performance of all organisations, i.e. an internal performance standard (Shleifer 1985; National Clinical Audit Advisory Group 2011). Hence, a provider will be considered to perform well when the observed achievement is better than a reference value derived from all providers. In many cases, this reference value is simply the average across all providers, i.e. $\theta^* = \frac{1}{J}\sum \theta_j$.

[54] Devlin et al. (2010) propose the use of a similar classification system to compare EQ-5D health profiles over time without resorting to making strong assumptions about patients' preferences.

some $k \in K$, or

3. is *non-comparable* to the comparator if $\theta_{jk} \geq \theta_{j'k}$ for some $k \in K$ and $\theta_{jk} \leq \theta_{j'k}$ for the remaining $k \in K$,

where $j \neq j'$ and $\theta_{j'k}$ denotes the performance level of the comparator, which may be either another provider or an absolute internal or external performance standard θ^*.

4.3 Methodology

4.3.1 Empirical approach

The aims of the empirical analysis are to obtain estimates of providers' performance θ_{jk} and of the correlation of θ_{jk} across each of the $K = 1, \ldots, 4$ performance dimensions, and to classify providers according to the dominance classification set out in section 4.2. We estimate multivariate multilevel models (MVMLMs) with achievement score Y_{ijk} observed for patients $i = 1, \ldots, n_j$ who are clustered in hospitals $j = 1, \ldots, J$. Multilevel (i.e. random intercept) models have become a staple tool in the field of performance assessment and allow us to i) adjust achievements for differences in case-mix across providers, ii) decompose unexplained variation in achievement into random (within-provider) variation at patient level and systematic (between-provider) variation at provider level, and iii) obtain more reliable (precision-weighted or shrunken) estimates of performance (Goldstein 1997; Normand et al. 1997; Ash et al. 2012).

The multivariate nature of the data is taken into account through correlated random terms at each level of the hierarchy. These random terms are assumed to be drawn from multivariate normal distributions (MVN) with unconstrained variance-covariance matrices (Zellner 1962; Hauck and Street 2006). Allowing for correlation across achievements is beneficial for several reasons. First, we can construct multivariate hypothesis tests of parameters of interest that take into account the correlation between dimensions and achieve correct coverage probabilities. We discuss this in detail below. Second, we can achieve efficiency gains and obtain more precise estimates of relevant parameters if either the components of X_{ijk} differ across k or non-identity link functions are employed for at least some of the regression equations (Zellner 1962; Thum 1997; Bailey and Hewson 2004). Finally, by utilising a maximum likelihood estimator, data about achievements that are missing for any particular performance domain can be assumed missing at random conditional on all modelled covariates *and* achievements (Little and Rubin 1987; Goldstein 1986).

Hospital achievements are measured using two continuous and two binary variables. In order to ascertain the conditional normality of error terms as imposed by the MVN assumption[55], we apply appropriate transformations (e.g. logarithmic) for the continuous achievement variables and specify probit models for the binary achievement variables. The latter can be motivated by considering each binary achievement variable as the observed realisation of a latent truncated Gaussian variable.

The empirical model to be estimated is specified as

$$Y^*_{ijk} = \alpha_k + X^{'}_{ijk}\beta_k + \theta_{jk} + \epsilon_{ijk} \tag{4.2}$$

with $Y^*_{ijk} = f(Y_{ijk})$ for $k = 1, 2$ and

$$Y_{ijk} = \begin{cases} 1 \text{ if } Y^*_{ijk} > 0 \\ 0 \text{ if } Y^*_{ijk} \leq 0 \end{cases}$$

for $k = 3, 4$.

The variable Y_{ijk} denotes the observed outcome, Y^*_{ijk} is the corresponding latent underlying variable, $f(.)$ is a transformation function chosen to normalise the conditional distribution of ϵ_{ijk}, X_{ijk} is a vector of explanatory variables whose components may differ across dimensions, α_k is an intercept term, θ_{jk} denotes a random effect at provider level and ϵ_{ijk} denotes the random error term at patient level. Both random

[55] In principle it is possible to use other multivariate distributions such as multivariate gamma. However, such models are not typically implemented in standard statistical software packages and are therefore rarely used in practice.

terms are assumed to be MVN distributed with mean vector zero and a $K \times K$ variance-covariance matrix, so that $\theta_{jk} \sim MVN(0, \Sigma)$ with

$$\begin{aligned} E(\theta_{jk}) &= 0 \\ var(\theta_{jk}) &= \tau_k^2 \\ cov(\theta_{jk}, \theta_{jk'}) &= \rho_\theta \tau_k \tau_{k'} \end{aligned}$$

for all $k \neq k^{'}$, and similarly $\epsilon_{ijk} \sim MVN(0, \Omega)$ with

$$\begin{aligned} E(\epsilon_{ijk}) &= 0 \\ var(\epsilon_{ijk}) &= \sigma_k^2 \text{ for } k = 1, 2 \\ var(\epsilon_{ijk}) &= 1 \text{ for } k = 3, 4 \\ cov(\epsilon_{ijk}, \epsilon_{ijk'}) &= \rho_\epsilon \sigma_k \sigma_{k'} \end{aligned}$$

for all $k \neq k^{'}$. The model reduces to a set of univariate models if all off-diagonal elements of Σ and Ω are zero, i.e. achievements are uncorrelated conditional on observed patient factors.

Estimation was performed in MLwiN 2.32 called from within Stata 13 using the `runmlwin` programme (Leckie and Charlton 2013).

4.3.2 Classification of provider effects and multivariate hypothesis tests

We compare providers against a common absolute performance standard, here defined as the expected performance of a (hypothetical) hospital of average performance α_k, i.e. the conditional mean. We base our assessment of provider performance on estimates of θ_{jk}, which represent the provider-specific deviation from this benchmark. These parameters are not directly estimated in a random effects framework but can be recovered in post-estimation using Empirical Bayes predictions techniques (Skrondal and Rabe-Hesketh 2009). We stack performance estimates into vector coordinates to denote the provider's location in the k-dimensional performance space with the origin being normalised to zero. A provider's dominance classification is then determined by comparing its estimated adjusted achievements to that of the performance standard across all four dimensions simultaneously. This leads to three possible classifications: dominant, dominated, or non-comparable.

In order to quantify the uncertainty around these possible classifications we take a Bayesian perspective and calculate the posterior probability that a given provider truly dominates [is dominated by; non-comparable to] the multidimensional performance standard. This involves calculating the area under the MVN probability density function that covers each of the three possibilities, for each provider.[56] Figure 4.1a illustrates this for the two-dimensional case with two highly correlated bivariate normal distributed achievements ($\rho = 0.6$). The centroid of the density is given by X and the ellipse shows the central 95% of this density. The density is dissected by two lines which intersect at the benchmark. The density covered by the areas A and B equal the probability of *dominating* or *being dominated by* the benchmark, whereas the density covered by area C gives the probability for the *non-comparable* outcome. To calculate these probabilities, we follow the simulation approach of O'Hagan et al. (2000). Our simulation involves drawing S repeated samples from the MVN posterior distribution of the provider-specific Empirical Bayes estimates of the mean vector θ_j and associated variance-covariance matrix Σ_j. We then apply the dominance criteria to each simulation and calculate posterior probabilities by averaging across simulations. Formally,

$$Pr(\text{dominant } | J = j) = \frac{1}{S} \sum_{s=1}^{S} \prod_{k=1}^{4} I(\theta_{jk}^{s} > 0) \tag{4.3}$$

$$Pr(\text{dominated } | J = j) = \frac{1}{S} \sum_{s=1}^{S} \prod_{k=1}^{4} I(\theta_{jk}^{s} < 0) \tag{4.4}$$

[56] Our problem is similar to that encountered in the context of cost-effectiveness analysis, where one wishes to compute the probability that a new treatment is cost-effective for a given level of willingness to pay (Van Hout et al. 1994; Briggs and Fenn 1998; O'Hagan et al. 2000).

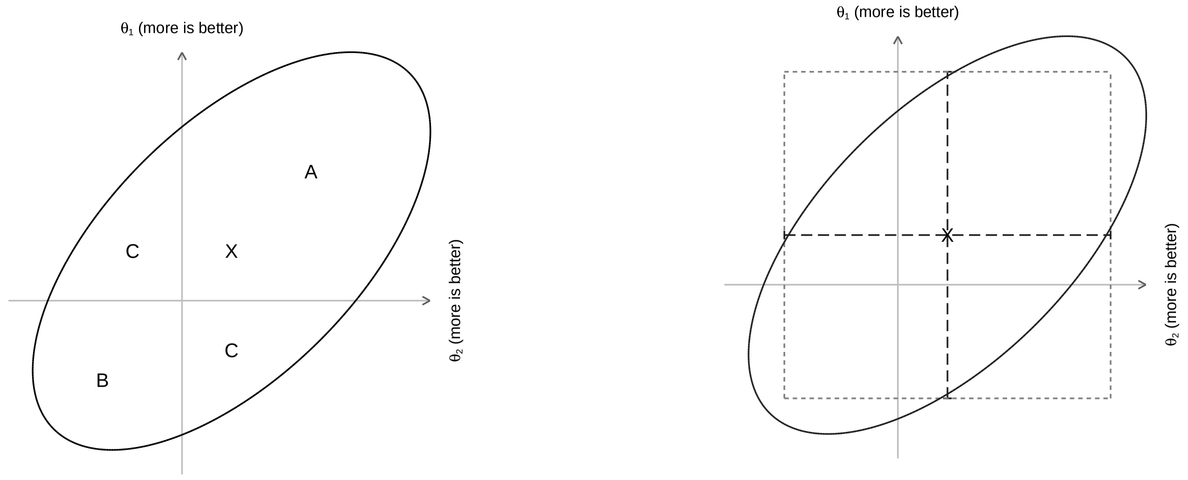
and by construction

$$Pr(\text{non-comparable } | J = j) = 1 - (Pr(\text{dominant } | J = j) + Pr(\text{dominated } | J = j)) \tag{4.5}$$

where S is the total number of simulations (here $S = 10,000$), θ^s_{jk} denotes the simulated provider-effect in simulation s, and I is an indicator function that takes the value of one if the condition is true and zero otherwise. This approach has several advantages over a series of univariate assessments: Most importantly, it accounts for the correlation between performance dimensions and thus achieves correct coverage of the confidence region (Briggs and Fenn 1998). Figure 4.1b illustrates the difference between probability statements if performances on both dimensions are incorrectly assumed to be independent. The dashed line outlines the resulting 'confidence box', which is formed by the end points of two independent 95% confidence intervals that are adjusted for multiple testing. Furthermore, because we make probability statements about a single quantity of interest, the provider's location in the k-dimensional performance space, we avoid such issues of multiple testing.

4.3.3 Risk-adjustment

Perhaps the primary reason that observed achievements differ across hospitals is because they treat different types of patients. To account for this, we develop specific risk-adjustment models for three of the performance dimensions. Based on previous research (Gutacker et al. 2013; Street et al. 2014), we identify a set of 'core' variables common to all models: patient age, gender, pre-treatment health status, primary diagnosis (coded as osteoarthritis (ICD-10: M15-19), rheumatoid arthritis (ICD-10: M05-06), or other) comorbidity burden, socio-economic status, and year of treatment. Other variables considered were time with symptoms, whether the patient lived alone, whether the patient required assistance filling in the PROM


θ_1 (more is better)
θ_2 (more is better)
A
C
X
C
B
θ_1 (more is better)
θ_2 (more is better)

(a) Areas covered if acknowledging correlation

(b) Difference between areas if achievements are assumed independent ('confidence box')

Legend: X denotes the centroid of the density. The solid ellipsoid line shows the inner 95% of the bivariate density with $\rho = 0.6$*, whereas the dashed line denotes the density covered by the confidence box that is formed by two independent 95% confidence intervals. The horizontal and vertical axes intersect at the benchmark and dissect each density into four areas, where the covered density of the area reflects the probability of dominating the benchmark (A), being dominated by the benchmark (B) or being non-comparable to the benchmark (C) (left panel).*

Figure 4.1: Example of area of probability density plane covered under different assumptions about the dependence of achievement scores

questionnaire, or whether she considered herself disabled.[57] Finally, in the length of stay model, we controlled for the healthcare resource group (HRG, the English equivalent of Diagnosis Related Groups) to which the patient was allocated.

Preliminary modelling of potential risk-adjusters was conducted on the basis of univariate multilevel regression models and visual inspection of LOWESS plots (for continuous variables) and box plots (for categorical variables). A significance level of $p < 0.05$ was required for variables to be retained. All continuous variables were first added linearly to the regression model and we subsequently explored whether squared terms improved the fit of the model. As expected, our exploratory work confirmed the importance of all core variables in explaining variation in each of the three performance dimensions. Time with symptoms, assistance and living alone did not explain variation in the probability of being re-admitted and were thus not included in the final model. Non-linear effects were found for age (all performance dimensions) and pre-treatment health status (only length of stay and post-operative OHS).

No risk-adjustment was performed in the analysis of waiting times because providers are expected to manage their waiting lists so as to balance high priority cases and those with less urgent need for admission.

4.3.4 Endogeneity due to patient selection of healthcare provider

Patients in the English NHS have a right to choose their provider of inpatient care for most elective procedures. This may lead to bias in the estimates of hospital performance if both the choice of hospital and the achievements for an individual patient are driven by common underlying factors that are not controlled for as part of X_{ijk}. This may arise if patients self-select into hospitals based on unobserved characteristics or providers cream-skim (Gowrisankaran and Town 1999; Geweke et al. 2003). Examples include unobserved severity, health literacy or other factors

[57] We only consider information contained in the pre-operative questionnaire since the e.g. need for assistance in filling in the post-operative questionnaire may be endogenous to the outcome of the care process.

that enter the personal health production function and are also determinants of hospital choice.

In order to test for bias due to patient selection and to obtain correct estimates of hospital performance, we estimate the model in (4.2) and perform two-stage residual inclusion (2SRI) as suggested by Terza et al. (2008). In the first stage, we estimate a multinomial choice model of hospital choice, where choice is assumed to be determined by the straight-line distance[58] from the patient's residence to the provider, an unobserved patient effect and random noise. Distance is commonly chosen in the literature as an instrumental variable as it is a major driver of hospital choice and is exogenously determined, on the reasonable assumption that patients do not choose where to live based on hospital performance (Gowrisankaran and Town 1999). The residual from this regression captures both the unobserved patient effect and random noise. In the second stage, we enter this residual as an additional regressor into each of the four achievement regression models. If the coefficients on the first-stage residuals are estimated to be statistically significantly different from zero this provides evidence of selection bias and the need for adjustments based on 2SRI (Terza et al. 2008).

4.4 Data

Our primary source of data is the Hospital Episode Statistics (HES) data warehouse, which contains detailed inpatient records for all patients receiving NHS-funded care in England. We extract information on all patients undergoing unilateral hip replacement (identified through the primary procedure code; see Department of Health (2008a)) in the period April 2009 to March 2012.[59] Patients were excluded if

[58] We also include distance^2 and distance^3 as well as an indicator for whether the hospital is the closest alternative. Hospitals with less than 30 patients were removed from the choice set. The patient's residence was approximated by the centroid of the lower super output area (LSOA) in which the patient lives. LSOAs are designed to include approximately 1,500 inhabitants, i.e. they are substantially smaller than US ZIP codes.

[59] HES records activity at the level of 'finished consultant episodes' (FCEs) and we link consecutive episodes within the hospital stay and across hospital transfers to form continuous inpatient spells (CIPS). A CIPS is deemed complete when the patient is discharged from one provider and not re-admitted to another provider within 2 days.

they were aged 17 or younger at the time of admission, underwent revision surgery, were admitted as emergencies or day-cases, or if information on important risk-adjustment variables was missing. Patients were also excluded if they attended a provider that treated fewer than 30 patients in the same financial year. We record any hospital admission occurring within 28 days after the initial admission for hip replacement surgery. All linkage was achieved using unique patient identifiers.

For each patient, we extract information on demographics and socio-economic background, medical characteristics and information pertaining to the admission process and the hospital stay itself. These data are used to construct three achievement measures: i) inpatient length of stay (top-coded at the 99th percentile), ii) emergency re-admission within 28 days of discharge for any condition (coded as 0=not re-admitted, 1=re-admitted), and iii) waiting time, measured as the time elapsed between the surgeon's decision to admit and the actual admission to hospital. Waiting time is categorised into waits of no more than 18 weeks (=0) and waits exceeding 18 weeks (=1) to mirror the contemporaneous waiting time performance standard in the English NHS.[60] We also derive the following risk-adjustment variables from HES: age, sex, comorbidity burden as measured by individual Elixhauser comorbidity conditions recorded in secondary diagnosis fields (Elixhauser et al. 1998), number of emergency admissions to hospital within the last year (coded as 0=none, 1=one or more), and patients' approximate socio-economic status based on level of income deprivation in the patient's neighbourhood of residence as measured by the Index of Multiple Deprivation 2004 (Noble et al. 2006).

We link HES records to data from the national Patient Reported Outcome Measures (PROM) survey. This survey invites all patients undergoing unilateral hip replacement to report their health status before and six months after surgery using the Oxford Hip Score (OHS) (Dawson et al. 1996).[61] The OHS is a reliable and validated

[60] The current performance standard is defined in terms of proportion of patients exceeding a waiting time of 18 weeks between the GPs referral and the admission (Department of Health 2015). Unfortunately, data on the time elapsed between the GPs referral and the surgeon's decision to admit are not recorded in HES. Our performance estimates will therefore be overstated.

[61] All patients are also invited to fill in the EuroQol-5D (EQ-5D) questionnaire, a generic health-related quality of life instrument (Brooks 1996). However, we focus on the OHS as it is better

measure of health status for hip replacement patients and consists of twelve questions regarding functioning and pain. For each item, the patient is asked to respond on a five-item scale. These items are summed up to generate an index score ranging from 0 (worst) to 48 (best). The post-operative OHS forms the fourth achievement measure and the pre-operative OHS score is used to control for initial health status at admission. Because we observe pre-operative health status in addition to the co-morbidity markers that are usually available in administrative records, our estimates of performance are more likely to indicate the true impact that providers have on performance measures ('value added') rather than reflect residual case-mix differences. The PROM survey also gathered additional information on duration of problems, and whether the patient lives alone, considered herself disabled, or required help filling in the questionnaire. Pre-operative survey responses are collected by paper questionnaire during the last outpatient appointment or on the day of admission, whereas follow-up responses are collected via mailed survey to the patient's home address. Participation in the PROM survey is voluntary for patients but mandatory for all providers of NHS-funded care. Approximately 60% of patients returned completed pre-operative questionnaires that can be linked to HES (Gutacker, Street et al. 2015).

4.5 Results

4.5.1 Descriptive statistics

The estimation sample consists of 95,955 patients treated in 252 providers during April 2009 and March 2012. Table 4.1 presents descriptive statistics. Patients are on average 67 years old, and approximately 41% of patients are male. The majority (68%) report having had problems with their hip joint for 1 to 5 years,

approximated by a continuous distribution and we do not seek to make comparisons across disease areas. Furthermore, the OHS is the relevant outcome measure for the newly introduced best practice tariff (a pay-for-performance scheme) in the English NHS that was introduced in April 2014 (Monitor and NHS England 2013). Previous comparisons have demonstrated that performance assessments based on the EQ-5D and OHS lead to similar conclusions (Neuburger, Hutchings, Meulen et al. 2013).

although about 8% of patients experienced symptoms for more than 10 years and 14% reported problems for less than 1 year. Approximately 39% of patients classify themselves as having a disability, and 27% live alone. Another 90,158 patients have been excluded from the analysis because of missing data, predominantly with respect to pre-operative health. These patients tend to be slightly older (68.7 vs 67.4 years), less likely to be male (39% vs 41%) and more likely to have been admitted as an emergency in the past year (11% vs 8%); see Appendix A4.1 for full descriptive statistics.

Figure 4.2 illustrates the empirical distributions of the achievement variables on their untransformed scales. The average post-operative OHS is 38.5 (SD=9.2) and the average length of stay is 5.4 days (SD=3.8), with both distributions showing substantial skew. Approximately 5.2% of patients were readmitted to hospital within 28 days of discharge, and about 17.5% of patients waited longer than 18 weeks to be admitted to hospital. There is a substantial proportion of missing responses in terms of post-operative OHS (15.2%) and, to lesser degrees, waiting time (4.0%) and length of stay (0.1%). Conversely, emergency re-admission status is recorded for all patients.

4.5.2 Provider heterogeneity and correlation between performance dimensions

All achievements are adjusted for case-mix. The estimated coefficients on risk-adjustment variables and associated standard errors are not the focus of this study and are reported in Table A4.2 in the Appendix. The first-stage residuals from the selection equation are jointly statistically significant ($\chi^2(4) = 14.97$; p<0.01) when entered into the main equations, suggesting that self-selection into hospital may bias performance estimates if uncontrolled for (see Table A4.3 in the Appendix for first-stage estimates). We therefore focus on results from models with adjustment for self-selection.

From the estimated variance-covariance matrices Σ and Ω we can calculate the correlation across performance estimates. The lower off-diagonal in Table 4.2 shows the correlation between performance estimates at provider level, whereas the upper

Table 4.1: Descriptive statistics

Description	N	Mean	SD
Achievement measures (Dependent variables)			
Post-operative OHS	81,336	38.50	9.21
Length of stay (in days)	95,878	5.36	3.75
Waiting time > 18 weeks	92,154	0.17	0.38
28-day emergency readmission	95,955	0.05	0.22
Patient characteristics (Control variables)			
Patient age (in years)	95,955	67.43	11.29
Patient gender (1=male, 0=female)	95,955	0.41	0.49
Pre-operative OHS	95,955	17.66	8.28
Primary diagnosis			
Osteoarthritis	95,955	0.93	0.25
Rheumatoid arthritis	95,955	0.01	0.07
Other	95,955	0.06	0.24
Number of Elixhauser comorbidities			
0	95,955	0.35	0.48
1	95,955	0.29	0.45
2-3	95,955	0.26	0.44
4+	95,955	0.10	0.31
Previously admitted as an emergency (1=yes, 0=no)	95,955	0.08	0.28
Socio-economic status	95,955	0.12	0.09
Disability (1=yes, 0=no)	95,955	0.39	0.49
Living alone (1=yes, 0=no)	95,955	0.27	0.44
Assistance (1=yes, 0=no)	95,955	0.21	0.41
Symptom duration			
< 1 year	95,955	0.14	0.35
1 - 5 years	95,955	0.68	0.47
6 - 10 years	95,955	0.11	0.31
> 10 years	95,955	0.08	0.26
Healthcare Resource Group			
HB12C - category 2 without CC	95,955	0.77	0.42
HB11C - category 1 without CC	95,955	0.10	0.29
HB12B - category 2 with CC	95,955	0.07	0.26
HB12A - category 2 with major CC	95,955	0.04	0.19
HB11B - category 1 with CC	95,955	0.01	0.11
other	95,955	0.02	0.12

Legend: N = Number of observations, SD = Standard deviation; OHS = Oxford Hip Score; CC = complications or co-morbidities.

Notes: Healthcare Resource Groups refer to major hip procedures for non-trauma patients in category 1 (HB12x) or category 2 (HB11x). Socio-economic status is approximated by the % of neighbourhood residents claiming income benefits. This characteristic is measured at neighbourhood level (lower super output area (LSOA)).

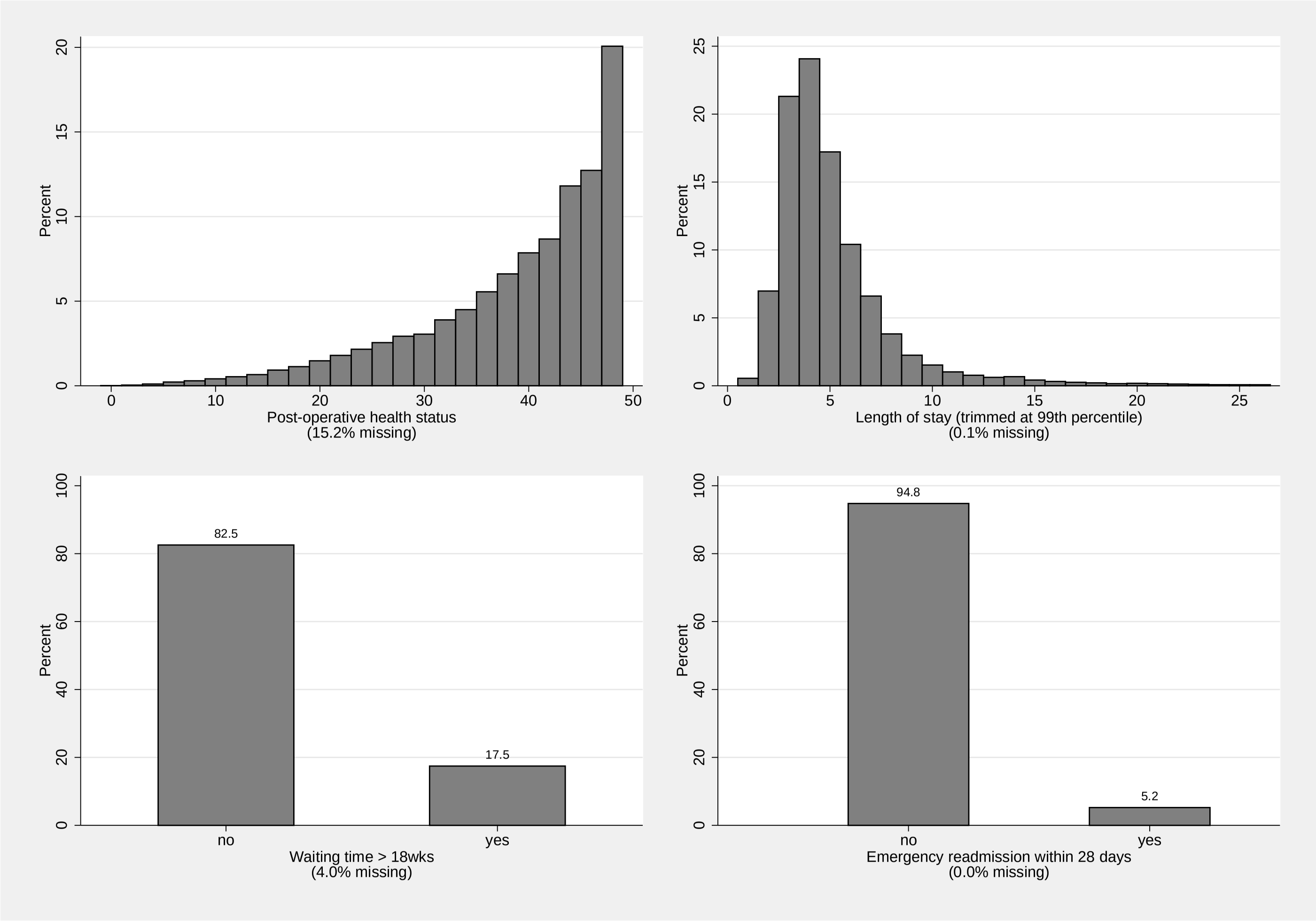


Figure 4.2: Empirical distribution of unadjusted achievement scores

Table 4.2: Correlation between performance dimensions

Performance dimension	(1)	(2)	(3)	(4)
Length of stay (1)	1.00	***-0.13***	***0.02***	***0.02***
Post-operative OHS (2)	**-0.34**	1.00	***-0.02***	***-0.07***
Waiting time > 18 wks (3)	**0.26**	**-0.31**	1.00	*0.00*
28-day emergency readmission (4)	0.03	**-0.49**	0.16	1.00

Notes: Lower triangle reports the correlation between random effects at provider level, whereas upper triangle (in italics) reports the correlation between random effects (i.e. the idiosyncratic error term) at patient level. Bold indicates that the correlation is statistically significantly different from zero at the 95% level.

off-diagonal shows the correlation at patient level. Bold numbers indicate that the correlation coefficient is statistically significantly different from zero (p<0.05; Huber-White standard errors).

We focus our discussion on the correlation between performance dimensions at provider level. Our results suggest significant correlations for four combinations of dimensions. Hospitals with shorter length of stay also realise better post-operative health status for their patients (ρ = -0.34; SE = 0.067; p<0.001). This is consistent with findings from randomised controlled trials that tested the effectiveness of so-called 'fast track' or 'enhanced recovery' pathways and found that hospitals that mobilise patients sooner after surgery were able to discharge them quicker and achieve better post-operative outcomes (Husted et al. 2008; Larsen et al. 2008; Paton et al. 2014). We also find evidence to suggest that hospitals with shorter length of stay also have a lower proportion of patients waiting more than 18 weeks to be admitted (ρ = 0.26; SE = 0.065; p<0.001), suggesting better management of capacity and of their waiting lists. This would be consistent with a queuing model of limited bed capacity, where prospective patients cannot be admitted until current patients are discharged. Hospitals that have better post-operative health outcomes also tend to have a lower proportion of patients waiting for more than 18 weeks (ρ= -0.31; SE = 0.071; p<0.001). Finally, the correlation between post-operative health status and probability of an emergency readmission within 28 days is negative and statistically significant (ρ = -0.49; SE = 0.078; p<0.001). Overall, these correlations indicate that inferences based on a series of univariate assessments would likely be misleading and that our MVMLM is preferable for this empirical analysis of provider performance.

It is also of interest to understand how much of the observed variability in adjusted achievement scores can be attributed to providers (Hauck et al. 2003). We calculate the intraclass correlation coefficient (ICC)[62] for each of the four performance dimensions with confidence intervals formed by the delta method. The largest ICC is observed for waiting times with 27.4% (SE = 0.020; p<0.001) of unexplained variation in achievements occurring between providers, followed by length of stay with approximately 13.3% (SE = 0.011; p<0.001). In contrast, the ICCs on the achievements post-operative OHS (1.7%; SE = 0.002; p<0.001) and emergency readmission (2.2%; SE = 0.003; p<0.001) are substantially smaller; implying that providers have less influence over these performance dimensions.

We have conducted sensitivity analyses with respect to a number of modelling choices (results are reported in Appendix Tables A4.4 to A4.6): First, we excluded privately own and operated providers (so called ‘independent sector treatment centres’ (ISTCs)) as these may be argued to operate under different production constraints (see below). The estimated covariance terms in Σ are somewhat attenuated and the correlations of waiting time with length of stay (p=0.174) and post-operative health status (p=0.857) are no longer statistically significant. Second, we included additional regressors based on patient risk factors averaged at provider level to correct for potential bias arising from correlation between X_{ij}’s and the hospital random effects (Mundlak 1978).[63] Due to convergence problems, we restricted this to patient age, pre-operative PROM score and level of income deprivation. Again, covariance terms are smaller in size but remain statistically significant. Finally, we restricted the risk-adjustment to variables that can be derived from routine administrative data, i.e. we excluded all variables based on the PROM survey. Results are

[62] The ICC for performance dimension k is $ICC_k = \frac{\tau_k^2}{\tau_k^2+\sigma_k^2}$.

[63] This bias is likely to be small. We compared coefficient estimates from fixed and random effects estimators using Hausman tests and found little practical difference between those estimates, although the tests all rejected the assumption of unbiasedness for the random effects approach. This is likely to be due to our large sample, where within effects swamp between effects and the Hausman test is over-powered. Results are reported in Appendix Tables A4.7 to A4.10.

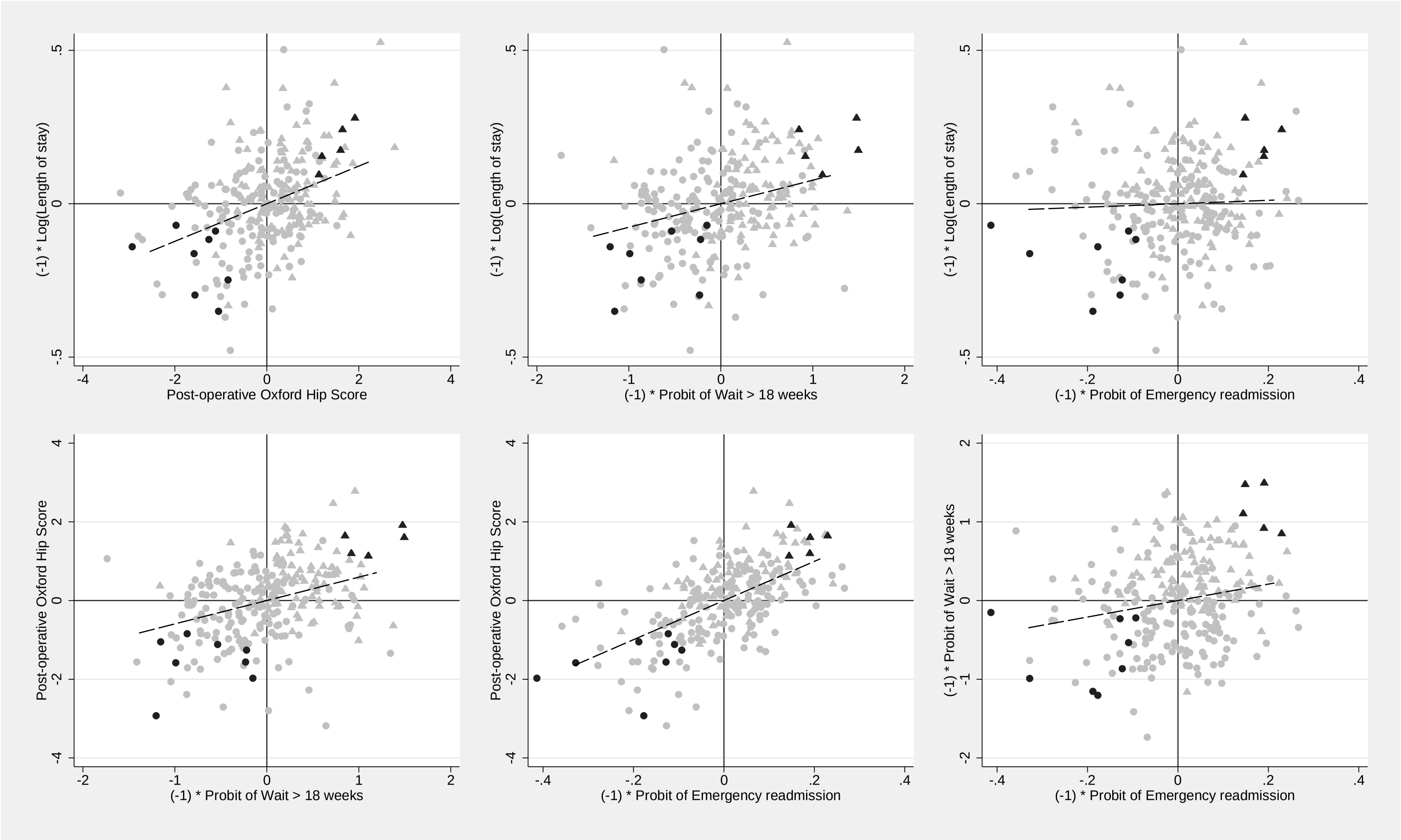
robust to this omission.

4.5.3 Provider performance assessment

We now turn to the assessment of multidimensional provider performance. Figure 4.3 shows the location of each provider in the four-dimensional performance space, where each panel presents scatter plots for two dimensions. The axes for all performance dimensions except post-operative health status are reversed (i.e. multiplied by -1) so that higher scores indicate better performance. Hence, providers in the NE quadrant perform better than the benchmark on both dimensions, whereas those in the SW quadrant perform worse. Providers that dominate or are dominated by the multidimensional benchmark with at least 90% probability are highlighted as darker points.

Figure 4.3 shows that we identify five dominant and eight dominated providers at a probability level of 90%. It turns out that all dominant providers are privately owned and operated treatment centres that perform mainly orthopaedic procedures, here marked as triangles, whereas all dominated providers are public NHS providers, marked as circles, that provide a wider mix of services, including emergency care. Note however that not all ISTCs are located in NE quadrant, and not all NHS providers are located in the SW quadrant. To test whether the observed performance advantage of ISTCs also holds on average, we re-estimated the models and included an indicator variable for private ownership. We found statistically significant effects on length of stay (beta=-0.100; SE = 0.020; p<0.001), post-operative health status (beta=1.205; SE = 0.157; p<0.001), probability of being readmitted (beta=-0.084; SE = 0.072; p<0.001), and the probability of waiting longer than 18 weeks (beta=-0.820; SE = 0.030; p=0.007).

Table 4.3 provides descriptive statistics for dominant and dominated providers in the financial year 2011/12. Both groups are comparable in terms of the annual volume of NHS-funded procedures provided. This suggests that volume-outcome effects may be less important in explaining overall performance differences. Conversely, we find that dominant providers operate in more competitive markets as



Notes: Each of the six panels shows bivariate plots of performance estimates. Higher scores imply better performance. Triangles indicate privately operated providers and circles indicate NHS providers.

Figure 4.3: Multidimensional performance estimates

indicated by the lower Herfindahl-Hirschman Index (HHI).[64] This finding is consistent with the theory of quality competition in price-regulated markets (Gaynor et al. 2015). Note however, that these comparisons are based on a small number of observations (J=13) and should be interpreted as associations. Ideally one would compare dominant ISTCs and dominated NHS hospitals across a wider range of characteristics (e.g. staffing ratios, experience of surgical teams, profit margin, etc.) to generate further hypotheses about the likely causal factors underlying those performance differences. Unfortunately, data limitations, especially with respect to ISTCs, prevent us from doing so.

Table 4.3: Characteristics of dominant and dominated providers (in 2011/12)

	Dominant (J=5)		Dominated (J=8)	
Description	Mean	SD	Mean	SD
Annual volume of hip replacements	361.60	198.16	365.38	190.04
Ownership (1=private, 0=NHS)	1.00	-	0.00	-
Herfindahl-Hirschman Index (HHI)	0.60	0.05	0.78	0.07

4.5.4 Comparison with approaches based on series of univariate probabilities

It is instructive to compare the results from our MVMLM assessment with two alternative approaches: 1) a series of four univariate multilevel regressions, and 2) an ‘intermediate’ MVMLM regression that takes into account the correlation between achievements during the estimation stage but treats performance estimates as independent. In both cases a provider is judged to be dominant [dominated] if all four individual probabilities of exceeding [falling short of] the benchmarks are greater or equal to a specified probability threshold (‘confidence box approach’). The second approach can thus be seen as an intermediate between a simple univariate approach and the full multivariate approach employed in this study.

[64] The HHI for provider j is calculated as the sum of the squared market shares of all providers $j^* = 1, \ldots, J^*$ that service LSOA $a = 1, \ldots, A$, here denoted as s_{aj}, weighted by the proportion of the provider’s observed total activity originating from this LSOA, s_{ja}, so that

$$HHI_j = \sum_a s_{ja} * [\sum_{j^*} (s_{aj^*})^2] \tag{4.6}$$

Hospital catchment areas are defined as all LSOAs within a radius of 30km around the hospital.

Table 4.4: Number of dominant/dominated providers under different estimation approaches and assumptions about the correlation between performance dimensions

Probability threshold Pr^*	(1) Univariate		(2) Intermediate multivariate		(3) Full multivariate	
	Dominant	Dominated	Dominant	Dominated	Dominant	Dominated
0.50	5	8	7	10	24	30
0.80	2	3	5	5	12	18
0.90	1	1	2	2	5	8
0.99	0	0	0	1	1	1

(1) Univariate approach - separate univariate models are estimated for each of the four performance dimensions and providers are considered dominant [dominated] if the independent probability of being dominant [dominated] exceeds $1-(1-Pr^*)/4$ on *each* of the four dimensions.
(2) Intermediate multivariate approach - multivariate model is estimated and providers are considered dominant [dominated] if the independent probability of being dominant [dominated] exceeds $1-(1-Pr^*)/4$ on *each* of the four dimensions. Correlation between performance dimensions is exploited in the estimation stage but ignored when forming probability statements.
(3) Fully multivariate approach - multivariate model is estimated and providers are considered dominant [dominated] if the probability of being dominant [dominated] on all four dimensions jointly exceeds Pr^*. See section 4.3.2 for details.

The univariate and intermediate multivariate approach both involve comparing four independent probabilities against a threshold value, which would lead to inflated risk of classifying providers as dominant [or dominated] when they are not (type I error). We adopt here the Bonferroni correction to adjust for multiple comparisons, i.e. we require (1-(1-Pr^*)/4)*100% probability on each of the four dimensions to designate a provider as dominant/dominated, where Pr^* equals the desired level of certainty.

Table 4.4 shows the number of providers identified as dominant/dominated under each of these approaches. At a probability threshold of 90% (Pr^*=0.9), the univariate and intermediate multivariate both identify just one or two dominant and dominated providers, which is fewer than the full MVMLM. The intermediate multivariate approach is more efficient than the univariate approach. This becomes apparent when applying an 80% threshold. At this probability threshold the univariate assessments identify two dominant and three dominated providers, whereas the intermediate MVMLM identifies five dominant and five dominated providers. The full MVMLM approach identifies 12 dominant and 18 dominated providers at the 80% threshold.

4.6 Discussion

Rarely are stakeholders explicit about the valuations they attach to different dimensions of performance, nor are these valuations likely to be identical. This renders the construction of a composite performance indicator that is appropriate for all audiences unfeasible. To circumvent this, we have set out a methodology for comparing healthcare providers in terms of their performance across a range of dimensions in a way that does not require valuation of each dimension and is consistent with economic theory. Building on previous literature, we analyse relative provider performance for each dimension and allow for correlation across dimensions (e.g. Hauck and Street 2006; Martin and Smith 2005; Hall and Hamilton 2004). We extend this literature by employing dominance criteria to compare providers against a multidimensional benchmark, and by constructing multivariate (rather than univariate) hypothesis tests of parameters that account for correlation between dimensions and thereby achieve correct coverage probabilities. Failure to perform multivariate tests can lead to incorrect inferences about multidimensional performance as we illustrate.

We have applied our MVMLM approach to study the performance of English providers of care to patients having hip replacement. By focusing on a single procedure, we can draw more robust conclusions about performance than studies conducted at hospital level. Our use of patient-level data allows us to employ multilevel models to control for a diverse range of patient characteristics and, thereby, to isolate the provider's impact on observed achievements. We study four dimensions of performance, namely long waiting times (>18 weeks), length of stay, 28-day readmission rates, and patient-reported health status after surgery. Achievements on some of these dimensions are correlated, implying that our multivariate estimation framework is appropriate. Our results do not suggest trade-offs between achievements on the four performance dimensions we studied. Instead, we observe positive, albeit weak, correlations. We wish to stress that these results do not necessarily imply a causal relationship between achievements, although some of our findings confirm those of randomised controlled trials conducted in routine care settings.[65] Nevertheless, this suggests that pairs of achievements are either a) driven by common underlying factors that enter both production functions, such as organisational effort, or b) that achievements on one dimension enable achievements on another. This information is of interest itself as it informs the debate whether incentive schemes can be simplified to reward providers on a subset of correlated measures, as suggested by Glazer et al. (2008), or whether regulators should instead ascertain performance across all individual performance dimensions of interest.

Our estimation yields, for each provider, one performance estimate per performance dimension, which together form a provider's performance profile. To translate this profile into a single statement about performance we employ a set of dominance criteria and classify providers into three groups: (i) dominant providers, which are 'positive deviants' that exhibit outstanding performance across all performance dimensions; (ii) dominated providers, which are 'negative deviants' with sub-standard performance; and (iii) the remainder. In this study of patients having hip replacement, all dominant providers were found to be privately operated treatment centres specialising in elective (i.e. non-emergency) hip and knee replacement, while all dominated providers were public NHS providers providing a wide range of services. ISTCs have previously been found to achieve on average better health outcomes than public providers (Browne et al. 2008; Chard et al. 2011) and to discharge patients earlier (Siciliani et al. 2013), and we can confirm these findings in our data. This may be the result of a more stream-lined production process: ISTCs typically focus exclusively on elective orthopaedic procedures, such as hip and knee replacement, whereas NHS providers offer a wide range of service, including emergency care. If the organisational set-up of ISTCs allows them to specialise, this may result in performance advantages. Our data do not allow us to unpack the reasons for the observed performance further, and we stress that performance assessment results should form the starting point for further investigations involving site visits and

[65] Importantly, these trials also provide evidence on the *direction* of the causal effect, i.e. what causes what.

qualitative analysis (Bradley et al. 2009). As with most regression analyses, general differences between types of providers can be identified using conditional mean comparisons, in which indicator variables are used to specify provider types. But our approach also allows us to identify positive and negative deviants *within* these broad categories of provider type. This is important as otherwise regulatory efforts may be accidently directed at those NHS hospitals that are found to perform relatively well; and vice versa for the identification of best practice in ISTCs.

The appeal of the dominance approach lies in the absence of strong assumptions about the various stakeholders' utility functions and its ability to reduce multiple performance estimates into a single assessment. However, this comes at a price. Because the approach requires providers to perform better than the benchmark on *all* dimensions, there is no scope to compensate for average or poor performance on one dimension through excellent performance on another. This very strict yardstick is difficult to achieve and so we identify only a small number of providers as dominant or dominated. Also, as the number of objectives under consideration increases it becomes increasingly more difficult to satisfy the dominance criteria (Pedraja-Chaparro et al. 1999). Nevertheless, although we have illustrated our methodology by analysing only four dimensions, it is generalisable to multiple dimensions.

These qualifications not withstanding, we advocate the dominance approach to multidimensional performance assessment as a useful addition to regulators' toolboxes.

Acknowledgments

We are grateful for comments and suggestions from Stirling Bryan, Chris Bojke, Richard Cookson, Bernard van den Berg, Pedro Saramago, Miqdad Asaria and those received during the HESG Summer 2012 meeting (Oxford), the 2014 iHEA and ECHE conference (Dublin), and the 34th Spanish Health Economics Association conference (Pamplona). This research was funded by the Department of Health in England under the Policy Research Unit in the Economics of Health and Social Care

Systems (Ref 103/0001). The views expressed are those of the authors and may not reflect those of the Department of Health.

5 Do patients choose hospitals that improve their health?

5.1 Introduction

Many European healthcare systems have recently extended patients' right to choose their provider of elective hospital care (Vrangbaek et al. 2012). Enhanced choice can accommodate patients' preferences for provider characteristics (e.g. proximity, quality or availability of amenities) and create market conditions that incentivise providers to compete (Besley and Ghatak 2003). Patients in the English National Health Service (NHS) have to be referred to inpatient services by their general practitioner, who acts as a gatekeeper, but are free to choose their preferred provider of care. Prices for hospital care are set nationally and patients do not bear the cost of treatment, so providers are expected to compete for elective patients on the basis of quality. Two prerequisites for such quality competition are that patients and their agents[66] have access to reliable, meaningful and understandable information about the quality of care offered by alternative providers, and that they act upon such information (Besley and Ghatak 2003; Marshall et al. 2004; Faber et al. 2009).

English patients can access comparative information on hospital quality through several channels, including the NHS Choices website, the Health & Social Care Information Centre (HSCIC) website and the Dr Foster Hospital Guide. These present information on risk-adjusted 28-day mortality and emergency readmission rates, calculated from routine hospital discharge data. Such indicators have been criticised as being incomplete and noisy measures of quality, revealing little about the

[66] These may include the patient's general practitioner (GP) as well as family, friends and others. Some patients may not be willing or able to make a choice and their referring GP may choose the most appropriate hospital for them, i.e. the GP acts as an agent to the patient. It is generally not possible to distinguish between decision makers using administrative data. For simplicity, we will henceforward denote the decision-maker as the patient.

changes in health that the vast majority of patients will experience as the result of treatment (Appleby and Devlin 2004; Lilford and Pronovost 2010). This is especially so for mortality rates for common elective operations such as hip (0.3%) and knee replacement (0.2%), which are generally very low (Berstock et al. 2014; Belmont et al. 2014).

New hospital quality measures that address these concerns are increasingly available. Since April 2009, all providers in the English NHS have been required to collect patient-reported outcome measures (PROMs) for all NHS-funded patients undergoing unilateral hip and knee replacement, varicose vein surgery or groin hernia repair (Department of Health 2008a). PROMs are validated questionnaires used to elicit patients' health status and health-related quality of life (HRQoL). Each eligible patient is invited to complete a PROM questionnaire before and three or six months after their surgery. The changes in scores can be interpreted as the improvement in patients' health and are used for hospital benchmarking (Nuttall et al. 2015; Gutacker et al. 2013).

Hospital quality measures derived from PROMs improve over 'failure' measures such as mortality or emergency readmission rates in several ways. First, they capture the entire spectrum of health (Appleby and Devlin 2004; Gutacker et al. 2013) and thus allow inferences about improvements in health as a consequence of treatment. Second, because post-operative health status is adjusted for pre-operative status, it can be argued that they adjust better for case-mix. Finally, PROMs reflect the patients' view on their health and health improvement. This, one may argue, makes them especially relevant for prospective patients who are about to choose their provider.

It has been the English Department of Health's expressed ambition to establish patients' self-reported outcomes as an important component of hospital quality assessment. It was also hoped that such information would be used *"by patients and GPs exercising choice"* (Department of Health 2008a, p.6). Consequently, provider-specific average risk-adjusted changes in health status have been disseminated online on a regular basis since the beginning of the national PROM programme (Health &

Social Care Information Centre 2013b). Some patients might access this information directly, whereas others might rely on their general practitioners, who act as their agent, to retrieve, interpret and communicate this information.

In this study we test whether hospital demand responds to PROM-based measures of hospital quality in addition to more conventional measures such as mortality and readmission rates. We estimate a hospital choice model for elective hip replacement surgery in the English NHS to identify how hospital choice responds to hospital and patient characteristics. Our focus is on two key aspects of hospital choice: 1) whether hospitals with better PROM-derived quality (as measured by the changes in patients' Oxford Hip Score (OHS)) face higher demand and 2) whether patients' response to quality differs according to their morbidity, as measured by the pre-operative health status, and other characteristics such as age or income deprivation. To address potential endogeneity we use lagged quality and waiting times. We also undertake robustness checks using hospital fixed effects and by comparing the effects of quality on choices by elective hip replacement patients with those by emergency hip replacement patients who we expect to be less sensitive to quality.

This is the first study which explores whether hospital demand responds to quality as measured by average patient health gains at provider level, which are derived from patient self-reported outcome measures. The existing literature has predominantely focused on failure measures such as mortality rates, either measured at aggregate hospital level or for specific conditions (Sivey 2008; Beckert et al. 2012; Moscone et al. 2012; Gaynor et al. 2012), readmission rates (Varkevisser et al. 2012; Moscone et al. 2012), as well as hospital reputation and other composite scores (Pope 2009; Varkevisser et al. 2010; Varkevisser et al. 2012; Ruwaard and Douven 2014); see Brekke et al. (2014) for an overview. These studies have typically reported a positive relation between quality and hospital demand. Second, we make novel use of pre-operative individual level PROMs data to explore such questions as whether sicker patients travel farther and choose hospitals with higher quality of care as often assumed in the literature (Gowrisankaran and Town 1999; Geweke et al. 2003). Previous studies have either relied on instrumental variable approaches to approximate the role of (unobserved) pre-operative health status on demand (Gowrisankaran and Town 1999; Geweke et al. 2003) or have used measures of comorbidity burden and past utilisation as proxies for health status. Our data allow us to explore this issue more directly. Third, our study contributes to the small literature on hospital choice in publicly funded health systems where demand is rationed by waiting time (Sivey 2012; Beckert et al. 2012; Moscone et al. 2012; Gaynor et al. 2012). Our analysis differs from Beckert et al. (2012), who also study choice of provider for hip replacement surgery in England, in that we use provider quality measures which are procedure-specific and more directly related to the quality of care provided[67], explore the role of pre-operative health status, and model the entire relevant market, including private providers of NHS-funded care.

The remainder of the chapter is structured as follows: in the next section we describe the data used in this study in more detail. Section 5.3 describes our econometric model and sets out our strategy to mitigate potential endogeneity bias. In Section 5.4 we present the estimated marginal utilities of hospital characteristics and show how these vary with observed patient characteristics. Section 5.5 presents the estimated effects of changes in providers' quality on their own demand and that of their competitors. Finally, the last section offers a discussion of the results.

5.2 Data

We use patient-level data from Hospital Episode Statistics (HES) for all elective admissions for patients aged 18 or over who underwent NHS-funded primary (i.e. non-revision) hip replacement surgery[68] between April 2010 and March 2013 in NHS or private providers. HES contains rich information on patients' demographic and

[67] Beckert et al. (2012) model hospital quality using hospital-wide mortality and MRSA infection rates. Aggregate hospital level quality indicators, such as the summary hospital mortality indicator (SHMI) used in the English NHS, do not correlate well with procedure-specific outcome measures (Gravelle et al. 2014). In 2010/11, the Pearson correlation coefficients between SHMI and the quality measures used in this study were -0.09 (OHS), -0.05 (emergency readmission rate) and 0.10 (mortality rate), respectively.

[68] See Department of Health (2008a) for procedure codes. We exclude patients that underwent revision surgery to ensure a more homogeneous sample and because these are believed to be likely to return to the place of initial surgery, independent of observed hospital attributes.

medical characteristics, small area of residence and on the hospital stay. Privately funded patients treated in the private sector are not included in HES and are excluded from our analysis.[69]

We derive a number of patient variables from HES: patients' age, gender, the number of emergency admissions during the 365 days prior to their hip replacement admission, and the number of Elixhauser comorbid conditions recorded in admissions in the previous year (Elixhauser et al. 1998; Gutacker, Bloor et al. 2015). We also obtain an identifier of the GP practice that the patient is registered with. These are available for all patients. We use the 2004 Index of Multiple Deprivation (Noble et al. 2006) to attribute to each patient the proportion of residents claiming means-tested benefits in their Lower Super Output Area (LSOA)[70], which we interpret as a measure of income deprivation. We measure a patient's distance from a hospital as the straight-line distance from the centroid of their LSOA.[71]

The PROM survey invites all NHS-funded hip replacement patients to report their health status and HRQoL before and six months after surgery using a paper-based questionnaire. The pre-operative questionnaire is administered by the hospital either as part of the admission process or during the last outpatient appointment preceding admission. The post-operative questionnaire is administered by a central agency and posted to the patient. Participation in the PROM survey is compulsory for providers but optional for patients. Approximately 60% of patients provide complete pre- and postoperative PROM questionnaires that can be linked to their HES record (Hutchings et al. 2014; Gutacker, Street et al. 2015).

Each PROM questionnaire contains three instruments: the Oxford Hip Score

[69] Approximately 11% of the English population have private (supplementary) insurance and approximately 16% of hip replacement surgeries are funded privately, either out-of-pocket or through private insurance (Commission on the Future of Health and Social Care in England 2014).

[70] HES records patients' locations in terms of the LSOA (2001 census boundaries) in which they reside. Each LSOA contains approximately 1,500 inhabitants and is designed to be homogeneous with respect to tenure and accommodation type.

[71] We determine a hospital's location on the basis of its headquarter's postcode (for NHS trusts) or the postcode of the individual hospital's site (for ISTCs). We do not model NHS hospital sites individually as quality information for these providers is only recorded at trust level and hospital site codes are often poorly recorded in HES data. This is likely to induce noise to our distance measure.

(OHS), the EuroQoL-5D (EQ-5D) descriptive system, and the EuroQol Visual Analogue Scale (EQ-VAS). The OHS is a condition-specific instrument that consists of 12 questionnaire items regarding hip-related functioning and pain (Dawson et al. 1996). Each item is scored on a five-point scale, with four indicating no problems and zero indicating severe problems. The overall score is calculated as the sum of all items and ranges from zero (worst) to 48 (best). Both EuroQol instruments are generic PROMs, i.e. they can be applied to different health conditions, and are described in detail elsewhere (Brooks 1996). Previous analysis showed substantial correlation between the EQ-5D and OHS (Neuburger, Hutchings, Meulen et al. 2013). Since the OHS is a condition-specific measure and hence plausibly more likely to affect hospital choice for hip replacements we focus on the OHS throughout this study. Also, the OHS is the relevant outcome measure for the newly introduced best practice tariff for hip replacement surgery (a pay-for-performance scheme) and we therefore expect providers to be more concerned with their performance on it (Monitor and NHS England 2013).

We use PROMs data in two ways. First, we obtained risk-adjusted hospital-specific PROM change scores for the OHS from the HSCIC website (Health & Social Care Information Centre 2013b). Data are reported by financial year, which run from April to March of the next year. The HSCIC excludes from these reports providers with less than 30 valid pre- and post-operative PROM returns due to concerns about statistical validity and patient anonymity. The case-mix adjustment methodology is reported elsewhere (Department of Health 2012a).[72] There is some evidence to suggest that the hospital-specific mean scores are robust to missing data (Gomes et al. 2015). Second, in some of our models, we use the information in the individual patients' pre-operative PROMs questionnaires to measure their pre-operative morbidity and investigate whether choice of provider is affected by pre-operative morbidity. Because patients can decline to participate or providers may fail to administer a questionnaire

[72] The adjustment takes into account a range of patient characteristics including age, sex, pre-operative PROM score, socio-economic status, comorbidity burden, whether the patient lives alone as well as other indicators of disability.

there is scope for missing data and selection bias, and we explore this in the empirical analysis for the subset of models which make use of pre-operative morbidity.

We calculate risk-adjusted hospital-specific 28-day emergency readmission and 28-day mortality rates after hip replacement as additional quality measures. These data are presented on patient information websites (such as NHS Choices). To compute them, we link our HES data to Office of National Statistics death records and apply the HSCIC case-mix adjustment as set out in the readmission outcome indicator specification (Health & Social Care Information Centre 2013a).[73]

We group providers into seven categories used by the National Patient Safety Agency: NHS small / medium / large non-teaching trust, NHS teaching trust, NHS specialised orthopaedic provider, NHS multi-service provider, and NHS Primary Care Trusts (PCTs).[74] We also distinguish NHS hospitals from Independent Sector Treatment Centres (ISTCs) which are private providers treating NHS patients.

Finally, we derive from HES the median time (in months) that patients in each hospital had to wait between the specialist's decision to add the patient to the waiting list and the admission (the inpatient wait). Patients in the English NHS do not pay for their care directly and waiting times thus serve as a rationing mechanism (Iversen and Siciliani 2011). We use the median rather than the mean because it is less affected by a small number of patients with very long wait and thus more representative of the expected waiting time for a prospective patient. We also conduct sensitivity analysis using the proportion of patients in this hospital that waited longer than 120 days.

[73] Both readmission and mortality rates are adjusted for age (in 5-yr bands), sex, socio-economic status, comorbidity burden as captured by the Charlson index and the number of emergency admissions in the last year.

[74] PCTs are responsible for purchasing care for their resident population and, with the exception of the Isle of Wight PCT, do not provide care themselves.

5.3 Methods

5.3.1 Model specification

We use a random utility choice model (McFadden 1974). Utility of patient $i = 1, \dots, N$ at provider $j = 1, \dots, J$ at time $t = 1, \dots, T$ is $U_{ijt} = V_{ijt} + \xi_{jt} + \epsilon_{ijt}$, where V_{ijt} depends on observable hospital characteristics and travel distance, ξ_{jt} are unobserved hospital characteristics, and ϵ_{ijt} is unobserved random utility. Patients choose from a set of hospitals $M_{it} \in J$. Assuming ϵ_{ijt} is iid extreme value yields the *multinomial logit* (MNL) model in which the probability that patient i chooses hospital j is

$$P_{ijt} = \exp \frac{V_{ijt} + \xi_{jt}}{\sum_{j \in M_{it}} V_{ij't} + \xi_{j't}} \tag{5.1}$$

We assume that all patients who require treatment are treated, i.e. there is no outside option.

In our baseline specification, utility is a linear additive function of the distance from the patient's residence to the hospital D_{ij}, distance squared D^2_{ij}, hospital quality metrics Q_{jt-1}, waiting time W_{jt-1}, and a vector of time-invariant hospital characteristics Z_j, so that

$$U_{ijt} = D^{'}_{ij}\beta_{d,i} + D^{2'}_{ij}\beta_{d^2,i} + Q^{'}_{jt-1}\beta_{q,i} + W^{'}_{jt-1}\beta_{w,i} + Z^{'}_{jt}\beta_{z,i} + \xi_{jt} + \epsilon_{ijt} \tag{5.2}$$

where ξ_{jt} and ϵ_{ijt} are unobserved. We assume that anticipated utility at a provider is based on its previous period's quality and waiting time because relevant information are available only with a lag (see section 5.3.2). Varkevisser et al. (2012) make a similar assumption. We also estimate models with contemporaneous waiting time and quality scores in sensitivity analyses.

We allow preferences to vary across patients according to their observed characteristics. Thus the marginal utility of quality for patient i is

$$\beta_{q,i} = \beta_q + X^{'}_i\delta_q \tag{5.3}$$

and similar for distance, waiting time, and other hospital characteristics. All continuous covariates in X_i are mean centred and base categories for categorical characteristics are set to their mode. Thus, the vectors of coefficients β_d, β_{d^2}, β_q, β_w, β_z reflect the preferences of an average/modal patient, hereafter referred to as the '*reference patient*'.

We also estimate models which allow for unobserved patient heterogeneity in tastes over quality, with

$$\beta_{q,i} = \beta_q + X_i^{'}\delta_q + \sigma_q\alpha_i \tag{5.4}$$

where σ_q is the standard deviation of a normal variable with mean zero and α_i is an unobserved patient effect. The latter may capture, for example, differences in the ability to access and interpret quality information. This *random coefficient multinomial logit* (RCMNL) or mixed logit model (Hensher and Greene 2003; Train 2003), unlike the MNL model, allows for unrestricted substitution patterns, thereby relaxing the assumption of independence of irrelevant alternatives (IIA).[75] If $\sigma_q = 0$ then the RCMNL model reduces to the MNL model in (5.2).

While the MNL model has a closed form solution that can be estimated via maximum likelihood, the RCMNL needs to be approximated through simulation. To reduce the computational burden[76] of the RCMNL model we assume uncorrelated normally distributed random coefficients for the quality metrics in Q_{jt-1} and no random coefficients for other covariates. The RCMNL model is estimated with maximum simulated likelihood using 50 Halton draws.

All models are estimated in Stata 13 with `clogit` and the user-written command `mixlogit` (Hole 2007b). Standard errors are clustered at the GP practice level to allow for agent-induced correlation across patients: patients in the same practice are expected to make more similar choices than patients in different practices if GPs have an influence on their decisions.

[75] The IIA states that the probability of choosing one hospital over another depends solely on the characteristics of these two hospitals and not on the characteristics of any other hospital. The standard MNL model imposes the IIA assumption, whereas the RCMNL does not.

[76] Even after imposing those constraints the RCMNL model with our baseline specification still took over 5 days to compute on a high-performance computing system.

5.3.2 Endogeneity

To interpret β_q as an unbiased estimate of the marginal utility of hospital quality (up to a linear transformation) requires that the unobserved hospital effect ξ_{jt} is uncorrelated with any of the independent variables, i.e. all observed variables are exogenous. This assumption may not hold for four reasons (Varkevisser et al. 2012; Gaynor et al. 2012; Brekke et al. 2014).

First, hospitals may learn by doing so that higher volume providers have higher quality (Luft et al. 1987; Gaynor et al. 2005). Thus changes in demand will also affect quality and induce simultaneity bias. Based on the institutional context of this study we argue that this concern can be dismissed. While volume-outcome effects have been reported for elective joint replacement surgery, these scale effects tend to occur only in very low volume hospitals that treat less than 100 patients per year (Judge et al. 2006; Mäkelä et al. 2011). The increasing incidence of hip replacement surgery in England and trends to aggregate services in high-volume hospitals mean that all NHS providers in our sample are comfortably above this threshold and has led commentators to suggest that volume effects are of little relevance in the English NHS (Judge et al. 2006). For private providers we cannot ascertain their true level of activity as treatment of non-NHS patients is not recorded in HES, but we expect those to perform a sufficient number of procedures to operate profitably. The average hospital in our sample treats over 300 patients per year.

Second, because of short run capacity constraints, changes in demand will also affect waiting time in the same period (Gaynor et al. 2012).[77] While our primary interest is not in the effect of waiting time on demand, we are concerned that any bias introduced through endogenous variables will filter through to our estimate of β_q (Wooldridge 2002). However, if, as we assume, demand depends on past, rather than current, quality and waiting time, then demand changes in period t cannot affect waiting time at $t-1$.

[77] It may also be that supply and demand are determined simultaneously, i.e. hospitals react to demand shocks by adjusting their supply, e.g. by performing more surgeries on weekends. We do not consider this in our model explicitly, although the use of lagged waiting time circumvents this problem as well.

Third, sicker patients may choose higher quality hospitals or hospitals may turn away or discourage patients with characteristics that make them less likely to achieve a large improvement in health status. If such systematic selection occurs and is not controlled for in the calculation of hospital quality scores then those scores would in part be determined by patients' choices or provider selection. However, provider quality scores are adjusted for a rich set of demographic, socio-economic, and morbidity patient characteristics, including, in the case of PROMs, the patients' self-reported pre-operative health status. Hence, we do not believe that unobserved patient selection is likely to bias the quality scores significantly.

Finally, there may be unobserved hospital characteristics that affect demand and are correlated with observed covariates (Jung et al. 2011). For example, hospitals in areas with better amenities may attract better staff thereby ensuring higher observed clinical quality but also unobserved interpersonal aspects of quality. Our assumption that patients use information on previous period quality and waiting times when choosing hospitals does not remove omitted variable bias operating through unobserved non-transitory hospital characteristics. However, the low correlations between the PROM quality measure and the conventional readmission and mortality measures suggest that omitted variables may not lead to serious bias. We undertake two types of sensitivity analyses to explore the size of the potential omitted variable bias. Our first approach is to estimate the choice model in (5.2) with alternative-specific time-invariant fixed effects (FEs) (Hodgkin 1996; Monstad et al. 2006; Sivey 2012). These hospital FEs capture the utility of non-transitory unobserved hospital characteristics. The coefficients on observed hospital characteristics are now identified solely through variation within providers over time, thereby removing any endogeneity bias operating through unobserved time-invariant characteristics. However, this approach is quite demanding of the data, and because we only observe providers over three years we expect this approach to result in imprecise estimates of the marginal utility of hospital quality. Also, because our market structure changes over time, due to the opening of new independent sector treatment centres, the FEs do not correspond to observed market shares in each time period. This may bias estimates if incumbent providers differ systematically from new entries. We therefore also estimate a model based on NHS trusts only, whose numbers are relatively stable over time.

Our second approach is to follow Pope (2009) (see also Gaynor et al. (2012)) and gauge the possible impact of unobserved hospital heterogeneity by using a control group of emergency hip replacement patients whose choice of provider is less responsive to quality and waiting time. The majority of emergency hip replacement patients suffer from a fractured neck of the femur as a result of a fall and official recommendations are that they should be treated within 48 hours (NICE 2011). Further delays are linked to worse outcomes (Moja et al. 2012). We therefore expect provider choice by emergency hip replacement patients to be less affected by publicly reported information on quality and more by distance to providers and time-invariant unobserved factors, such as long-standing reputation or dimensions of accessibility not captured by our distance measure (e.g. parking charges or connection to the public transport system).

If we assume that emergency patients' demand is entirely inelastic to observed quality and they do not wait[78], but value the same unobserved hospital characteristics as elective patients, then their true utility is given by

$$U_{ijt}^{Emer} = D_{ij}^{'}\beta_{d,i}^{Emer} + D_{ij}^{2'}\beta_{d^2,i}^{Emer} + \xi_{jt} + \epsilon_{ijt} \tag{5.5}$$

If we estimate the model specified in (5.2) for emergency patients and find $\widehat{\beta}_q^{Emer} \neq 0$, we conclude that $cov(Q_{jt-1}, \xi_{jt}) \neq 0$. Moreover, if we assume that elective and emergency patients have the same preferences for unobserved hospital characteristics, then the effect of quality on elective demand, purged of omitted variable bias, is $\beta_q^{\Delta} = \beta_q^{Elec} - \beta_q^{Emer}$. Since coefficients in separate MNL models may be scaled differently, we estimate a pooled model for elective and emergency patients

[78] Elective waiting time and associated supply constraints do not apply to emergency patients, i.e. there is always sufficient capacity to treat an emergency patient. Given the urgent nature of the condition, patients will usually be treated within hours of arrival, not weeks or months. Explorations of our data revealed that elective waiting time is only weakly correlated with the volume of emergency patients, suggesting that supply for these distinct groups is separate.

by interacting all covariates with an indicator variable for emergency. This forces the scaling to be the same. The coefficients on the interaction terms are estimates of β_k^Δ for $k \in [d, d^2, q, w, z]$.

If emergency patients are also sensitive to elective quality[79], or emergency quality that correlates with it, or if unobserved hospital characteristics have different effects on choices by emergency and elective patients and are correlated with observed quality, then β_q^Δ can no longer be interpreted as the unbiased effect of quality on elective demand. If unobserved hospital factors are not correlated with quality, then β_k^Δ reflects the differences in preferences in two distinct groups of patients: those that require urgent care and have less time to compare hospitals, and those that have sufficient time to reach an informed decision. In this case, we expect that $\beta_q^\Delta > 0$: elective patients will be more sensitive to quality than emergency patients.

5.3.3 Elasticities, changes in demand and willingness to travel

The estimated coefficients on quality are estimates of the marginal utility from quality. Since the utility function is unique only up to a linear transformation, the coefficients only convey information about the sign of marginal utility of hospital characteristics and hence about the sign of the effect of quality on demand. The ratio of estimated marginal utilities (the negative of the marginal rate of substitution) is unaffected by linear transformations and so provides quantitative and comparable information about patient preferences. We estimate the reference patient's willingness to travel (WTT) for a one standard deviation (SD) increase in quality as

$$WTT = \frac{\partial D_{ij}}{\partial Q_j}|_{U_{ij}} SD(Q) = -\frac{\partial U_{ij}}{\partial D_{ij}} / \frac{\partial U_{ij}}{\partial Q_j} SD(Q) = \frac{-\beta_q}{\beta_d + 2\beta_{d^2} D} SD(Q) \qquad (5.6)$$

where D is the median distance to hospitals in patients' choice sets. We estimate standard errors by the delta method (Hole 2007a). WTT is the extra distance in kilometres that the reference patient located the median distance away from a

[79] As with elective patients, we do not observe who chooses the hospital for emergency hip replacement. This may be the patient, a family member, GP, or the ambulance crew.

provider would be willing to travel to that provider if its quality was increased by $SD(Q)$, where $SD(Q)$ is averaged across hospitals and years.

We are also interested in whether providers could attract more patients by improving their quality. Expected demand at provider j is $Y_{jt} = \sum_{i \in S_{jt}} P_{ijt}$, where S_{jt} is the set of patients whose choice set includes provider j, i.e. for whom $j \in M_{it}$. Following Santos et al. (2015) we calculate the average partial effect of a one SD increase in quality on provider j's demand, i.e. demand responsiveness to quality, as

$$\frac{\partial Y_{jt}}{\partial Q_{jt-1}} SD(Q) = SD(Q) \sum_{i \in S_{jt}} \frac{\partial P_{ijt}}{\partial Q_{jt-1}} = SD(Q) \sum_{i \in S_{jt}} \beta_q P_{ijt}(1 - P_{ijt}) \qquad (5.7)$$

We report the mean of (5.7) over all providers and years.

We calculate the elasticity of demand of provider j with respect to own quality as

$$E_{jt}^{Q_{jt-1}} = \sum_{i \in S_{jt}} \frac{\partial P_{ijt}}{\partial Q_{jt-1}} \frac{Q_{jt-1}}{Y_{jt}} = \sum_{i \in S_{jt}} \beta_q P_{ijt}(1 - P_{ijt}) \frac{Q_{jt-1}}{\sum_{i \in S_{jt}} P_{ijt}} \qquad (5.8)$$

We report the mean of (5.8), weighted by providers' predicted demand $\sum_{i \in S_{jt}} P_{ijt}$.

Finally, we compute the cross-elasticity of demand for provider j with respect to the quality of provider $j^{'}$ as

$$E_{jt}^{Q_{j'}} = \sum_{i \in S_{jt} \cap S_{j^{'}t}} \frac{\partial P_{ijt}}{\partial Q_{j't-1}} \frac{Q_{j't-1}}{\sum_{i \in S_{jt}} P_{ijt}} = - \sum_{i \in S_{jt} \cap S_{j^{'}t}} \beta_q P_{ijt} P_{ij't} \frac{Q_{j't-1}}{\sum_{i \in S_{jt}} P_{ijt}} \qquad (5.9)$$

with $j \neq j^{'}$. Note that for some combinations of j and $j^{'}$ the cross-elasticity is zero because no patients have both providers in their choice sets.

5.4 Results

5.4.1 Descriptive statistics

Our main sample is 173,773 elective hip replacement patients treated in 230 providers during the period April 2010 to March 2013.[80] Their average age is 68 years and 40% are male (Table 5.1). The average pre-operative OHS is 17.5 and 9% of patients have been admitted to hospital as an emergency at least once during the preceding 365 days (average number of admissions = 0.13). Self-reported pre-operative OHS is only weakly correlated with past emergency utilisation (ρ = -0.10) and the number of comorbidities (ρ = -0.14). This suggests that past emergency utilisation and comorbidity burden are poor proxies for current health status[81] as experienced by the patient.

On average, within 30km patients have a choice of 8 providers, with over 90% of patients having access to at least two different providers. Even within 10km there are on average 1.6 hospitals and over 20% of patients can choose between two or more providers. To reduce computational burden we restrict patient choice sets to the 50 nearest providers.[82] The 741 patients (or 0.04% of the sample) who chose a provider outside this set were dropped from the analysis.

Patients live on average 14.7 kilometres from their chosen hospital. Figure 5.1 shows that just over half (53.7%) of patients bypassed the local hospital and nearly a fifth (18.3%) bypassed the nearest three hospitals. On average, patients travel 5.4 km (SD=14.8) beyond their nearest hospital to be treated.[83]

[80] The number of providers varied slightly over this period because of mergers, changes in coding and market entry, especially with respect to private facilities. There were 157 providers in 2010/11, 202 in 2011/12, and 212 in 2012/13, of which 18 (11.5%) in 2010/11, 62 (30.7%) in 2011/12, and 78 (36.8%) in 2012/13 are privately operated.

[81] We also calculated the correlations between these measures and the EQ-5D utility score, which one may argue is a more holistic measure of health-related quality of life. The correlations are similar: ρ = -0.10 for past utilisation, and ρ = -0.14 for comorbidity burden.

[82] Choice sets are deliberately chosen to be large to avoid introducing selection bias. Not many patients may search out information on all 50 hospitals' characteristics before making a choice. However, given the strong preference for hospitals nearby and the assumption of IAA, including extra alternatives should not affect the model estimates.

[83] These numbers are somewhat higher than those reported by Beckert et al. (2012), presumably because our data also cover private providers treating NHS-funded patients.

Table 5.1: Descriptive statistics - elective sample

Variable	Obs	Mean	SD	ICC
Patient characteristics				
Distance travelled (in km)	173,773	14.7	17.7	
Distance travelled past closest provider (in km)	173,773	5.4	14.8	
Number of providers within 10km radius	173,773	1.6	1.7	
Number of providers within 30km radius	173,773	8.5	7.3	
Age	173,773	68.0	11.5	
Male	173,773	0.40	0.49	
Past utilisation	173,773	0.13	0.49	
Number of Elixhauser conditions	173,773	0.43	0.94	
Income deprivation	173,773	0.12	0.09	
Pre-operative Oxford Hip Score[a]	71,614	17.5	8.2	
Provider characteristics				
Observed volume	571	304.3	209.1	94.7%
Waiting time (in months)	571	2.5	1.1	77.4%
Change in Oxford Hip Score	571	19.8	1.4	57.0%
28-day emergency readmission rate (in %)	571	5.65	2.41	36.8%
28-day mortality rate (in %)	571	0.17	0.36	3.4%

Obs = Observations; SD = Standard deviation; ICC = Intraclass correlation coefficient.
Notes: Patient characteristics for patients choosing provider between April 2010 and March 2013. Provider waiting time, change in Oxford Hip Score, readmission rate, mortality rate are for financial years 2009/10 to 2011/12. Provider characteristics are unweighted.

[a] Responders to PROM survey that were treated between April 2010 and March 2012.

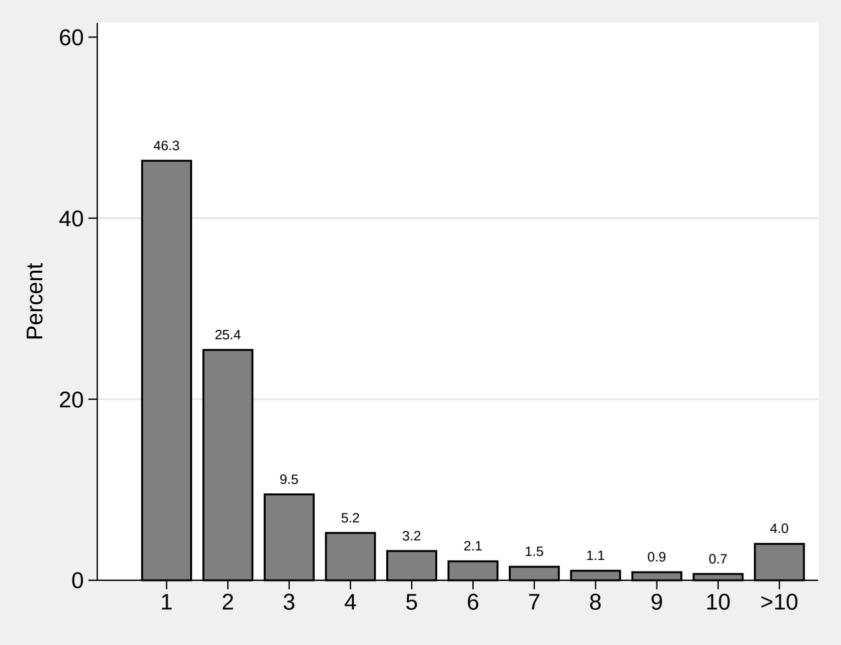


Figure 5.1: Percentage of elective patients who went to their Nth nearest hospital

The hospital waiting time and quality scores are lagged by one year and are for financial years 2009/10 to 2011/12. The risk-adjusted OHS health again has a mean of 19.8 with a SD of 1.4. There are much larger coefficients of variation for hospital emergency re-admission and mortality rates. The average waiting time at provider level is 2.5 months, which is substantially lower than in previous years (see Appendix Figure A5.1 and Siciliani et al. 2014). The provider OHS change scores are only weakly correlated with waiting time (ρ=-0.30), readmission rates (ρ=-0.28) and mortality rates (ρ=-0.05). This suggests that choice models that are restricted to mortality and readmission rates may not even indirectly pick up the effect of PROM measures on demand.

The intra-class correlation coefficient (ICC) shows that just over half of the observed variation in OHS change scores is between providers (ICC=57%) rather than over time.[84] Between-provider variation is markedly greater for waiting times (ICC=77%). Most of the variation in readmission rates and mortality is within providers.

5.4.2 Regression results

5.4.2.1 Main effects

The results from for the RCMNL model (see Appendix Table A5.1) suggest no significant variation in the random coefficients on the quality metrics. Hausman tests also did not reject the IIA assumption. We therefore concentrate on the MNL models reported in Tables 5.2 to 5.5.

Table 5.2 is our preferred specification with distance, lagged waiting time, the three lagged quality metrics and indicators for the type of provider as well as interactions with patient age, gender, past utilisation, comorbidity, and local area income deprivation (we explore interactions with pre-operative OHS in section 5.4.2.2). This specification does not include hospital FEs. The main effects are the estimated

[84] These ICCs differ from those reported in previous chapters, which focused on variation in individual patients' scores across providers.

marginal utilities for the reference patient with mean or modal characteristics. The reference patient prefers shorter distances with the marginal disutility from distance declining with distance. She prefers specialised providers to non-specialised providers. She is also more likely to choose a public provider over a private provider after accounting for distance, waiting time and quality.[85]

Table 5.2: Estimated marginal utilities

Variable	Est	SE
Main effects		
Distance (in km)	-0.184	0.002***
Distance^2	0.000	0.000***
NHS trust - medium	-0.530	0.030***
NHS trust - multi-service	-0.603	0.099***
NHS trust - small	-0.791	0.038***
NHS trust - specialist	1.023	0.072***
NHS trust - teaching	-0.445	0.033***
Independent sector treatment centre	-1.467	0.045***
Primary care trust	-1.159	0.206***
Waiting time (in months)	0.013	0.015
Change in Oxford Hip Score	0.118	0.008***
28-day emergency readmission rate (in %)	-0.052	0.004***
28-day mortality rate (in %)	-0.031	0.026
Interaction with distance		
x Patient age	-0.002	0.000***
x Male	0.002	0.001
x Past utilisation	-0.003	0.002
x Comorbidity count	-0.004	0.001***
x Income deprivation	-0.186	0.017***
Interaction with waiting time		
x Patient age	0.003	0.000***
x Male	-0.009	0.009
x Past utilisation	-0.006	0.012
x Comorbidity count	-0.018	0.007**
x Income deprivation	0.046	0.084
Interaction with change in Oxford Hip Score		
x Patient age	0.001	0.000*
x Male	-0.007	0.005
x Past utilisation	-0.010	0.007
x Comorbidity count	-0.009	0.003**
x Income deprivation	-0.455	0.047***
Interaction with 28-day emergency readmission rate		

continued

[85] During our study period ISTC were funded through block contracts and paid to provide care to a pre-specified number of NHS patients. However, most ISTCs did not fulfil their quotas although they generally had low waiting times (Naylor and Gregory 2009). Our results are consistent with this observation and suggest a positive preference for public providers by NHS-funded hip replacement patients. Brown et al. (2015) also found evidence of a pro-public preference in the stated preferences of the general public in New Zealand.

Table 5.2: Estimated marginal utilities

Variable	Est	SE
x Patient age	-0.0003	0.000*
x Male	0.000	0.003
x Past utilisation	0.011	0.003**
x Comorbidity count	0.001	0.002
x Income deprivation	0.125	0.026***
Interaction with 28-day mortality rate		
x Patient age	-0.001	0.001
x Male	-0.053	0.022*
x Past utilisation	0.046	0.026
x Comorbidity count	-0.009	0.015
x Income deprivation	-0.025	0.168
WTT(OHS change)	1.287	0.085***
WTT(Readmission rate)	-0.981	0.079***
WTT(Mortality rate)	-0.086	0.072
Number of patients	173,032	
Number of providers	230	
BIC	460,994	
Pseudo R^2	0.637	

*** p<0.001; ** p<0.01; * p<0.05

Notes: Conditional logit model of choice of hospital for elective hip replacement patients treated between April 2010 and March 2013. OHS change, waiting time, readmission rate and mortality rate are lagged by one year. Coefficients are marginal utilities. WTT is the ratio of the coefficient on the quality variable to the marginal utility of distance evaluated at the median distance (in km). Interaction terms with distance^2 and provider type not reported (available on request). Standard errors are clustered at GP practice level.

Reference patient demand is increasing with the OHS change score and falling with emergency admission rates. The estimated WTT for a one SD increase in OHS is 1.3 km or 8.7% of the average distance travelled to the chosen provider. The WTT for a SD decrease in emergency readmission rates is 1.0km. There is no statistically significant effect of procedure-specific mortality rates on demand. Nor does the waiting time affect choice of provider, which may be a result of the historically short waiting time during our study period.[86]

Results are robust to the use of contemporaneous rather than lagged waiting time and quality (Appendix Table 7.4, model 1). Contemporaneous waiting time has a positive but statistically insignificant coefficient. When we use the proportion of patients waiting longer than 120 days as a waiting time measure the coefficient is

[86] A similar argument has been made by Brown et al. (2015), who estimate the waiting time elasticity of demand in New Zealand to be -0.004, much lower than values of -0.07 to -0.14 previously reported in the literature (Martin et al. 2007).

negative and statistically significant (Appendix Table 7.4, model 2). The coefficients on the quality measures are almost unaffected by the use of contemporaneous waiting time and quality.

The HSCIC also produces hospital quality scores based on the case-mix adjusted change in the EQ-5D utility score. This is highly correlated with the OHS change score (Neuburger, Hutchings, Meulen et al. 2013) and when we estimate the baseline specification with EQ-5D substituted for OHS we find similar WTT (Appendix Table 7.4, model 3). Results are also robust to exclusion of independent sector treatment centres from patient choice sets (Appendix Table 7.4, model 4).

5.4.2.2 Patient heterogeneity

The coefficients on the interaction terms in the lower parts of Table 5.2 suggest that preferences vary across types of patient. We find, like other studies (Propper et al. 2007; Beckert et al. 2012), that older patients dislike distance more. They care less about waiting time and get greater marginal utility from improvements in the OHS change score, reductions in emergency readmissions and reductions in mortality rates. There is little difference between the preferences of male and female patients except that male patients have a greater dislike for providers with higher mortality. Preferences vary little by morbidity as measured by past emergency admissions. In contrast, patients with more comorbidities have a greater dislike of distance and waiting time, but care less about readmission rates. Finally, patients from neighbourhoods with greater income deprivation care more about distance and less about quality.

The existence of detailed patient reported pre-operative health status measures in our dataset allows us to explore in more detail whether patients in worse health status are more sensitive to quality and more willing to travel, as commonly assumed in the literature on hospital quality (Gowrisankaran and Town 1999; Geweke et al. 2003). The correlations between patients' pre-operative OHS and their routinely available morbidity measures are low, suggesting that they measure different aspects of the patient's condition at the time of admission.

The first model in Table 5.3 is the same as our preferred specification but with additional patient pre-operative OHS interactions. Interaction terms with other patient characteristics are suppressed for brevity. Due to data limitations, we focus on patients treated during April 2010 and March 2012. We find that healthier patients are more willing to travel. Although the marginal utility from higher quality is similar for healthier patients, the reduced distance cost for these patients implies they are more willing to travel for higher quality. Healthier patients are also more likely to choose a private provider, which is consistent with observed differences in intake across provider types (Browne et al. 2008).

The fact that pre-operative OHS data are available for only about 60% of patients raises concerns about response bias if unobserved factors affect propensity to respond and utility from providers.[87] To investigate if responders to the pre-operative PROM questionnaire have different preferences to non-responders we re-estimate the preferred specification of Table 5.2 for our full sample (responders and non-responders) but interact a dummy variable for responder status with all the main and interacted explanatory variables; pre-operative health status is not modelled.

The pre-operative PROM questionnaire is administered *after* the patient has chosen the provider. Hence, it is unclear whether the response indicator variable reflects patient preferences or whether the choice determines the response indicator. For example, private providers have higher response rates than NHS hospitals (Gomes et al. 2015; Gutacker, Street et al. 2015) and also tend to have higher observed quality and shorter waiting times. We address this concern by including the observed provider pre-operative response rate as a provider characteristic when modelling the choices of responders and non-responders. This variable is informative about the individual's propensity to fill in a pre-operative PROM questionnaire given the chosen provider.[88] We find that responders and non-responders have generally very

[87] We are not concerned about the implications of response rates for the hospital level case-mix adjusted OHS change scores as these have been shown to be robust to variations in response rate (Gomes et al. 2015).

[88] As a check, we first re-estimate the responder only model with the addition of provider pre-operative response rates. The results are robust to this sensitivity analysis, with the WTT of 1.4km (SE=0.102) for a standard deviation increase in PROM quality being slightly larger than in our

Table 5.3: Choice models allowing for patient pre-operative Oxford Hip Score

	Patients with pre-op OHS (1)		Patients with pre-op OHS (2)		All patients (3)					
					Responders (3a)		Non-responders (3b)		Difference (3c)	
Variable	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Main effects										
Distance (in km)	-0.185	0.002***	-0.185	0.002***	-0.185	0.002***	-0.188	0.007***	0.003	0.007
Distance^2	0.000	0.000***	0.000	0.000***	0.000	0.000***	0.000	0.000**	0.000	0.000
NHS trust - medium	-0.526	0.037***	-0.646	0.037***	-0.641	0.037***	-0.558	0.039***	-0.083	0.034*
NHS trust - multi-service	-0.902	0.133***	-0.973	0.131***	-0.965	0.130***	-0.519	0.104***	-0.446	0.123***
NHS trust - small	-0.820	0.045***	-0.907	0.045***	-0.902	0.044***	-0.857	0.045***	-0.045	0.041
NHS trust - specialist	1.052	0.079***	0.869	0.082***	0.856	0.082***	0.973	0.089***	-0.117	0.070
NHS trust - teaching	-0.445	0.039***	-0.503	0.039***	-0.489	0.039***	-0.608	0.039***	0.119	0.038**
Independent sector treatment centre	-1.373	0.065***	-1.499	0.063***	-1.515	0.063***	-1.670	0.067***	0.155	0.059**
Primary care trust	-0.978	0.223***	-1.301	0.223***	-1.298	0.224***	-1.272	0.235***	-0.026	0.232
Waiting time (in months)	-0.011	0.020	0.035	0.020	0.033	0.020	-0.054	0.023*	0.087	0.018***
Change in Oxford Hip Score	0.161	0.010***	0.139	0.010***	0.137	0.010***	0.104	0.011***	0.033	0.010**
28-day emergency readmission rate (in %)	-0.050	0.006***	-0.046	0.006***	-0.046	0.006***	-0.052	0.006***	0.006	0.006
28-day mortality rate (in %)	-0.135	0.035***	-0.068	0.032*	-0.067	0.032*	-0.014	0.036	-0.053	0.033
Response rate			2.044	0.093***	2.038	0.092***	-2.287	0.080***	4.325	0.086***
Interaction with pre-operative Oxford Hip Score										
x Distance (in km)	0.001	0.000***	0.001	0.000***						
x Distance^2	0.000	0.000***	0.000	0.000***						
x NHS trust - medium	0.000	0.002	-0.001	0.002						
x NHS trust - multi-service	-0.004	0.007	-0.006	0.007						
x NHS trust - small	-0.002	0.002	-0.002	0.002						
x NHS trust - specialist	0.017	0.004***	0.016	0.004***						
x NHS trust - teaching	-0.005	0.002*	-0.008	0.002***						
x Independent sector treatment centre	0.038	0.003***	0.035	0.003***						
x Primary care trust	-0.002	0.008	-0.005	0.008						
x Waiting time (in months)	0.004	0.001***	0.004	0.001***						
x Change in Oxford Hip Score	0.001	0.001*	0.000	0.001						
x 28-day emergency readmission rate (in %)	0.000	0.000	0.000	0.000						
x 28-day mortality rate (in %)	0.002	0.002	0.003	0.002						
x Response rate			0.020	0.005***						
WTT(OHS change)	1.717	0.111***	1.475	0.113***	1.465	0.112***	1.048	0.136***	0.417	0.133**
WTT(Readmission rate)	-0.941	0.107***	-0.867	0.105***	-0.879	0.105***	-0.932	0.132***	0.054	0.127
WTT(Mortality rate)	-0.474	0.122***	-0.238	0.112*	-0.224	0.107*	-0.045	0.113	-0.180	0.107
Number of patients	71,329		71,329		113,751					
Number of providers	206		206		206					
BIC	182,407		179,628		283,989					
Pseudo R^2	0.649		0.654		0.657					

*** $p<0.001$; ** $p<0.01$; * $p<0.05$

Notes: Conditional logit model of choice of hospital for elective hip replacement patients treated between April 2010 and March 2012. OHS change, waiting time, readmission rate and mortality rate are lagged by one year. Coefficients are marginal utilities. WTT is the ratio of the coefficient on the quality variable to the marginal utility of distance evaluated at the median distance (in km). Models in (1) and (2) are for patients reporting a pre-operation OHS. Model in (3) is for all patients and interacts a dummy variable for reporting a pre-operation OHS. Interaction effects are reported in (3c). All models also contain a full set of interactions of age, gender, past utilisation, Elixhauser comorbidities, and deprivation with hospital characteristics and distance (not reported). Standard errors are clustered at GP practice level.

similar revealed preferences, with the exception of preferences for waiting times (non-responders prefer shorter waiting times) and PROM quality (responders derive more utility from health gains and are thus more willing to travel for it). There is no difference with respect to the disutility from travel distance, readmission rates or mortality.

5.4.3 Omitted variable bias

We also explore the possible impact of omitted hospital characteristics on our estimates of marginal utility for quality and other hospital characteristics. We compare preferences of elective and emergency patients estimated from pooled choice models with a full set of emergency patient dummy variables interacted with all explanatory variables. There are 73,629 emergency patients in our sample. Only 20% of emergency patients bypassed the nearest provider (see Appendix Figure A5.2). Descriptive statistics for this patient group are reported in Appendix Table A5.3. Emergency patients' choice sets are the 50 closest providers who carried out hip replacement surgery on at least 30 emergency patients in this year. This rules out private and specialised providers who only treat elective hip replacement patients. 708 (1.0%) emergency patients were dropped because they attended a provider not in their choice set. All main effects still pertain to the elective reference patient.

We report results for two different specifications. The first model in Table 5.4 compares emergency patients with elective patients who choose NHS or independent providers. However, there are some marked differences in observed characteristics between those two groups. For example, emergency patients are on average 12 years older than elective patients and have over twice as many recorded comorbidities. Hence in the second model reported in Table 5.5 we compare a set of elective and emergency patients matched exactly on age, gender, past emergency admissions, number of comorbidities, income deprivation and year of treatment. Additionally, we restrict the elective patient sample to those who used an NHS provider that treats

preferred specification (full results available on request).

at least 30 elective and emergency patients in that year; hence the choice sets are identical for elective and emergency conditional on location.

Table 5.4: Comparison of marginal utilities for elective and emergency patients

	Elective patients		Emergency patients		Difference	
Variable	Est	SE	Est	SE	Est	SE
Distance (in km)	-0.184	0.002***	-0.217	0.004***	-0.033	0.003***
Distance^2	0.000	0.000***	0.001	0.000***	0.000	0.000***
NHS trust - medium	-0.530	0.030***	-0.571	0.045***	-0.041	0.039
NHS trust - multi-service	-0.603	0.099***	-0.935	0.164***	-0.332	0.145*
NHS trust - small	-0.791	0.038***	-0.823	0.050***	-0.032	0.044
NHS trust - specialist	1.023	0.072***	n/a		n/a	
NHS trust - teaching	-0.445	0.033***	-0.609	0.045***	-0.164	0.042***
Independent sector treatment centre	-1.467	0.045***	n/a		n/a	
Primary care trust	-1.159	0.206***	-1.274	0.258***	-0.115	0.176
Waiting time (in months)	0.013	0.015	-0.010	0.022	-0.023	0.021
Change in Oxford Hip Score	0.118	0.008***	0.048	0.013***	-0.070	0.012***
28-day emergency readmission rate (in %)	-0.052	0.004***	-0.046	0.008***	0.006	0.007
28-day mortality rate (in %)	-0.031	0.026	0.056	0.056	0.087	0.057
WTT(OHS change)	1.285	0.085***	0.523	0.143***	-0.763	0.126***
WTT(Readmission rate)	-0.978	0.079***	-0.870	0.150***	0.109	0.134
WTT(Mortality rate)	-0.085	0.072	0.155	0.155	0.241	0.157
Number of patients	173,032		72,921			
Number of providers	230		138			
BIC	570,669					
Pseudo R^2	0.689					

*** $p<0.001$; ** $p<0.01$; * $p<0.05$

Notes: Conditional logit model of choice of hospital for elective and emergency hip replacement patients treated between April 2010 and March 2013. OHS change, waiting time, readmission rate and mortality rate are lagged by one year. Coefficients are marginal utilities for the 'reference patient'. Elective and emergency patients are not matched on observed characteristics but the 'reference patient' in both patient populations is defined according to the average characteristics of the elective patient sample. WTT is the ratio of the coefficient on the quality variable to the marginal utility of distance evaluated at the median distance (in km). Model is estimated with a full set of dummy variables interacted with hospital characteristics and other interaction terms. All models also contain a full set of interactions of age, gender, past utilisation, Elixhauser comorbidities, and deprivation with hospital characteristics and distance (not reported). Standard errors are clustered at GP practice level.

Table 5.5: Comparison of marginal utilities for elective and emergency patients - matched sample

	Elective patients		Emergency patients		Difference	
Variable	Est	SE	Est	SE	Est	SE
Distance (in km)	-0.220	0.004***	-0.215	0.004***	0.005	0.005
Distance^2	0.001	0.000***	0.000	0.000***	0.000	0.000***
NHS trust - medium	-0.709	0.050***	-0.560	0.046***	0.149	0.053**
NHS trust - multi-service	-0.710	0.175***	-0.921	0.182***	-0.211	0.213
NHS trust - small	-0.880	0.058***	-0.794	0.053***	0.086	0.062
NHS trust - teaching	-0.468	0.053***	-0.598	0.048***	-0.130	0.058*
Primary care trust	-1.133	0.292***	-1.429	0.314***	-0.296	0.282
Waiting time (in months)	-0.087	0.027**	-0.032	0.024	0.055	0.029
Change in Oxford Hip Score	0.092	0.015***	0.034	0.014*	-0.058	0.016***
28-day emergency readmission rate (in %)	-0.061	0.009***	-0.046	0.008***	0.015	0.010
28-day mortality rate (in %)	-0.050	0.061	0.070	0.063	0.120	0.079
WTT(OHS change)	0.796	0.128***	0.354	0.142*	-0.441	0.149**
WTT(Readmission rate)	-0.893	0.129***	-0.814	0.152***	0.079	0.165
WTT(Mortality rate)	-0.084	0.102	0.140	0.127	0.224	0.148
Number of patients	32,274		32,274			
Number of providers	138		138			
BIC	107,831					
Pseudo R^2	0.771					

*** p<0.001; ** p<0.01; * p<0.05

Notes: Conditional logit model of choice of hospital for elective and emergency hip replacement patients treated between April 2010 and March 2013. OHS change, waiting time, readmission rate and mortality rate are lagged by one year. Coefficients are marginal utilities for the 'reference patient'. Elective and emergency patients are matched exactly on observed characteristics (age, gender, past emergency utilisation in last year (none, once, or more), income deprivation of neighbourhood, number of Elixhauser comorbit conditions, year of treatment) and the 'reference patient' in both patient populations is defined according to the average (prior to matching) characteristics of the elective patient sample. Choice sets include only providers that treat at least 30 elective and 30 emergency hip replacement patient in this period. WTT is the ratio of the coefficient on the quality variable to the marginal utility of distance evaluated at the median distance (in km). Model is estimated with a full set of dummy variables interacted with hospital characteristics and other interaction terms. All models also contain a full set of interactions of age, gender, past utilisation, Elixhauser comorbidities, and deprivation with hospital characteristics and distance (not reported). Standard errors are clustered at GP practice level.

Both models suggest that emergency patients care less about provider OHS changes but have similar preferences over the more traditional quality measures using readmission and mortality rates. In the second specification, with closely matched patients, the estimated marginal utility of OHS changes (β_q^{Emer}=0.034) is just over one third of that for elective patients (β_q^{Elec}=0.092) and significant at p<0.05. If we assume that emergency patients' demand is entirely inelastic to variation in observed elective quality and that the estimated association for emergency patients is a result of omitted variables that affect emergency and elective patients in the same way, then the difference in the marginal utility of OHS changes (β_q^{Δ}=0.058) can be interpreted as a lower bound estimate of the true effect of OHS change score on elective patient utility. The WTT for a one SD increase in OHS change scores then is 0.4km (SE=0.149), which is smaller than that reported in Table 5.2.

Table 5.6 shows the main effects from our preferred specification estimated with additional hospital FEs. We find that PROM quality still has a statistically significant effect on demand, whereas emergency readmission rates no longer do. The WTT to travel for PROM quality is however 87% lower than that calculated from the results in Table 2 (0.2km vs 1.3km). This is likely to be due to the fixed effect absorbing part of the effect of time-invariant quality on choice. Results are broadly similar when patients' choice sets are restricted to NHS hospitals, although we now find a counter-intuitive positive effect of waiting time on demand.

Table 5.6: Choice model controlling for unobserved time-invariant hospital effects

	All providers (1)		NHS providers only (2)	
	Est	SE	Est	SE
Distance (in km)	-0.202	0.002***	-0.231	0.003***
Distance^2	0.000	0.000***	0.001	0.000***
Waiting time (in months)	0.021	0.024	0.053	0.012***
Change in Oxford Hip Score	0.017	0.006**	0.016	0.007*
28-day emergency readmission rate (in %)	0.005	0.004	0.000	0.004
28-day mortality rate (in %)	0.038	0.020	0.021	0.026
WTT(OHS change)	0.168	0.060**	0.124	0.055*
WTT(Readmission rate)	0.089	0.066	-0.0001	0.053
WTT(Mortality rate)	0.095	0.051	0.031	0.039
Number of patients	173,032		148,629	
Number of providers	230		144	
BIC	411,541		260,299	
Pseudo R^2	0.678		0.742	

*** $p<0.001$; ** $p<0.01$; * $p<0.05$

Notes: Conditional logit model of choice of hospital for elective hip replacement patients treated between April 2010 and March 2013. OHS change, waiting time, readmission rate and mortality rate are lagged by one year. Coefficients are marginal utilities. WTT is the ratio of the coefficient on the quality variable to the marginal utility of distance evaluated at the median distance (in km). Model in (1) does not impose restrictions on the type of provider in patients' choice sets. Model in (2) is based on a restricted choice set of NHS providers, thereby excluding patients that selected ISTCs. All models include indicator variables for hospitals (not reported). All models also contain a full set of interactions of age, gender, past utilisation, Elixhauser comorbidities, and deprivation with hospital characteristics and distance (not reported). Standard errors are clustered at GP practice level.

5.5 The economic effects of quality on demand

We use the results from choice models to illustrate the effect of quality differentiation on hospital demand. Column four and five of Table 5.7 provide the marginal utilities of the different quality measures and the willingness to travel for a one SD increase in these measures. The sixth and seventh columns show the average total and relative change in demand from a one SD increase in quality, and column eight gives the own quality demand elasticities. We base our calculations on the estimates for our preferred specification in Table 5.2. This should be kept in mind when interpreting the results presented in this section.

Table 5.7: Effect sizes of hospital quality measures

	Observed		Marginal utility	Effect of SD increase in quality			Elasticity of demand
Quality indicator	Mean	SD		WTT	Demand change	% Demand change	
Change in Oxford Hip Score	19.8	1.4	0.118	1.3	33.9	9.4	1.3
Emergency readmission rate (in %)	-5.6	2.4	-0.052	-1.0	-25.3	-7.0	-0.2
Mortality rate (in %)	-0.2	0.4	-0.031	-0.1	-2.2	-0.6	0.0

Notes: All calculations based on estimated marginal utilities reported in Table 5.2. WTT is the ratio of the coefficient on the quality variable to the marginal utility of distance evaluated at the median distance (in km). Changes in volume and elasticities are averaged across hospital-year observations and are weighted by predicted demand $\widehat{Y}_{ijt} = \sum_{i \in M_{it}} P_{ijt}$.

The expected increase in demand for a SD increase in OHS is approximately 34 patients, or 9.4% of predicted demand at current quality levels. Increases in readmission and mortality rates are associated with decreases in demand, although the association of mortality and demand is not statistically significant. The effect of a one SD increase in OHS is larger than that of a one SD decrease in readmission rate.

There is substantial variation across providers in the effect of OHS change scores on own demand (Figure 5.2). The estimated elasticities range from 0.2 to 2.4 (mean = 1.3). About 42% of the variation in elasticities is explained by the amount of competition a provider faces, here measured by the Herfindahl-Hirschman Index (HHI).[89] Providers in more competitive areas (low HHI) face larger quality elasticities than those in less competitive areas (high HHI), with elasticities falling by approximately 0.29 per 0.1 increase in HHI (assuming a linear effect; $p<0.001$) (Figure 5.3). Markets are more competitive in areas where independent sector treatment centres are active.

[89] See FN 64.

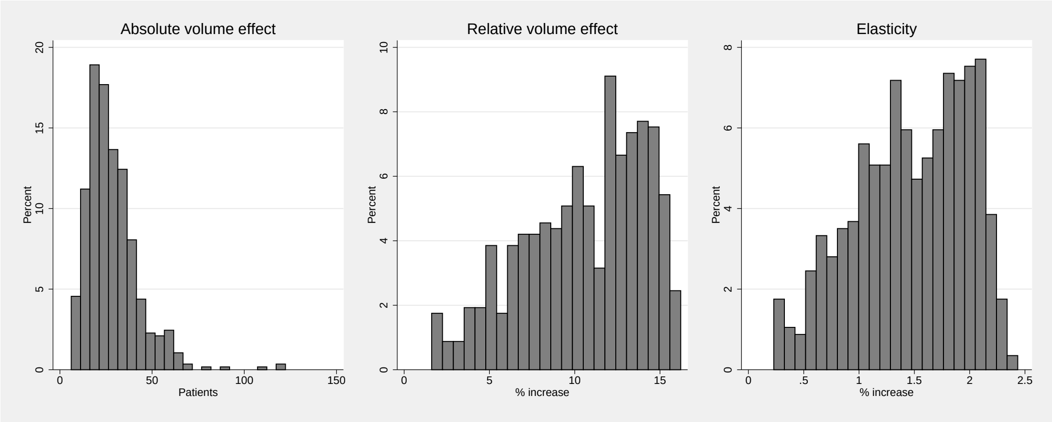
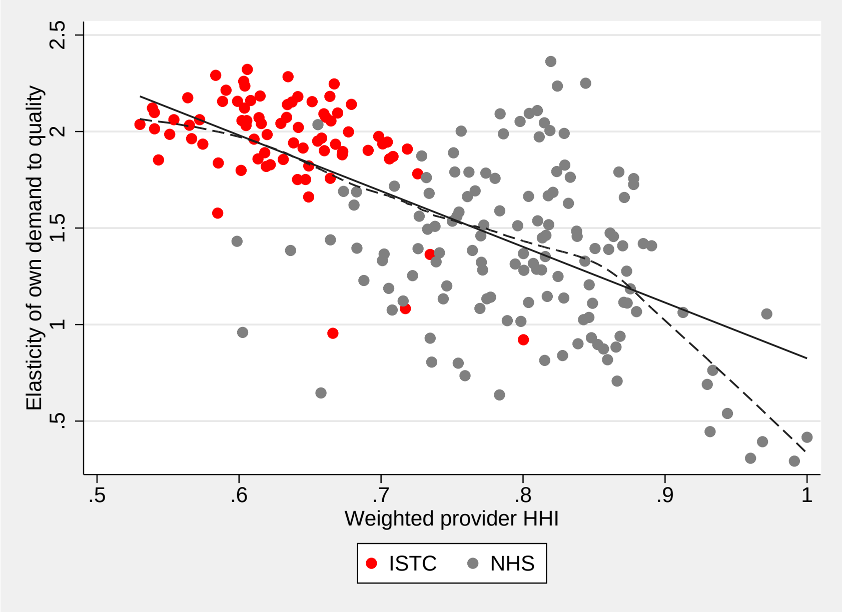


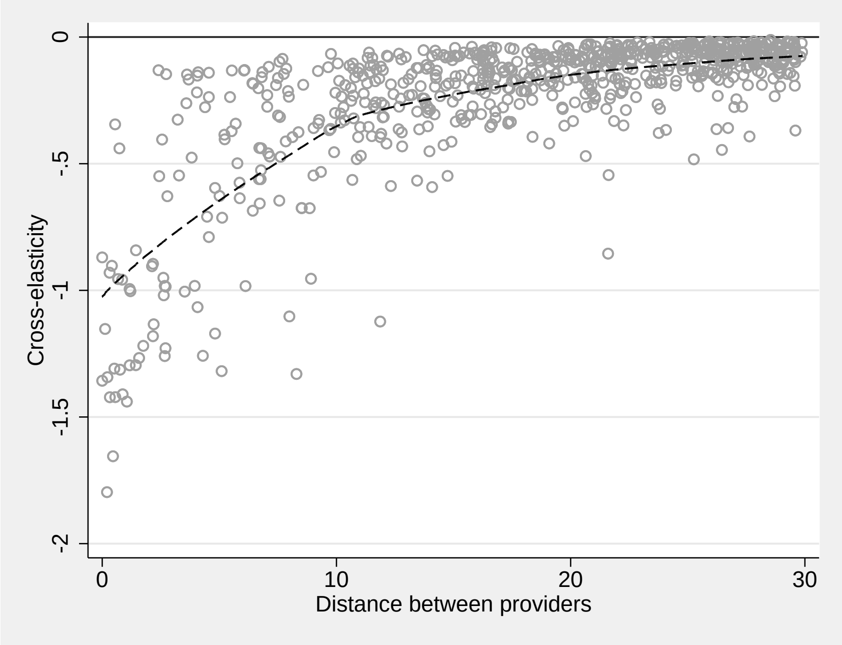
Figure 5.2: Distribution of changes in hospital demand as a result of a SD increase in Oxford Hip Score change scores and quality elasticity of demand



Notes: Solid line shows best linear fit (Intercept = 3.71 (SE=0.16), slope = -2.89 (SE=0.22), R^2 = 0.42). Dashed line shows LOWESS curve.

Figure 5.3: Differences in quality elasticity of demand between providers in competitive (low HHI) and non-competitive (high HHI) markets

We also examine the effect of changes in the quality of other providers on a provider's demand. Higher cross-quality demand elasticities make it more likely that increases in one provider's quality will trigger an increase in the quality of other providers. Figure 5.4 shows how cross-quality elasticities decline rapidly as the distances between providers increase. Whereas a 1% increase in a competitor's PROM quality is associated with a -0.63% reduction in demand if the competitor is located within 10 km, this reduces to -0.23% when the competitor is 30km away.



Notes: Dashed line shows LOWESS curve.

Figure 5.4: Percentage change in demand as a result of percentage change in competitor's quality

5.6 Discussion

The collection of patient-reported outcome measures has been introduced in England with the ambition that these new metrics of hospital quality would influence patient choice of hospital (Department of Health 2008a). This study is the first to test the relationship between observed hospital PROM quality and demand for elective hip replacement surgery. It uses data on observed choices for all NHS-funded patients treated between April 2010 and March 2013 in private and public hospitals in

England. In order to address potential endogeneity bias we implement an empirical strategy based on lagged explanatory variables, hospital fixed effects and a control group design based on demand for emergency hip replacement.

Our results suggest that elective hospital demand is statistically significantly associated with observed quality as measured by PROMs and other metrics. While individual patients are not very sensitive to quality differences — the estimated willingness to travel for a standard deviation increase in PROM quality is less than 1.3km — the number of potential patients in a hospital's market implies that the average hospital can attract an increase in elective activity of approximately 34 new patients, or 9% of existing activity levels, if it finds ways to improve PROM quality by one standard deviation. Hospital demand is more responsive to a one standard deviation of PROM quality than one standard deviation of emergency readmission rates, and there is no statistically significant association with mortality rates after hip replacement surgery.

Our findings that choice responds to quality suggest that providers could compete on quality to attract additional demand. However, the change in activity that would arise after a change in quality may be modest. First, a standard deviation increase in OHS (equivalent to 1.4 points) would be a substantial improvement in quality for any provider and difficult to achieve. For comparison, the average year-on-year improvement in hospital PROM scores is 0.196 OHS points, or less than 15% of the observed standard deviation. Second, we show that the effect of quality changes on the providers' ability to attract patients away from local competitors diminishes rapidly as distance increases. This may result in local quasi-monopolies where quality improvements have little effect on demand. Finally, our estimated effect is likely to be an upper bound estimate and our analysis on emergency patients suggests that the coefficient of demand to quality could be up to 30% smaller. Taken together, the incentive effect of patients 'voting with their feet' and demanding higher quality is likely to be limited. Of course, whether or not providers engage in quality competition based on published PROM scores depends primarily on whether they *perceive* their demand to be elastic to quality changes and on how much they value their reputation. We cannot answer these questions with our data.

There are several policy levers which may be used to ensure that PROM quality information is used to inform hospital choice (Marshall et al. 2004; Faber et al. 2009). Many patients may still not know about hospital PROM scores and more active dissemination to the general public may be required (e.g. by adding the information to the Choose & Book system). Some patients may find it difficult to access this information, for example if they do not have access to the internet. There is a lack of evidence on the extent to which patients and general practitioners are aware of this information and consider it as part of their decision-making process. Similarly, the information may not be sufficiently meaningful to them in its current format. A recent study by Hildon et al. (2012) showed that a high proportion of patients and doctors do not consider the reported PROMs to have an intuitive metric and thus struggle to interpret provider scores. Finally, some patients may not consider variation between hospitals sufficiently large to be considered important. Some of these points may resolve over time, whereas others require targeted policy intervention to improve the dissemination of quality information.

We also explore whether patient preferences vary according to observed and unobserved patient characteristics. We find that the preference for PROM quality increases with age and decreases with income deprivation, comorbidity burden and past utilisation. Qualitatively similar results are obtained for preferences for quality as approximated by emergency readmission rates. Interestingly, we do not find evidence that preferences for quality vary with pre-operative health status as reported by the patient herself. But because healthier patients are more willing to travel, they have ceteris paribus a higher willingness to travel for quality. Hence, the *'distance bias'* described by Gowrisankaran and Town (1999) is likely to occur not because more morbid patients request higher quality, but because they derive different disutility from travel. This finding may be specific to the condition under study as osteoarthritis and other conditions that require hip replacement reduce patients' mobility, and more severely morbid patients thus may be less able or willing to travel.

There remains scope for further research. For example, we cannot disentangle whether the estimated effect is driven by patients' choices versus general practitioners' choices acting on their behalf. We conjecture it is due to both. We also did not test whether the first release of PROM information in 2009/10 constituted news to patients and their agents and how this changed their behaviour. For example, analysing the effect of the public release of cardiovascular surgery report cards on New York hospitals' market share, Dranove and Sfekas (2008) show that the effect is larger if the signal about hospital quality contradicts prior beliefs, and that failure to account for prior beliefs may lead to downward biased estimates of the quality elasticity of demand. Because PROM scores have been collected and disseminated for all providers in England, there is no natural control group to isolate the causal effect of information release. Finally, our findings may be specific to the condition under study. Patients undergoing surgery with considerable risk of peri-operative mortality may be more sensitive to quality information since the cost of choosing an inferior provider would likely be more significant (see e.g. Gaynor et al. 2012).

In conclusion, the results reported in this chapter provide some first evidence to suggest that hospital demand for hip replacement responds to hospital quality as captured by changes in patient-reported health status.

Acknowledgements

We are grateful for comments and suggestions from Carsten Crede, Andrew Street, Karen Bloor, Richard Cookson and those received during presentations at the Universities of Birmingham, Newcastle, York and Melbourne, the King's Fund, the 5th Network of Industrial Economists Doctoral Colloquium (Nottingham), and the Leeds 2015 meeting of the Health Economic Study Group. This research was funded by the Department of Health in England under the Policy Research Unit in the Economics of Health and Social Care Systems (Ref 103/0001). The views expressed are those of the authors and not necessarily those of the funders.

Hospital Episode Statistics are copyright ©2015, re-used with the permission of

6 Conclusions

This thesis presents four empirical studies that explore the use of performance measures based on changes in patient self-reported health status and HRQoL to assess the costs and quality of care provided by hospitals in the English NHS. Such information is useful to principals (e.g. patients, regulators, purchasers of care) that want to incentivise and hold to account their agents (here: hospital providers) so as to reduce the potential for rent extraction. In what follows, I shall first briefly summarise the main findings of these analyses and discuss policy implications, and then make suggestions for further research.

6.1 Summary of key findings and implications for policy

In Chapter 2 we explore the empirical relationship between hospitals' costs and quality of care, here measured by provider mean changes in patients' PROM scores, for four surgical procedures. Healthcare providers often argue that resource utilisation increases in quality, i.e. one needs to invest more to get better care. Regulators are typically less informed about the production process and thus cannot assess these claims. Our analysis provides little empirical evidence to support the notion that quality is necessarily costly. Indeed, assuming a linear relationship we estimate that higher case-mix adjusted costs are generally associated with lower quality, although most of these estimates are not statistically significant. We find some evidence of a non-linear, U-shaped relationship between case-mix adjusted costs and changes in PROMs for hip replacement patients, as previously found by e.g. Hvenegaard et al. (2011) in different contexts. However, controlling for the extra costs of quality only has a small effect on providers' estimated relative costs. Hence, hospitals (i.e.

agents) may be trying to exploit their information advantage over purchasers and regulators of care (i.e. principals) when claiming that cost variation is the result of quality variation, not rent extraction.

There are two main implications for policy. First, even nearly a decade[90] after the introduction of Payment by Results (PbR), there remains substantial variation in resource utilisation amongst providers of the same care. This is not readily explained by a differential effectiveness of the treatment provided, as measured through changes in patients' self-reported health, or by differences in observable patient or provider characteristics that can be assumed exogenous to the provider. This raises questions about the ability of the current reimbursement system to incentivise cost containment and, over time, standardise resource utilisation. Farrar et al. (2009) have shown that the introduction of PbR was associated with a decrease in length of stay on average. However, they did not explore whether resource use has become more standardised after the introduction of PbR. More generally, there is a lack of longitudinal analyses to establish whether hospital costs are still converging or whether PbR has already 'lost its bite'. Given the substantial variation in reported reference costs, one would expect the latter.

Second, if cost and quality are negatively related - for at least some providers and levels of quality - this implies scope to improve the efficiency of the service. This raises the question why (semi-)altruistic providers have not yet amended their care processes to achieve better health outcomes at lower costs. One possible explanation is that providers are not aware of best practice. Another explanation is that the immediate costs of service re-design outweigh the perceived short- to medium-term benefits (Smith 2015). Addressing this might require a different funding model than simply paying per unit of activity, or may not be solvable solely through market-based incentives. If so, policy makers would be well advised to combine incentives to reduce costs and/or improve quality with information for providers on how to do so

[90] The PbR system was rolled out to all elective procedures and all NHS trusts at the beginning of the financial year 2005/6 (Department of Health 2011a). However, for a number of elective interventions, including hip replacement surgery, the PbR system had already been implemented in the financial year 2003/4.

without harming service provision.

In Chapter 3 we discuss the appropriate derivation of multidimensional performance measures to assess provider performance, which might be particularly important in the context of multiple stakeholders. Specifically, we propose a methodology to analyse EQ-5D data to inform prospective patients and local managers about the relative performance of hospital providers in improving different aspects of their patients' health. We argue that, in this specific setting, analysing each of the five dimensions independently is more appropriate than the current practice of analysing EQ-5D utility scores. This is because the EQ-5D utility scores are based on average preferences of the UK general population and these may differ from those of individual patients. Also, disaggregated information may be more useful for local managers to identify problems in the care process. Our empirical analysis shows that provider variation in outcomes is more pronounced on those EQ-5D dimensions that receive low weights in the UK general population tariff. Performance estimates based on utility scores may therefore understate between-provider differences since variability on dimensions with low weights will feature less prominently than variability on dimensions with higher weights.

In constructing and publishing composite performance scores, policy makers and those responsible for the public dissemination of such data should give more thought to the role that value sets play therein. For example, the rationale for using the UK general population preferences for constructing EQ-5D utility scores is well recognised in the context of technology adoption into the reimbursement catalogue of tax-funded healthcare systems (Siegel et al. 1997; Brazier et al. 2005). But this does not mean that the same rationale applies when these data are used to compare hospitals and inform a wide range of stakeholders. An intermediate solution to the two extremes contrasted in this thesis (composite scores based on general population preferences vs. no aggregation) may be to elicit the preferences of the relevant population of decision-makers, in this context groups of prospective patients. Such an approach would still involve averaging across individuals and, therefore, might lead to mismatch between patients' own preferences and those reflected in the composite score. However, patients undergoing the same procedure are likely to be more homogeneous (e.g. in terms of age, mobility requirements, and expectations) and the mismatch may therefore be smaller. To avoid costly elicitation exercises, policy makers may want to draw on the existing data that has been collected as part of the national PROM survey. For example, the PROM survey also collects VAS data alongside the EQ-5D health profile and these two datasets could be mapped to obtain (non-utility) weights.[91]

The approach advocated in Chapter 3 has a number of attractive features, not least that it makes no assumptions about patients' preferences regarding performance on the different health domains and is therefore consistent with general welfare theory. However, a drawback is that it results in multiple performance statements, which patients and other recipients of information may find difficult to comprehend or synthesise. To overcome this issue, in Chapter 4 we propose the use of dominance criteria to identify providers that excel or perform poorly across all relevant performance dimensions simultaneously. Dominance criteria require only weak assumptions about the preferences of the relevant information recipients and are therefore consistent with the normative arguments put forward in Chapter 3.

We demonstrate the feasibility and utility of this approach in Chapter 4. Specifically, we study the multidimensional performance of providers of hip replacement surgery with respect to length of stay, readmission rates, waiting times and changes in patients' health status. We find that all providers identified as dominant are privately operated independent sector treatment centres (ISTCs), whereas all those dominated by the benchmark are NHS trusts. We also find evidence of a statistically significant negative association between length of stay and patient outcomes, somewhat analogous to the findings of Chapter 2. These results should be understood as a starting point for further in-depth analysis of why those identified privately operated ISTCs produce excellent results across all performance dimensions, and in how far their best practices could be transferred to NHS hospitals.

[91] Greiner et al. (2003) use EQ-5D health profile and VAS data from eleven population surveys in six Western European countries to calculate a European tariff.

In Chapter 5 we explore the use of PROM-based performance information by patients choosing hospitals. To this end, we estimate a hospital choice model for all NHS patients undergoing hip replacement surgery in England between April 2010 and March 2013. We find that patients are more likely to choose a hospital with better PROM scores, and that this finding is robust to a number of alternative specifications. However, the effect of quality on individual choices is small, which is consistent with the existing literature on consumer choice in healthcare (see Brekke et al. (2014) and examples cited therein). However, because the market for hip replacement surgery is large, providers can still attract greater demand if they find ways to improve outcomes more than their competitors. Public release of performance information may thus stimulate quality competition. Yet, providers' ability to attract patients away from competitors diminishes rapidly with distance. This may result in local quasi-monopolies in which the public release of performance information may incentivise providers to improve their quality primarily through concerns about their reputation, not fears over loss of activity.

If the public release of performance information is only effective in stimulating quality competition in some regions of the country (i.e. urban areas with many competitors) or for some specific conditions, market based incentive mechanism may be insufficient to motivate providers in other contexts. As patients seem unwilling to travel far for better care, this may contribute to inequalities in population health. Hence, policy makers and regulators may want to consider other, non-market based mechanisms to ensure that all providers strive to deliver high quality care. The newly introduced best practice tariff (BPT) for elective hip and knee replacement rewards providers on the basis of their relative effect on patients' health (Monitor and NHS England 2013), and may be one of many alternative financial and non-financial vehicles to improve quality.

Finally, I would like to draw attention to two general issues that are of relevance to policy makers and those administering the national PROM programme. The first issue is data quality. For PROMs to become a credible indicator of hospital performance, analyses must be based on a sufficiently large number of responses to reduce the risk of selection bias. Throughout this thesis, we have given little attention to the issue of missing data and associated biases. However, in other work we have explored the impact of missing data on provider performance estimates and found those to be robust to non-response (Gomes et al. 2015). Nevertheless, low participation rates and non-response may undermine the credibility of PROM-based quality indicators if providers perceive them as non-representative or imprecise. The decision to link bonus payments to participation rates in the orthopaedic BPT is commendable, although we have cautioned elsewhere that the requirement of 50% participation may not be sufficiently high to motivate providers to put enough effort into data collection (Gutacker, Street et al. 2015).

The second issue is the lack of evidence on the cost-effectiveness of collecting and disseminating comparative PROM data. Our analyses show that PROM data can be used to assess the quality performance of providers in routine care settings. The national PROMs programme is currently focussing on less than 4% of all elective hospital activity[92] in the English NHS, so there is scope to roll it out to other areas. However, in doing so, policy makers must consider whether the benefits of collecting and disseminating performance data outweigh the cost of collection. While the direct costs can be quantified with relative ease — Maynard and Bloor (2010) report unit costs of approximately £6.50 — the benefits to patients and the public, e.g. in the form of better information on provider quality or reassurance of fitness to practice, and other indirect costs have not yet been rigorously assessed. The research reported on in Chapter 5 gives some first insights into the potential benefit of collecting and disseminating PROM data, but more efforts are required to establish the cost-effectiveness of collection and public reporting of performance data.

[92] Based on primary procedure code and data for FY2013/14 (Health & Social Care Information Centre 2015).

6.2 Suggestions for further research

The research presented in this thesis can be extended in several directions. I would like to highlight four areas of further research that I consider especially fruitful. In some cases, this will require combining methodologies from the fields of regulation and economic evaluation.

First, our finding that better patient health outcomes are associated with lower resource utilisation requires further investigation. One area of concern is potential endogeneity between both. This may arise due to a number of mechanisms, including unobserved confounding or simultaneity. The analysis in Chapter 4 is more robust to confounding than previous studies in this area due to the availability of good pre-operative health status information and adjustment for selection into hospital. Also, other sources of evidence, like those from evaluations of enhanced recovery pathways, support our empirical findings (Husted et al. 2008; Larsen et al. 2008; Paton et al. 2014). However, further econometric analyses based on suitable instrumental variables would be useful. More generally, research is required to understand the relationship between cost and quality for more conditions and identify the underlying factors that are amenable to policy, both to set an informed benchmark and help providers achieve it. In many cases, econometric analysis will be able to identify highly or poorly performing providers[93] but further qualitative research in those institutions will be required to understand how resources are utilised to their best effect.

Second, the available PROM data collected as part of the national PROM programme allows the assessment of short term (i.e. three or six months) health benefits but still falls short of the ambition to measure changes in patients' health trajectories over time. The National Joint Registry has begun to collect follow-up PROM data at one, three and five years after surgery (National Joint Registry 2011). These data should prove useful in understanding the longer term effects that providers have on their patients' health. They may also help to alleviate concerns that the

[93] For example by using the methodology developed in Chapter 4.

follow-up period in the national PROM programme is too short to capture all relevant benefits (Browne et al. 2013). However, such information will be of limited use for performance management purposes since it cannot be used to detect and respond to substandard care in a timely fashion. Extrapolation techniques within a modelling framework may be more suitable for this purpose but it remains to be seen how precise such predictions would be and whether they would be accepted by providers as reliable measures of their performance.[94]

Third, I pointed out in Chapter 1 that performance information can be used to inform future contracts. Yet, in practice such information is mainly used retrospectively, e.g. to adjust payments according to observed outcomes or challenge providers about their quality of care. If performance information is to be used by purchasers of care to determine which provider to contract with in the future, past performance needs to be predictive of future performance. Leckie and Goldstein (2009) examine the predictive ability of school league tables and found past performance to be largely unrelated to current performance. With respect to PROMs, Varagunam et al. (2014) found poor to moderate agreement between providers' performance classifications (better than expected, as expected, worse than expected) over time for hip and knee replacement surgery, and low or non-existent agreement for the two other procedures. However, it should be noted that Varagunam et al. (2014)'s analysis did not exploit the longitudinal nature of the data or adjust for regression to the mean (Jones and Spiegelhalter 2009). More research is required into the intertemporal stability of provider performance estimates to inform policy makers about their suitability for prospective contracting purposes and prospective patients about their utility for informing hospital choice.

Finally, the orthopaedic BPT, introduced in April 2014, will provide an excellent opportunity to evaluate the cost-effectiveness of using financial incentives to improve patients health outcomes. Providers will receive an 11% bonus on top of the tariff

[94] Conversely, such information has proven useful in establishing the cost-effectiveness of different procedures in routine care settings. See Coronini-Cronberg et al. (2013) for an example in general surgery.

for hip and knee replacement surgery if they do not perform statistically significantly worse than the national average and achieve at least 50% participation in the pre-operative PROM survey (Monitor and NHS England 2013; Gutacker, Street et al. 2015). Future research could evaluate the overall effect of this high-powered incentive scheme on patient outcomes and contrast its effectiveness in motivating previously highly and poorly performing providers. Other interesting aspects include the incentive for providers to limit PROM survey participation — since this reduces the probability to be detected as performing unsatisfactorily for a given level of statistical significance — and to engage in patient selection if the case-mix adjustment is perceived to be incomplete.

In summary, the data collected as part of the national PROMs programme have given me the opportunity to explore variation in patients' health outcomes following treatment as a suitable performance indicator to assess and incentivise hospital providers. The findings are encouraging but more work is required to make the best use of these data to reduce information asymmetries and ensure more effective use of resources in the English NHS.

7 Appendices

7.1 Appendix to Chapter 2

Table A2.1: Effect of health gain on costs under different GLM specifications

Measure of health gain	Log / Gamma		Log / Poisson		Identity / Gaussian	
	Est	SE	Est	SE	Est	SE
Knee replacement						
EQ-5D	-13.0	20.8	-14.8	19.6	-15.3	20.0
EQ-VAS	-80.4	37.1*	-78.9	36.8*	-77.3	36.4*
OKS	-60.8	57.9	-61.8	55.4	-64.3	56.2
Hip replacement						
EQ-5D	-15.5	22.7	-18.4	24.9	-19.3	25.5
EQ-VAS	-52.2	27.8	-51.8	30.1	-51.4	30.3
OHS	26.2	62.8	30.5	68.6	28.4	69.5
Groin hernia repair						
EQ-5D	-6.1	12.7	-6.6	12.5	-7.4	12.8
EQ-VAS	-15.2	17.6	-13.8	17.1	-14.1	16.9
Varicose vein surgery						
EQ-5D	16.7	9.4	12.9	7.7	11.9	6.4
EQ-VAS	35.7	15.9*	32.6	14.3*	32.0	13.4*
AVVQ	52.3	29.7	45.0	25.8	44.5	24.5

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Notes: All effects are marginal effects on the untransformed scale, i.e. they express changes in costs for a unit change in health gain. Specification of linear predictor mirrors that of models reported in Table 2.3. However, the models reported here do not account for a provider random effect due to convergence problems. Standard errors are clustered at provider level.

Table A2.2: Comparison of fixed and random effects estimates - knee replacement

	Fixed effects			Random effects			Difference in estimates	
Variable	Est	SE		Est	SE		Absolute	Relative
Intercept	5817.09	18.29	***	5909.60	121.43	***	-92.51	-1.59%
Patient aged 61-66	24.63	16.46		24.47	16.47		0.16	0.63%
Patient aged 67-72	-10.14	16.06		-10.35	16.07		0.21	-2.04%
Patient aged 73-77	14.30	16.82		14.09	16.83		0.22	1.51%
Patient aged >77	71.55	16.73	***	71.21	16.73	***	0.34	0.48%
Male patient	-11.39	10.13		-11.41	10.13		0.02	-0.16%
Admission: Transferred in from another provider	1716.19	113.01	***	1715.27	113.04	***	0.91	0.05%
Discharge: Other	127.02	37.68	***	127.90	37.68	***	-0.88	-0.70%
Discharge: Died in hospital	-353.97	126.40	**	-353.39	126.44	**	-0.58	0.16%
Discharge: Transferred out to another provider	324.02	39.13	***	324.02	39.14	***	0.00	0.00%
Under the care of multiple consultants	-124.41	37.14	***	-125.21	37.15	***	0.80	-0.65%
Number of secondary procedure codes	164.31	6.84	***	164.32	6.84	***	0.00	0.00%
Number of secondary diagnosis codes	24.63	3.01	***	24.66	3.01	***	-0.04	-0.15%
Weighted Charlson comorbidity index	-3.20	7.94		-3.16	7.94		-0.04	1.25%
Income deprivation - 1st quintile (lowest)	5.34	16.87		4.83	16.88		0.51	9.60%
Income deprivation - 2nd quintile	7.55	16.37		6.99	16.37		0.56	7.37%
Income deprivation - 3rd quintile	30.70	15.94		30.04	15.94		0.66	2.15%
Income deprivation - 4th quintile	2.42	16.11		1.82	16.11		0.60	24.66%
HRG: HB21A	2551.06	23.86	***	2552.48	23.87	***	-1.43	-0.06%
HRG: HB21B	418.57	23.39	***	418.80	23.39	***	-0.23	-0.05%
HRG: HB23C	-2982.70	36.30	***	-2983.13	36.31	***	0.43	-0.01%
HRG: HR05Z	275.60	70.50	***	275.37	70.52	***	0.23	0.08%
HRG: HB23B	-1450.01	73.44	***	-1450.61	73.46	***	0.60	-0.04%
HRG: HR06A	-691.33	75.86	***	-691.43	75.88	***	0.09	-0.01%
HRG: HR04C	800.19	77.93	***	800.68	77.96	***	-0.49	-0.06%
HRG: HR22C	-2941.17	81.64	***	-2940.31	81.67	***	-0.87	0.03%
HRG: HB22B	1662.69	95.86	***	1662.51	95.89	***	0.17	0.01%
Main diagnosis: Rheumatoid arthritis	-74.10	28.05	**	-74.95	28.06	**	0.86	-1.15%
Main diagnosis: other	-73.74	79.70		-74.42	79.73		0.68	-0.92%
Procedure: Total knee replacement (revision)	-190.49	44.08	***	-190.19	44.09	***	-0.30	0.16%
Procedure: Unicompartmental replacement (primary)	-582.51	49.00	***	-581.94	49.01	***	-0.56	0.10%
Procedure: Unicompartmental replacement (revision)	-1745.90	172.57	***	-1746.09	172.62	***	0.20	-0.01%
Procedure: Hybrid prosthetic replacement (primary)	-14.61	79.07		-14.55	79.09		-0.06	0.42%
Procedure: Hybrid prosthetic replacement (revision)	-441.28	274.64		-441.08	274.72		-0.20	0.04%
Hausman test statistic	79.66		***					
Sample size	60,780			60,780				

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Notes: Model differs from that reported in Table 2.3 in that provider characteristics are not modelled.

Table A2.3: Comparison of fixed and random effects estimates - hip replacement

	Fixed effects			Random effects			Difference in estimates	
Variable	Est	SE		Est	SE		Absolute	Relative
Intercept	5706.86	33.67	***	5709.74	118.76	***	-2.89	-0.05%
Patient aged 60-66	18.05	19.41		17.59	19.41		0.46	2.52%
Patient aged 67-72	53.44	19.17	**	52.90	19.18	**	0.54	1.02%
Patient aged 73-77	42.57	19.94	*	42.03	19.95	*	0.54	1.28%
Patient aged >77	82.68	19.61	***	81.95	19.61	***	0.73	0.89%
Male patient	-7.49	12.16		-7.50	12.17		0.02	-0.23%
Admission: Transferred in from another provider	1490.31	74.64	***	1490.71	74.68	***	-0.40	-0.03%
Discharge: Other	229.21	37.08	***	231.56	37.09	***	-2.35	-1.03%
Discharge: Died in hospital	-170.86	121.75		-171.05	121.81		0.19	-0.11%
Discharge: Transferred out to another provider	144.95	39.05	***	145.19	39.06	***	-0.23	-0.16%
Under the care of multiple consultants	-353.45	37.56	***	-353.59	37.58	***	0.14	-0.04%
Number of secondary procedure codes	175.34	6.92	***	175.59	6.92	***	-0.25	-0.14%
Number of secondary diagnosis codes	46.27	3.48	***	46.28	3.49	***	-0.01	-0.03%
Weighted Charlson comorbidity index	-16.22	8.62		-16.30	8.62		0.08	-0.49%
Income deprivation - 1st quintile (lowest)	-8.49	22.33		-9.03	22.33		0.55	-6.45%
Income deprivation - 2nd quintile	-6.77	18.91		-7.31	18.91		0.53	-7.84%
Income deprivation - 3rd quintile	16.51	18.97		15.91	18.97		0.60	3.65%
Income deprivation - 4th quintile	17.89	18.75		17.45	18.76		0.43	2.42%
HRG: HB12C	566.80	51.59	***	567.50	51.61	***	-0.70	-0.12%
HRG: HR05Z	518.44	28.88	***	519.55	28.89	***	-1.12	-0.22%
HRG: HB12B	2170.62	30.41	***	2171.47	30.43	***	-0.85	-0.04%
HRG: HB12A	459.69	54.98	***	461.14	55.01	***	-1.45	-0.32%
HRG: HR04C	886.68	42.83	***	887.23	42.84	***	-0.54	-0.06%
HRG: HB11C	3720.61	75.05	***	3722.72	75.08	***	-2.11	-0.06%
HRG: HR04B	5192.20	81.98	***	5192.11	82.02	***	0.09	0.00%
HRG: HA13C	-840.66	145.96	***	-840.01	146.03	***	-0.65	0.08%
HRG: HB13Z	-1843.20	134.81	***	-1839.50	134.87	***	-3.70	0.20%
HRG: other	260.35	60.51		260.49	60.53		-0.14	-0.05%
Main diagnosis: Rheumatoid arthritis	-53.77	28.44		-54.61	28.45		0.84	-1.56%
Main diagnosis: other	-49.33	79.89		-49.99	79.92		0.66	-1.34%
Procedure: Total hips replacement (revision)	33.08	51.22		31.92	51.24		1.16	3.50%
Procedure: Total prosthetic replacement (primary)	177.94	102.80		178.01	102.85		-0.08	-0.04%
Procedure: Total prosthetic replacement (revision)	-434.52	111.29	***	-431.39	111.34	***	-3.12	0.72%
Procedure: Hybrid prosthetic replacement (primary)	9.52	20.40		9.42	20.40		0.10	1.05%
Procedure: Hybrid prosthetic replacement (revision)	-383.84	74.73	***	-384.92	74.77	***	1.08	-0.28%
Procedure: Other hip replacement (primary)	-1363.55	610.28	*	-1365.21	610.55	*	1.65	-0.12%
Procedure: Other hip replacement (revision)	-2885.47	365.62	***	-2885.86	365.79	***	0.39	-0.01%
Hausman test statistic	84.91		***					
Sample size	53,235			53,235				

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Notes: Model differs from that reported in Table 2.3 in that provider characteristics are not modelled.

Table A2.4: Comparison of fixed and random effects estimates - hernia repair

	Fixed effects			Random effects			Difference in estimates	
Variable	Est	SE		Est	SE		Absolute	Relative
Intercept	1288.54	11.31	***	1326.17	32.78	***	-37.63	-2.92%
Patient aged 44-56	13.45	7.17		13.34	7.18		0.11	0.81%
Patient aged 57-65	28.78	7.12	***	28.71	7.12	***	0.07	0.24%
Patient aged 66-74	64.50	7.37	***	64.50	7.38	***	0.00	0.00%
Patient aged >74	147.01	7.67	***	146.97	7.68	***	0.04	0.03%
Male patient	-22.78	8.49	**	-22.77	8.50	**	-0.01	0.06%
Admission: Transferred in from another provider	660.87	70.14	***	660.39	70.16	***	0.48	0.07%
Discharge: Other	377.00	36.80	***	376.42	36.81	***	0.57	0.15%
Discharge: Died in hospital	751.35	100.55	***	751.47	100.58	***	-0.12	-0.02%
Discharge: Transferred out to another provider	513.41	68.07	***	513.25	68.09	***	0.15	0.03%
Under the care of multiple consultants	206.41	25.79	***	206.50	25.80	***	-0.10	-0.05%
Number of secondary procedure codes	78.41	4.17	***	78.31	4.17	***	0.10	0.13%
Number of secondary diagnosis codes	33.93	2.05	***	33.98	2.05	***	-0.04	-0.13%
Weighted Charlson comorbidity index	54.98	4.42	***	54.92	4.42	***	0.06	0.11%
Income deprivation - 1st quintile (lowest)	-16.02	7.45	*	-16.74	7.45	*	0.73	-4.54%
Income deprivation - 2nd quintile	-8.68	8.02		-9.28	8.02		0.60	-6.90%
Income deprivation - 3rd quintile	-7.67	6.94		-8.16	6.94		0.49	-6.44%
Income deprivation - 4th quintile	-4.92	6.99		-5.25	6.99		0.33	-6.76%
HRG: FZ18B	85.14	7.85	***	85.11	7.86	***	0.04	0.04%
HRG: FZ18A	304.69	15.17	***	306.01	15.17	***	-1.33	-0.44%
HRG: FZ18D	-84.38	23.92	***	-84.38	23.93	***	0.00	-0.01%
HRG: FZ19Z	718.74	107.91	***	718.24	107.94	***	0.49	0.07%
HRG: FZ17C	427.21	38.42	***	429.70	38.43	***	-2.49	-0.58%
HRG: FZ12C	114.52	40.31	**	114.77	40.33	**	-0.25	-0.22%
HRG: LB34B	-257.09	42.07	***	-257.47	42.08	***	0.38	-0.15%
HRG: FZ17B	857.68	63.15	***	857.39	63.17	***	0.28	0.03%
HRG: FZ12A	2586.45	68.98	***	2589.58	69.00	***	-3.13	-0.12%
HRG: other	2078.32	22.05		2078.66	22.04		-0.34	-0.02%
Procedure: Exicision of inguinal hernia sac	-920.12	104.20		-919.90	104.24		-0.22	0.02%
Procedure: Inguinal hernia surgery	60.22	8.35		60.29	8.35		-0.07	-0.11%
Procedure: Femoral hernia surgery	4.98	15.56		4.98	15.56		0.00	0.04%
Procedure: Incisional hernia surgery (OPCS: T25)	23.35	47.57		23.37	47.59		-0.02	-0.07%
Procedure: Incisional hernia surgery (OPCS: T26)	1134.40	129.60		1143.91	129.62		-9.51	-0.84%
Laprascopic surgery	-57.79	7.21		-57.60	7.21		-0.18	0.32%
Hausman test statistic	69.91		***					
Sample size	57,343			57,343				

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

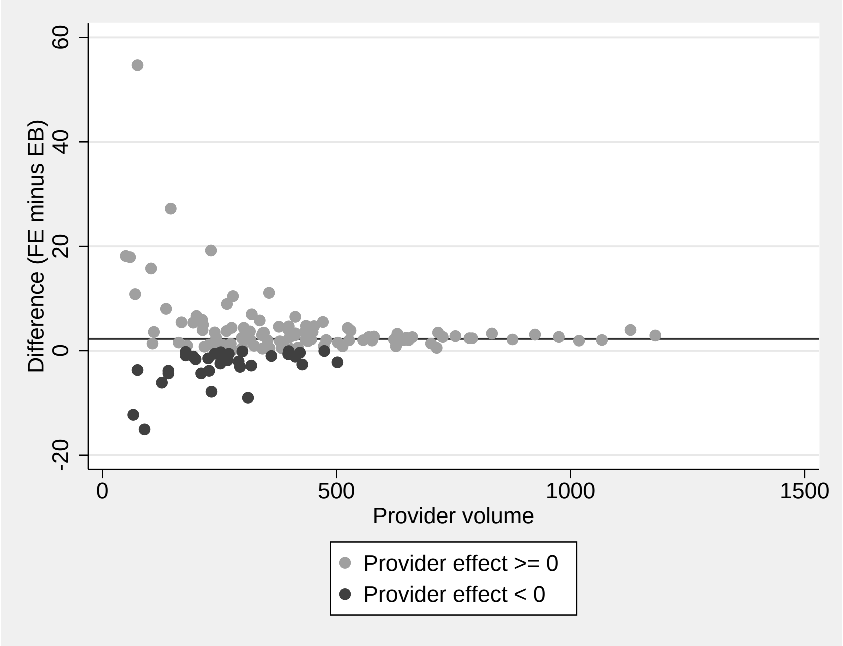
Notes: Model differs from that reported in Table 2.3 in that provider characteristics are not modelled.

Table A2.5: Comparison of fixed and random effects estimates - varicose vein surgery

	Fixed effects			Random effects			Difference in estimates	
Variable	Est	SE		Est	SE		Absolute	Relative
Intercept	1178.70	9.38	***	1175.41	36.28	***	3.29	0.28%
Patient aged 37-44	20.78	8.12	*	20.78	8.12	*	0.00	0.00%
Patient aged 45-53	10.40	8.12		10.47	8.12		-0.07	-0.72%
Patient aged 54-63	23.55	8.11	**	23.63	8.11	**	-0.07	-0.31%
Patient aged >63	32.10	8.50	***	32.27	8.50	***	-0.18	-0.55%
Male patient	21.00	5.20	***	21.12	5.20	***	-0.11	-0.54%
Admission: Transferred in from another provider	43.54	126.89		43.45	126.91		0.09	0.21%
Discharge: Other	124.24	67.60		123.07	67.61		1.17	0.94%
Discharge: Died in hospital	-	-		-	-		-	-
Discharge: Transferred out to another provider	-362.97	269.15		-363.91	269.17		0.94	-0.26%
Under the care of multiple consultants	295.66	57.62	***	296.06	57.63	***	-0.41	-0.14%
Number of secondary procedure codes	10.89	3.36	**	10.85	3.36	**	0.04	0.39%
Number of secondary diagnosis codes	22.80	3.26	***	22.77	3.26	***	0.03	0.15%
Weighted Charlson comorbidity index	-2.05	8.84		-2.02	8.84		-0.03	1.38%
Income deprivation - 1st quintile (lowest)	-25.32	8.67	**	-25.58	8.67	**	0.26	-1.03%
Income deprivation - 2nd quintile	-9.39	8.53		-9.46	8.53		0.07	-0.78%
Income deprivation - 3rd quintile	-11.59	8.09		-11.77	8.09		0.18	-1.55%
Income deprivation - 4th quintile	-14.51	8.10		-14.50	8.10		0.00	0.02%
HRG: QZ09B	290.57	8.62	***	290.41	8.62	***	0.15	0.05%
HRG: QZ10A	2.31	12.30		2.36	12.30		-0.05	-2.29%
HRG: QZ08B	160.88	11.09	***	160.94	11.09	***	-0.07	-0.04%
HRG: QZ09A	402.59	25.23	***	402.51	25.23	***	0.08	0.02%
HRG: QZ07B	427.10	24.90	***	427.10	24.91	***	0.01	0.00%
HRG: QZ08A	203.97	34.14	***	203.69	34.14	***	0.28	0.14%
HRG: QZ05B	-48.64	67.58		-50.10	67.59		1.47	-3.02%
HRG: QZ07A	677.48	76.49	***	677.39	76.49	***	0.10	0.01%
HRG: QZ02A	3467.88	196.57	***	3470.28	196.59	***	-2.40	-0.07%
HRG: other	165.80	91.46		165.22	91.44		0.57	0.35%
Main diagnosis: Varicose ulcer	83.71	18.65	***	83.66	18.65	***	0.06	0.07%
Main diagnosis: Varicose vein with inflammation	-1.61	14.64		-1.31	14.63		-0.30	18.57%
Main diagnosis: Varicose vein with inflammation and ulcer	5.20	43.98		5.43	43.98		-0.22	-4.26%
Main diagnosis: other	15.09	45.71		14.38	45.71		0.71	4.69%
Procedure: Radiofrequency ablation	-110.92	10.42		-111.37	10.41		0.45	-0.41%
Procedure: Endovenous laser treatment of long saphenous vein	-103.67	8.90		-103.88	8.90		0.20	-0.20%
Procedure: Subfascial endoscopic perforator surgery	-47.19	46.45		-47.63	46.45		0.44	-0.93%
Hausman test statistic	38.61							
Sample size	23,077							

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Notes: Model differs from that reported in Table 2.3 in that provider characteristics are not modelled.



Notes: Graph shows difference between fixed effects (FE) and empirical Bayes (EB) estimates of provider effects from analysis of costs of hip replacement surgery (adjusted for patient characteristics). Effects are shrunken towards weighted average of provider effects, i.e. the global average.

Figure A2.1: Difference between shrunken and not shrunken provider performance estimates

Table A2.6: Relationship between health outcome and costs - provider-level constraints not modelled

	Constant MCQ		Non-constant MCQ				
Variable	H		H		H^2		Test of joint significance
	Estimate	SE	Estimate	SE	Estimate	SE	$\chi^2(2)$
Knee replacement							
EQ-5D	-8.4	21.9	25.7	30.4	-0.7	0.7	1.22
EQ-VAS	-74.2	36.2 *	-113.3	46.9 *	5.5	4.6	6.05 *
OKS	-54.7	67.2	343.7	217.8	-14.3	7.4	4.96
Hip replacement							
EQ-5D	-22.0	31.6	-321.7	154.5 *	3.8	2.0	5.16
EQ-VAS	-47.4	40.5	-109.0	59.5	4.1	3.6	3.53
OHS	31.0	84.8	-2362.5	759.3 **	61.0	19.5 **	9.78 **
Groin hernia repair							
EQ-5D	-8.3	13.7	-67.3	32.8 *	3.4	1.8	4.21
EQ-VAS	-11.4	16.5	-10.2	17.5	1.2	3.7	0.67
Varicose vein surgery							
EQ-5D	-3.1	9.2	3.6	19.5	-0.3	0.8	0.41
EQ-VAS	14.2	11.7	22.6	10.6 *	4.3	1.8 *	10.2 **
AVVQ	9.1	18.7	145.0	66.6 *	-7.8	3.8 *	4.76

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$
Standard errors are robust to heteroscedasticity.
OHS/OKS = Oxford Hip/Knee Score; AVVQ = Aberdeen varicose vein questionnaire; MCQ = marginal costs of quality; H = health outcome, i.e. the change in health status after treatment

Table A2.7: Relationship between health outcome and costs - excluding specialised orthopaedic providers

	Constant MCQ		Non-constant MCQ				
Variable	H		H		H^2		Test of joint significance
	Estimate	SE	Estimate	SE	Estimate	SE	$\chi^2(2)$
Knee replacement							
EQ-5D	-6.4	22.8	33.1	31.7	-0.8	0.7	1.60
EQ-VAS	-78.9	37.1 *	-117.6	44.8 **	5.4	4.6	7.13 *
OKS	-55.7	67.2	388.6	230.5	-15.8	7.8 *	5.18
Hip replacement							
EQ-5D	-23.1	32.3	-310.6	166.7	3.6	2.1	4.16
EQ-VAS	-57.3	39.8	-105.6	59.1	3.2	3.5	3.65
OHS	37.0	88.1	-2,450.9	779.1 **	63.2	20.2 **	9.90 **

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$
Standard errors are robust to heteroscedasticity.
OHS/OKS = Oxford Hip/Knee Score; MCQ = marginal costs of quality; H = health outcome, i.e. the change in health status after treatment

Table A2.8: Provider effects and performance assessment by PROM instrument and specification

		Provider effects[a,b]		Change in provider effects			Magnitude of	
					Stat. sign.		adjustment[a,c]	
PROM	MCQ	Min	Max	Stat. insign.	Econ. Insign.	Econ. sign.	Min	Max
Knee replacement								
EQ-5D	constant	-3,947	7,164	140	0	0		
	non-constant	-3,802	7,147	133	6	1	731	731
EQ-VAS	constant	-4,720	6,538	78	62	0		
	non-constant	-4,634	6,649	76	62	2	275	369
OKS	constant	-4,076	6,941	132	7	0		
	non-constant	-4,002	6,872	126	12	1	785	785
Hip replacement								
EQ-5D	constant	-4,278	7,482	130	8	0		
	non-constant	-4,209	7,176	109	23	7	-1,346	393
EQ-VAS	constant	-4,543	7,770	112	26	0		
	non-constant	-4,638	7,860	114	22	2	-234	512
OHS	constant	-4,262	7,131	137	0	1	-279	-279
	non-constant	-4,080	6,562	74	30	34	-3,828	404
Groin hernia repair								
EQ-5D	constant	-874	1,705	146	0	0		
	non-constant	-860	1,720	142	3	3	-250	-93
EQ-VAS	constant	-877	1,727	143	0	0		
	non-constant	-877	1,729	145	0	0		
Varicose vein surgery								
EQ-5D	constant	-711	1,260	124	0	0		
	non-constant	-711	1,244	121	1	2	-35	197
EQ-VAS	constant	-802	1,243	116	0	8	-106	115
	non-constant	-789	1,249	108	3	13	-327	98
AVVQ	constant	-821	1,231	122	0	2	-74	91
	non-constant	-675	1,189	112	3	9	-78	454

PROM = patient-reported outcome measure; MCQ = marginal costs of quality; OHS/OKS = Oxford Hip and Knee scores; AVVQ = Aberdeen varicose vein questionnaire

[a] All effects are in GBP

[b] After adjusting for health outcomes

[c] Only for econ. and stat. significant changes. Negative numbers indicate improvements in estimated provider performance.

7.2 Appendix to Chapter 3

Table A3.1: Correlation between performance estimates on EQ-5D dimensions

	Mobility	Self-Care	Usual Activities	Pain / Discomfort	Anxiety / Depression
Mobility	1.000				
Self-Care	0.343	1.000			
Usual Activities	0.707	0.450	1.000		
Pain/Discomfort	0.532	0.346	0.561	1.000	
Anxiety/Depression	0.236	0.294	0.311	0.380	1.000

Notes: Based on J=230 hospital performance estimates. All correlations are statistically significantly different from zero at p<0.01.

7.3 Appendix to Chapter 4

Table A4.1: Descriptive statistics for included and excluded observations

	Included			Excluded		
Description	N	Mean	SD	N	Mean	SD
Achievement measures (Dependent variables)						
Length of stay (in days)	95,878	5.34	3.64	90,017	6.05	4.37
Waiting time > 18 weeks	92,154	0.17	0.38	84,937	0.20	0.40
28-day emergency readmission	95,955	0.05	0.22	90,158	0.06	0.24
Patient characteristics (Control variables)						
Patient age (in years)	95,955	67.43	11.29	90,158	68.66	11.64
Patient gender (1=male, 0=female)	95,955	0.41	0.49	89,944	0.39	0.49
Primary diagnosis						
Osteoarthritis	95,955	0.06	0.24	90,158	0.08	0.28
Rheumatoid arthritis	95,955	0.93	0.25	90,158	0.91	0.29
Other	95,955	0.01	0.07	90,158	0.01	0.08
Number of Elixhauser comorbidities						
0	95,955	0.35	0.48	90,158	0.34	0.47
1	95,955	0.29	0.45	90,158	0.28	0.45
2-3	95,955	0.26	0.44	90,158	0.26	0.44
4+	95,955	0.10	0.31	90,158	0.13	0.33
Previously admitted as an emergency (1=yes, 0=no)	95,955	0.08	0.28	90,158	0.11	0.31
Socio-economic status	95,955	0.12	0.09	90,158	0.13	0.10
Healthcare Resource Group						
HB12C - category 2 without CC	95,955	0.77	0.42	90,158	0.75	0.43
HB11C - category 1 without CC	95,955	0.10	0.29	90,158	0.08	0.28
HB12B - category 2 with CC	95,955	0.07	0.26	90,158	0.07	0.26
HB12A - category 2 with major CC	95,955	0.04	0.19	90,158	0.04	0.21
HB11B - category 1 with CC	95,955	0.01	0.11	90,158	0.01	0.10
other	95,955	0.02	0.12	90,158	0.04	0.19

Legend: N = Number of observations, SD = Standard deviation; CC = complications or co-morbidities.

Notes: Healthcare Resource Groups refer to major hip procedures for non-trauma patients in category 1 (HB12x) or category 2 (HB11x). Socio-economic status is approximated by the % of neighbourhood residents claiming income benefits. This characteristics is measured at neighbourhood level (lower super output area (LSOA)).

Table A4.2: Estimated coefficients and standard errors from multivariate regression model

	Length of stay		Post-operative OHS		Waiting time $>$ 18 weeks		28-day emergency readmission	
Variable	Est	SE	Est	SE	Est	SE	Est	SE
Constant	2.078	0.052***	27.154	0.823***	-1.335	0.053***	-1.609	0.119***
FY 2010/11	-0.096	0.011***	0.043	0.072	0.114	0.045*	-0.008	0.019
FY 2011/12	-0.203	0.015***	0.229	0.085**	0.208	0.050***	-0.051	0.020*
Pre-operative OHS	-0.011	0.001***	0.599	0.016***			-0.005	0.001***
Pre-operative OHS^2	0.000	0.000***	-0.009	0.000***				
Patient age	-0.027	0.002***	0.208	0.025***			-0.014	0.004***
Patient age^2	0.000	0.000***	-0.002	0.000***			0.000	0.000***
Male patient	-0.074	0.004***	0.908	0.062***			0.142	0.015***
Primary diagnosis: Rheumatoid arthritis	0.026	0.022	-0.486	0.529			-0.086	0.113
Primary diagnosis: Other	0.035	0.009***	-1.169	0.187***			0.079	0.028**
Elixhauser comorbidities: 1	0.025	0.004***	-0.456	0.068***			0.061	0.017***
Elixhauser comorbidities: 2-3	0.068	0.004***	-1.433	0.083***			0.148	0.017***
Elixhauser comorbidities: 4+	0.153	0.007***	-2.826	0.133***			0.285	0.023***
Previously admitted as an emergency	0.071	0.005***	-0.613	0.124***			0.137	0.023***
Socio-economic status	0.003	0.001**	-0.523	0.027***			0.011	0.005*
Disabled	-0.036	0.003***	2.586	0.080***			-0.065	0.016***
Living alone	0.111	0.005***	-0.368	0.071***				
Symptom duration: 1 - 5 years	0.020	0.004***	-0.654	0.077***				
Symptom duration: 6 - 10 years	0.039	0.005***	-1.335	0.121***				
Symptom duration: $>$ 10 years	0.055	0.007***	-1.712	0.159***				
Assistance in filling in PROM questionnaire	0.067	0.005***	-0.545	0.097***				
HRG: HB11C - category 1 without CC	0.037	0.006***						
HRG: HB12B - category 2 with CC	0.127	0.006***						
HRG: HB12A - category 2 with major CC	0.495	0.011***						
HRG: HB11B - category 1 with CC	0.122	0.016***						
HRG: other	0.376	0.031***						
First-stage residual	0.001	0.000	0.012	0.006*	0.001	0.001	0.003	0.001*
Var(θ_j)	0.025	0.002***	1.203	0.141***	0.378	0.038***	0.023	0.003***
Var(ϵ_{ij})	0.162	0.001***	68.563	0.340***	1.000		1.000	
Number of observations	95,955							

*** p< 0.001; ** p<0.01; * p<0.05

Legend: Est = Estimate; SE = Huber-White standard error (robust to unknown heteroscedasticity); OHS = Oxford Hip Score; HRG = Healthcare Resource Group; CC = Complications and comorbidities; FY = Financial year (April - March).

Notes: Socio-economic status is approximated by the % of neighbourhood residents claiming income benefits. This characteristic is measured at neighbourhood level (lower super output area (LSOA)).

Table A4.3: Estimated coefficients and standard errors - multinomial hospital choice model (first-stage)

Variable	Est	SE
Closest hospital	0.185	0.014***
Distance to hospital	-0.197	0.003***
Distance^2	0.001	0.0001***
Distance^3	-0.00002	0.000002***
Number of patients	95,955	
Number of providers	252	
Pseudo R^2	0.706	
$\chi^2(4)$	120,930	

*** p< 0.001; ** p<0.01; * p<0.05

Legend: Est = Estimate; SE = Huber-White standard error

Notes: Distance to hospital is measured as the straight-line distance from the centroid of the patient's lower super output area (LSOA) to the provider's headquarter (NHS trust) or hospital site (ISTCs). Distance is measured in kilometres.

Table A4.4: Correlation between performance dimensions - excluding ISTCs

Performance dimension	(1)	(2)	(3)	(4)
Length of stay (1)	1.00	***-0.13***	***0.02***	***0.02***
Post-operative OHS (2)	**-0.27**	1.00	***-0.02***	***-0.07***
Waiting time > 18 wks (3)	0.11	-0.02	1.00	*0.00*
28-day emergency readmission (4)	-0.03	**-0.46**	-0.02	1.00

Notes: Lower triangle reports the correlation between random effects at provider level, whereas upper triangle (in italics) reports the correlation between random effects (i.e. the idiosyncratic error term) at patient level. Bold indicates that the correlation is statistically significantly different from zero at the 95% level.

Table A4.5: Correlation between performance dimensions - accounting for provider average risk factors

Performance dimension	(1)	(2)	(3)	(4)
Length of stay (1)	1.00	***-0.13***	***0.02***	***0.02***
Post-operative OHS (2)	**-0.21**	1.00	***-0.02***	***-0.07***
Waiting time > 18 wks (3)	**0.19**	**-0.17**	1.00	*0.00*
28-day emergency readmission (4)	-0.08	**-0.35**	0.07	1.00

Notes: Lower triangle reports the correlation between random effects at provider level, whereas upper triangle (in italics) reports the correlation between random effects (i.e. the idiosyncratic error term) at patient level. Bold indicates that the correlation is statistically significantly different from zero at the 95% level.

Table A4.6: Correlation between performance dimensions - risk-adjustment based on HES data only

Performance dimension	(1)	(2)	(3)	(4)
Length of stay (1)	1.00	***-0.16***	***0.01***	***0.02***
Post-operative OHS (2)	**-0.41**	1.00	***-0.01***	***-0.07***
Waiting time > 18 wks (3)	**0.28**	**-0.37**	1.00	*0.00*
28-day emergency readmission (4)	0.04	**-0.47**	**0.17**	1.00

Notes: Lower triangle reports the correlation between random effects at provider level, whereas upper triangle (in italics) reports the correlation between random effects (i.e. the idiosyncratic error term) at patient level. Bold indicates that the correlation is statistically significantly different from zero at the 95% level.

Table A4.7: Comparison of fixed and random effects estimates - Length of stay

	Fixed effects		Random effects		Difference in estimates	
Variable	Est	SE	Est	SE	Absolute	Relative
Constant	2.072	0.029***	2.081	0.031***	-0.009	-0.43%
FY 2010/11	-0.096	0.003***	-0.096	0.003***	0.001	-0.63%
FY 2011/12	-0.203	0.004***	-0.204	0.004***	0.001	-0.31%
Pre-operative OHS	-0.011	0.001***	-0.011	0.001***	0.000	-0.29%
Pre-operative OHS^2	0.000	0.000***	0.000	0.000***	0.000	-0.18%
Patient age	-0.027	0.001***	-0.027	0.001***	0.000	-0.22%
Patient age^2	0.000	0.000***	0.000	0.000***	0.000	-0.10%
Male patient	-0.074	0.003***	-0.074	0.003***	0.000	0.18%
Primary diagnosis: Rheumatoid arthritis	0.034	0.006***	0.034	0.006***	0.000	0.57%
Primary diagnosis: Other	0.026	0.018	0.026	0.018	0.000	1.10%
Elixhauser comorbidities: 1	0.025	0.003***	0.025	0.003***	0.000	0.75%
Elixhauser comorbidities: 2-3	0.068	0.004***	0.068	0.004***	0.000	-0.01%
Elixhauser comorbidities: 4+	0.152	0.005***	0.152	0.005***	0.000	-0.12%
Previously admitted as an emergency	0.071	0.005***	0.071	0.005***	0.000	-0.32%
Socio-economic status	0.003	0.001**	0.003	0.001**	0.000	-4.41%
Disabled	-0.035	0.003***	-0.036	0.003***	0.000	-0.48%
Living alone	0.111	0.003***	0.111	0.003***	0.000	-0.14%
Symptom duration: 1 - 5 years	0.021	0.004***	0.021	0.004***	0.000	-0.68%
Symptom duration: 6 - 10 years	0.039	0.005***	0.040	0.005***	0.000	-0.77%
Symptom duration: > 10 years	0.056	0.006***	0.056	0.006***	0.000	-0.59%
Assistance in filling in PROM questionnaire	0.067	0.004***	0.067	0.004***	0.000	-0.03%
HRG: HB11C - category 1 without CC	0.037	0.005***	0.037	0.005***	0.000	-1.31%
HRG: HB12B - category 2 with CC	0.130	0.005***	0.130	0.005***	0.000	-0.23%
HRG: HB12A - category 2 with major CC	0.502	0.007***	0.503	0.007***	-0.001	-0.15%
HRG: HB11B - category 1 with CC	0.123	0.012***	0.124	0.012***	-0.001	-0.65%
HRG: other	0.388	0.011***	0.389	0.011***	-0.001	-0.20%
First-stage residual	0.001	0.000**	0.001	0.000**	0.000	0.30%
Hausman test statistic	75.17	***				
Sample size	95,878		95,878			

*** $p < 0.001$; ** $p<0.01$; * $p<0.05$

Legend: Est = Estimate; SE = Standard error; OHS = Oxford Hip Score; HRG = Healthcare Resource Group; CC = Complications and comorbidities; FY = Financial year (April - March).

Table A4.8: Comparison of fixed and random effects estimates - Post-operative OHS

	Fixed effects		Random effects		Difference in estimates	
Variable	Est	SE	Est	SE	Absolute	Relative
Constant	27.100	0.746***	26.708	0.747***	0.392	1.45%
FY 2010/11	0.031	0.077	0.071	0.077	-0.040	-132.25%
FY 2011/12	0.221	0.078**	0.272	0.077***	-0.050	-22.76%
Pre-operative OHS	0.593	0.014***	0.598	0.014***	-0.005	-0.92%
Pre-operative OHS^2	-0.009	0.000***	-0.009	0.000***	0.000	-1.01%
Patient age	0.209	0.022***	0.220	0.022***	-0.010	-5.00%
Patient age^2	-0.002	0.000***	-0.002	0.000***	0.000	-3.93%
Male patient	0.905	0.061***	0.897	0.061***	0.008	0.90%
Primary diagnosis: Rheumatoid arthritis	-1.257	0.142***	-1.066	0.136***	-0.191	15.19%
Primary diagnosis: Other	-0.547	0.437	-0.541	0.437	-0.006	1.07%
Elixhauser comorbidities: 1	-0.459	0.074***	-0.461	0.074***	0.002	-0.47%
Elixhauser comorbidities: 2-3	-1.429	0.079***	-1.448	0.079***	0.019	-1.33%
Elixhauser comorbidities: 4+	-2.801	0.115***	-2.845	0.115***	0.044	-1.58%
Previously admitted as an emergency	-0.594	0.117***	-0.622	0.117***	0.027	-4.57%
Socio-economic status	-0.499	0.022***	-0.540	0.022***	0.041	-8.19%
Disabled	2.568	0.065***	2.601	0.065***	-0.033	-1.30%
Living alone	-0.361	0.071***	-0.366	0.071***	0.005	-1.48%
Symptom duration: 1 - 5 years	-0.662	0.085***	-0.680	0.085***	0.018	-2.68%
Symptom duration: 6 - 10 years	-1.343	0.118***	-1.365	0.118***	0.022	-1.66%
Symptom duration: > 10 years	-1.727	0.138***	-1.752	0.138***	0.024	-1.41%
Assistance in filling in PROM questionnaire	-0.526	0.080***	-0.520	0.080***	-0.006	1.13%
First-stage residual	0.011	0.005*	0.014	0.005**	-0.003	-23.37%
Hausman test statistic	173.71	***				
Sample size	81,336		81,336			

*** p< 0.001; ** p<0.01; * p<0.05

Legend: Est = Estimate; SE = Standard error; OHS = Oxford Hip Score; FY = Financial year (April - March).

Table A4.9: Comparison of fixed and random effects estimates - Waiting time > 18 weeks

	Fixed effects		Random effects		Difference in estimates	
Variable	Est	SE	Est	SE	Absolute	Relative
Constant	0.028	0.003***	0.027	0.003***	0.001	3.07%
FY 2010/11	0.051	0.003***	0.050	0.003***	0.001	1.67%
FY 2011/12	0.000	0.000	0.000	0.000	0.000	6.96%
First-stage residual	0.144	0.002***	0.123	0.008***	0.021	14.74%
Hausman test statistic	39.14	***				
Sample size	92,154		92,154			

*** p< 0.001; ** p<0.01; * p<0.05

Based on linear probability model. Estimated coefficients indicate increase in probability associated with unit change in covariate.

Legend: Est = Estimate; SE = Standard error; FY = Financial year (April - March).

Table A4.10: Comparison of fixed and random effects estimates - 28-day emergency readmission

	Fixed effects		Random effects		Difference in estimates	
Variable	Est	SE	Est	SE	Absolute	Relative
Constant	0.072	0.015***	0.076	0.015***	-0.003	-4.78%
FY 2010/11	-0.001	0.002	-0.001	0.002	0.000	-80.54%
FY 2011/12	-0.005	0.002**	-0.006	0.002**	0.000	-8.84%
Pre-operative OHS	0.000	0.000***	-0.001	0.000***	0.000	-4.12%
Patient age	-0.002	0.000***	-0.002	0.000***	0.000	-4.64%
Patient age^2	0.000	0.000***	0.000	0.000***	0.000	-2.23%
Male patient	0.015	0.001***	0.015	0.001***	0.000	0.49%
Primary diagnosis: Rheumatoid arthritis	0.008	0.003*	0.007	0.003*	0.001	7.32%
Primary diagnosis: Other	-0.011	0.010	-0.011	0.010	0.000	0.70%
Elixhauser comorbidities: 1	0.005	0.002**	0.005	0.002**	0.000	-1.18%
Elixhauser comorbidities: 2-3	0.014	0.002***	0.014	0.002***	0.000	-2.14%
Elixhauser comorbidities: 4+	0.034	0.003***	0.035	0.003***	-0.001	-1.97%
Previously admitted as an emergency	0.019	0.003***	0.020	0.003***	-0.001	-3.64%
Socio-economic status	0.001	0.001*	0.002	0.001**	-0.001	-46.55%
Disabled	-0.006	0.002***	-0.006	0.002***	0.000	-2.88%
First-stage residual	0.000	0.000*	0.000	0.000	0.000	25.20%
Hausman test statistic	46.10	***				
Sample size	95,955		95,955			

*** p< 0.001; ** p<0.01; * p<0.05

Based on linear probability model. Estimated coefficients indicate increase in probability associated with unit change in covariate.

Legend: Est = Estimate; SE = Standard error; OHS = Oxford Hip Score; FY = Financial year (April - March).

7.4 Appendix to Chapter 5

Table A5.1: Mixed logit choice model

	Mean		Standard deviation	
Variable	Est	SE	Est	SE
Distance (in km)	-0.184	0.002***		
Distance^2	0.000	0.000***		
NHS trust - medium	-0.530	0.030***		
NHS trust - multi-service	-0.603	0.099***		
NHS trust - small	-0.791	0.038***		
NHS trust - specialist	1.023	0.072***		
NHS trust - teaching	-0.445	0.033***		
Independent sector treatment centre	-1.467	0.045***		
Primary care trust	-1.159	0.206***		
Waiting time (in months)	0.013	0.015		
Change in Oxford Hip Score	0.118	0.008***	0.000	0.001
28-day emergency readmission rate (in %)	-0.052	0.004***	0.000	0.001
28-day mortality rate (in %)	-0.031	0.026	-0.002	0.005
WTT(OHS change)	1.287	0.085***		
WTT(Readmission rate)	-0.981	0.079***		
WTT(Mortality rate)	-0.086	0.072		
Number of patients	173,032			
Number of providers	230			
BIC	461,041			

*** p<0.001; ** p<0.01; * p<0.05

Notes: Random coefficient (mixed) multinomial logit model of choice of hospital for elective hip replacement patients treated between April 2010 and March 2013. OHS change, waiting time, readmission rate and mortality rate are lagged by one year. Coefficients are marginal utilities. Random coefficients are specified for OHS change, readmission rate and mortality rate and estimates. WTT is the ratio of the coefficient on the quality variable to the marginal utility of distance evaluated at the median distance (in km). Interaction terms with patient characteristics not reported (available on request). Standard errors are clustered at GP practice level. The mean coefficients do differ from those reported in Table 5.2 for the conditional logit model if there is less rounding.

Table A5.2: Sensitivity analyses

	Model (1)		Model (2)		Model (3)		Model (4)	
Variable	Est	SE	Est	SE	Est	SE	Est	SE
Main effects								
Distance (in km)	-0.183	0.002***	-0.184	0.002***	-0.184	0.002***	-0.207	0.003***
Distance^2	0.000	0.000***	0.000	0.000***	0.000	0.000***	0.001	0.000***
NHS trust - medium	-0.508	0.032***	-0.530	0.030***	-0.540	0.031***	-0.643	0.036***
NHS trust - multi-service	-0.656	0.090***	-0.615	0.099***	-0.612	0.097***	-0.604	0.098***
NHS trust - small	-0.740	0.037***	-0.792	0.038***	-0.790	0.038***	-0.909	0.042***
NHS trust - specialist	1.019	0.071***	1.002	0.072***	1.065	0.073***	1.047	0.083***
NHS trust - teaching	-0.502	0.033***	-0.432	0.033***	-0.412	0.032***	-0.513	0.037***
Independent sector treatment centre	-1.472	0.068***	-1.522	0.037***	-1.398	0.044***		
Primary care trust	-1.071	0.200***	-1.153	0.206***	-1.111	0.206***	-1.195	0.213***
Waiting time (in months)	0.033	0.018	-0.193	0.076*	0.007	0.015	-0.029	0.018
Change in PROM score	0.104	0.007***	0.115	0.008***	4.935	0.284***	0.121	0.010***
28-day emergency readmission rate (in %)	-0.054	0.004***	-0.052	0.004***	-0.054	0.004***	-0.054	0.006***
28-day mortality rate (in %)	-0.044	0.024	-0.032	0.026	-0.026	0.026	0.033	0.035
Interaction with distance								
x Patient age	-0.001	0.000***	-0.002	0.000***	-0.002	0.000***	-0.002	0.000***
x Male	0.001	0.001	0.002	0.001	0.002	0.001	0.005	0.002**
x Past utilisation	-0.003	0.002	-0.003	0.002	-0.003	0.002	-0.001	0.002
x Comorbidity count	-0.005	0.001***	-0.004	0.001***	-0.003	0.001***	-0.001	0.001
x Income deprivation	-0.179	0.015***	-0.186	0.017***	-0.189	0.018***	-0.241	0.022***
Interaction with waiting time								
x Patient age	0.003	0.000***	0.008	0.002**	0.003	0.000***	0.003	0.001***
x Male	-0.018	0.010	-0.068	0.051	-0.010	0.009	0.001	0.012
x Past utilisation	-0.005	0.012	0.049	0.059	-0.007	0.012	0.008	0.013
x Comorbidity count	-0.036	0.007***	-0.057	0.033	-0.018	0.007**	-0.016	0.007*
x Income deprivation	0.032	0.086	1.180	0.458*	0.083	0.084	0.007	0.107
Interaction with change in PROM score								
x Patient age	0.001	0.000***	0.001	0.000*	0.047	0.010***	0.001	0.000
x Male	0.004	0.005	-0.007	0.005	-0.300	0.214	-0.002	0.006

continued

Table A5.2: Sensitivity analyses

	Model (1)		Model (2)		Model (3)		Model (4)	
Variable	Est	SE	Est	SE	Est	SE	Est	SE
x Past utilisation	-0.014	0.006*	-0.008	0.007	-0.327	0.240	-0.012	0.007
x Comorbidity count	-0.009	0.003**	-0.009	0.003**	-0.475	0.128***	-0.004	0.004
x Income deprivation	-0.389	0.046***	-0.448	0.047***	-18.634	1.698***	-0.568	0.056***
Interaction with 28-day emergency readmission rate								
x Patient age	0.000	0.000*	0.000	0.000*	0.000	0.000	-0.001	0.000**
x Male	0.002	0.003	0.000	0.003	0.000	0.003	-0.003	0.004
x Past utilisation	0.012	0.004**	0.012	0.003***	0.012	0.004***	0.010	0.004**
x Comorbidity count	0.003	0.002	0.001	0.002	0.001	0.002	0.003	0.002
x Income deprivation	0.082	0.026**	0.124	0.026***	0.138	0.026***	0.121	0.034***
Interaction with 28-day mortality rate								
x Patient age	-0.002	0.001*	-0.001	0.001	-0.001	0.001	-0.001	0.001
x Male	-0.051	0.021*	-0.053	0.022*	-0.049	0.022*	-0.058	0.031
x Past utilisation	0.039	0.030	0.049	0.026	0.047	0.026	0.060	0.032
x Comorbidity count	-0.001	0.015	-0.010	0.015	-0.009	0.015	-0.004	0.018
x Income deprivation	-0.059	0.172	-0.023	0.168	-0.049	0.167	-0.230	0.210
WTT(PROM change)	1.174	0.086***	1.261	0.086***	1.284	0.076***	1.078	0.089***
WTT(Readmission rate)	-1.042	0.081***	-0.983	0.080***	-1.034	0.079***	-0.824	0.089***
WTT(Mortality rate)	-0.125	0.067	-0.091	0.073	-0.074	0.074	0.055	0.059
Number of patients	176,471		173,032		171,737		148,629	
Number of providers	233		230		225		144	
BIC	473,568		460,956		450,787		299,221	
Pseudo R^2	0.645		0.637		0.639		0.701	

*** $p<0.001$; ** $p<0.01$; * $p<0.05$

Notes: Conditional logit model of choice of hospital for elective hip replacement patients treated between April 2010 and March 2013. PROM change, waiting time, readmission rate and mortality rate are lagged by one year if not otherwise stated. Coefficients are marginal utilities. WTT is the ratio of the coefficient on the quality variable to the marginal utility of distance evaluated at the median distance (in km). Interaction terms with distance^2 and provider type not reported (available on request). Standard errors are clustered at GP practice level.

Model (1) - PROM change, waiting time, readmission rate and mortality rate are contemporaneous. Based on observed choices for patients treated between April 2009 and March 2012. Since April 2012 PROM scores have been reported separately for primary and revision hip replacement surgeries, so that our measures of PROM quality are no longer comparable.

Model (2) - Proportion of patients waiting longer than 120 days substituted for waiting time (both lagged).

Model (3) - Lagged EQ-5D change scores substituted for lagged OHS change scores.

Model (4) - Patients' choice sets exclude independent sector treatment centres.

Table A5.3: Descriptive statistics - emergency sample

Variable	Obs	Mean	SD	ICC
Patient characteristics				
Distance travelled (in km)	73,629	14.2	27.1	
Distance travelled past closest provider (in km)	73,629	4.2	25.4	
Number of providers within 10km radius	73,629	1.0	1.4	
Number of providers within 30km radius	73,629	5.4	5.4	
Age	73,629	80.9	9.8	
Male	73,629	0.27	0.44	
Past utilisation	73,629	0.65	1.17	
Number of Elixhauser conditions	73,629	0.99	1.56	
Income deprivation	73,629	0.14	0.10	
Provider characteristics				
Observed volume	394	186.9	87.1	80.7%
Waiting time (in months)	394	3.0	0.7	46.1%
Change in Oxford Hip Score	394	19.4	1.3	49.1%
28-day emergency readmission rate (in %)	394	5.99	2.20	38.2%
28-day mortality rate (in %)	394	0.20	0.25	5.3%

Obs = Observations; SD = Standard deviation; ICC = Intraclass correlation coefficient.
Notes: Patient characteristics for patients choosing provider between April 2010 and March 2013. Provider waiting time, change in Oxford Hip Score, readmission rate, mortality rate are based on elective patients treated by the respective providers and are for financial years 2009/10 to 2011/12.

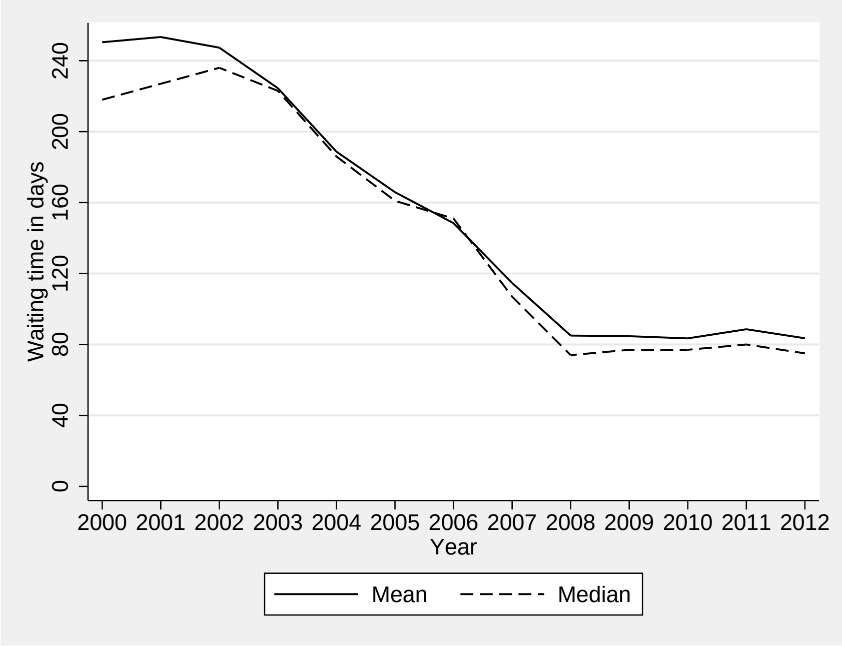


Figure A5.1: Waiting time for elective hip replacement surgery in England - 2000 to 2012

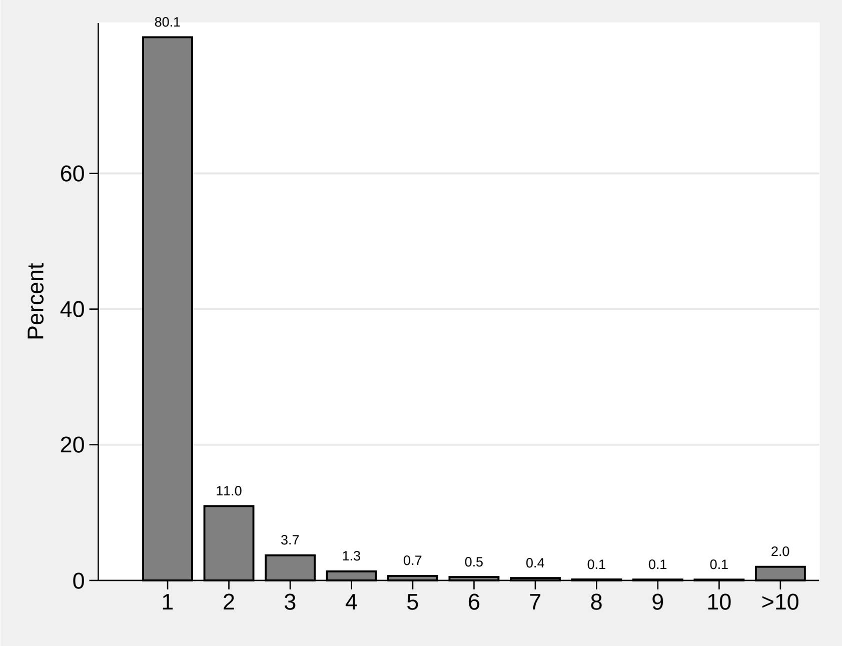


Figure A5.2: Percentage of emergency patients who went to their Nth nearest hospital

References

Aitkin, M. and N. Longford (1986). 'Statistical modelling issues in school effectiveness studies'. *Journal of the Royal Statistical Society, Series A*, 149: 1–43.

Appleby, J., E. Poteliakhoff J, K. Shah and N. Devlin (2013). 'Using patient-reported outcome measures to estimate cost-effectiveness of hip replacements in English hospitals'. *Journal of the Royal Society of Medicine*, 106: 323–331.

Appleby, J. and N. Devlin (2004). *Measuring success in the NHS: using patient assessed health outcomes to manage the performance of health care providers*. The King's Fund. London.

Appleby, J., C. Ham, C. Imison and M. Jennings (2010). *Improving NHS productivity*. The King's Fund, London.

Arrow, K. (1963). 'Uncertainty and the Welfare Economics of Medical Care'. *American Economic Review*, 53: 941–973.

– (1968). 'The Economics of Moral Hazard: Further Comment'. *American Economic Review*, 58: 537–539.

– (1986). 'Agency and the market'. In: *Handbook of mathematical economics*. Ed. by K. Arrow and M. Intriligator. Vol. 3. Amsterdam: Elsevier.

Ash, A., S. Fienberg, T. Louis, S.-L. T. Normand, T. Stukel and J. Utts (2012). *Statistical issues in assessing hospital performance*. Centre for Medicare & Medicaid Services.

Atkinson, T. (2005). *Atkinson Review: Final report - Measurement of Government Output and Productivity for the National Accounts*. Basingstoke: Palgrave MacMillan.

Austin, P., D. Alter and J. Tu (2003). 'The Use of Fixed- and Random-Effects Models for Classifying Hospitals as Mortality Outliers: A Monte Carlo Assessment'. *Medical Decision Making*, 23: 526–539.

Bailey, T. and P. Hewson (2004). 'Simultaneous modelling of multiple traffic safety performance indicators by using a multivariate generalized linear mixed model'. *Journal of the Royal Statistical Society. Series A*, 167: 501–517.

Basu, A. and A. Manca (2012). 'Regression Estimators for Generic Health-Related Quality of Life and Quality-Adjusted Life Years'. *Medical Decision Making*, 32: 56–69.

Beck, N. (2005). 'Multilevel Analysis of Comparative Data: A Comment'. *Political Analysis*, 13: 457–458.

Beckert, W., M. Christensen and K. Collyer (2012). 'Choice of NHS-funded hospital services in England'. *The Economic Journal*, 122: 400–417.

Belmont, P. J., G. P. Goodman, B. R. Waterman, J. O. Bader and A. J. Schoenfeld (2014). 'Thirty-Day Postoperative Complications and Mortality Following Total Knee Arthroplasty - Incidence and Risk Factors Among a National Sample of 15,321 Patients'. *The Journal of Bone & Joint Surgery*, 96: 20–26.

Bernal-Delgado, E., T. Christiansen, K. Bloor, C. Mateus, A. Yazbeck, J. Munck and J. Bremner (2015). 'ECHO: health care performance assessment in several European health systems'. *European Journal of Public Health*, 25: 3–7.

Berstock, J. R., A. D. Beswick, E. Lenguerrand, M. R. Whitehouse and A. W. Blom (2014). 'Mortality after total hip replacement surgery: A systematic review'. *Bone and Joint Research*, 3: 175–182.

Berwick, D. (2008). 'The science of improvement'. *Journal of the American Medical Association*, 299: 1182–1184.

Besley, T. and M. Ghatak (2003). 'Incentives, Choice and Accountability in the Provision of Public Services'. *Oxford Review of Economic Policy*, 19: 235–249.

Bjorgul, K., W. Novicoff and K. Saleh (2010). 'Evaluating comorbidities in total hip and knee arthroplasty: available instruments'. *Journal of Orthopaedics and Traumatology*, 11: 203–209.

Boadway, R. and N. Bruce (1984). *Welfare economics*. Oxford: Blackwell.

Bojke, C., A. Castelli and O. Nizalova (2011). ‘Exploring the concept of ‘avoidable mortality’ as a quality indicator for NHS hospital output: the case of circulatory diseases in England’. HESG York 2011.

Bradford, W. D., A. N. Kleit, M. A. Krousel-Wood and R. N. Re (2001). ‘Stochastic Frontier Estimation of Cost Models within the Hospital’. *The Review of Economics and Statistics*, 83: 302–309.

Bradley, E., L. Curry, S. Ramanadhan, L. Rowe, I. Nembhard and H. Krumholz (2009). ‘Research in action: using positive deviance to improve quality of health care’. *Implementation Science*, 4: 25.

Braeutigam, R. R. and M. V. Pauly (1986). ‘Cost Function Estimation and Quality Bias: The Regulated Automobile Insurance Industry’. *RAND Journal of Economics*, 17: 606–617.

Brazier, J., R. Akehurst, A. Brennan, P. Dolan, K. Claxton, C. McCabe, M. Sculpher and A. Tsuchiya (2005). ‘Should patients have a greater role in valuing health states?’ *Applied Health Economics and Health Policy*, 4: 201–208.

Brekke, K., H. Gravelle, L. Siciliani and O. Straume (2014). ‘Patient Choice, Mobility and Competition Among Health Care Providers’. In: *Health Care Provision and Patient Mobility*. Ed. by R. Levaggi and M. Montefiori. Springer.

Breslow, N. and D. Clayton (1993). ‘Approximate Inference in Generalized Linear Mixed Models’. *Journal of the American Statistical Association*, 88: 9–25.

Briggs, A. and P. Fenn (1998). ‘Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane’. *Health Economics*, 7: 723–740.

Brooks, R. (1996). ‘EuroQol: the current state of play’. *Health Policy*, 37: 53–72.

Brown, P., L. Panattoni, L. Cameron, S. Knox, T. Ashton, T. Tenbensel and J. Windsor (2015). ‘Hospital sector choice and support for public hospital care in New Zealand: Results from a labeled discrete choice survey’. *Journal of Health Economics*, 43: 118–127. DOI: 10.1016/j.jhealeco.2015.06.004.

Browne, J. P., H. Bastaki and J. Dawson (2013). ‘What is the optimal time point to assess patient-reported recovery after hip and knee replacement? a systematic review and analysis of routinely reported outcome data from the English patient-reported outcome measures programme'. *Health and Quality of Life Outcomes*, 11: 128.

Browne, J., L. Jamieson, J. Lewsey, J. van der Meulen, L. Copley and N. Black (2008). 'Case-mix & patients' reports of outcome in Independent Sector Treatment Centres: Comparison with NHS providers'. *BMC Health Services Research*, 8: 78.

Bryk, A. and S. Raudenbush (1988). 'Toward a More Appropriate Conceptualization of Reseach on School Effects: A Three-Level Hierarchical Linear Model'. *American Journal of Education*, 97: 65–108.

Busse, R., J. Schreyögg and P. C. Smith (2008). 'Variability in healthcare treatment costs amongst nine EU countries - results from the HealthBASKET project'. *Health Economics*, 17: S1–S8.

Carey, K. (2015). 'Measuring the Hospital Length of Stay/Readmission Cost Trade-Off Under a Bundled Payment Mechanism'. *Health Economics*, 24: 790–802.

Carey, K. and J. F. Burgess (1999). 'On measuring the hospital cost/quality trade-off'. *Health Economics*, 8: 509–520.

– (2000). 'Hospital Costing: Experience from the VHA'. *Financial Accountability & Management*, 16: 289–308.

Carey, K. and T. Stefos (2011). 'Measuring the cost of hospital adverse patient safety events'. *Health Economics*, 20: 1417–1430.

Chamberlain, G. (1982). 'Multivariate regression models for panel data'. *Journal of Econometrics*, 18: 5–46.

Chard, J., M. Kuczawski, N. Black and J. van der Meulen (2011). 'Outcomes of elective surgery undertaken in independent sector treatment centres and NHS providers in England: audit of patient outcomes in surgery'. *British Medical Journal*, 343: d6404.

Charlson, M. E., P. Pompei, K. L. Ales and C. R. MacKenzie (1987). 'A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation'. *Journal of Chronic Diseases*, 40: 373–383.

Chauhan, D. and J. Sussex (2008). *Report of the OHE Commission on NHS Outcomes, Performance and Productivity*. Office for Health Economics. London.

Chua, C., A. Palangkaraya and J. Yong (2010). 'A two-stage estimation of hospital quality using mortality outcome measures: an application using hospital administrative data'. *Health Economics*, 19: 1404–1424.

Clement, J., V. Valdmanis, G. Bazzoli, M. Zhao and A. Chukmaitov (2008). 'Is more better? An analysis of hospital outcomes and efficiency with a DEA model of output congestion'. *Health Care Management Science*, 11: 67–77.

Clement, N. D., A. Muzammil, D. MacDonald, C. R. Howie and L. C. Biant (2011). 'Socioeconomic status affects the early outcome of total hip replacement'. *Journal of Bone & Joint Surgery, British Volume*, 93-B: 464–469.

Coles, J. (2010). *PROMs risk adjustment methodology - guide for general surgery and orthopaedic procedures*. Northgate Informations Solutions Ltd & CHKS Ltd.

Commission on the Future of Health and Social Care in England (2014). *The UK private health market*. Appendix to 'A new settlement for Health and Social Care'. London: The King's Fund.

Contoyannis, P., A. M. Jones and N. Rice (2004). 'The dynamics of health in the British Household Panel Survey'. *Journal of Applied Econometrics*, 19: 473–503.

Coronini-Cronberg, S., J. Appleby and J. Thompson (2013). 'Application of patient-reported outcome measures (PROMs) data to estimate cost-effectiveness of hernia surgery in England'. *Journal of the Royal Society of Medicine*, 106: 278–287.

Cutler, D. M., R. S. Ilckman and M. B. Landrum (2004). 'The Role of Information in Medical Markets: An Analysis of Publicly Reported Outcomes in Cardiac Surgery'. *American Economic Review*, 94: 342–346.

Dafny, L. S. (2005). 'How Do Hospitals Respond to Price Changes?' *American Economic Review*, 95: 1525–1547.

Daidone, S. and F. D'Amico (2009). 'Technical efficiency, specialization and ownership form: Evidences from a pooling of Italian hospitals'. *Journal of Productivity Analysis*, 32: 203–216.

Daidone, S. and A. Street (2011). *Estimating the costs of specialised care*. CHE Research Paper 61. Centre for Health Economics, University of York.

Daniels, M. and S.-L. T. Normand (2006). 'Longitudinal profiling of health care units based on continuous and discrete patient outcomes'. *Biostatistics*, 7: 1–15.

Dawson, J., R. Fitzpatrick, A. Carr and D. Murray (1996). 'Questionnaire on the perceptions of patients about total hip replacement'. *Journal of Bone & Joint Surgery, British Volume*, 78-B: 185–190.

Dawson, J., R. Fitzpatrick, D. Murray and A. Carr (1998). 'Questionnaire on the perceptions of patients about total knee replacement'. *Journal of Bone & Joint Surgery, British Volume*, 80-B: 63–69.

De Wit, G. A., J. J. V. Busschbach and F. T. De Charro (2000). 'Sensitivity and perspective in the valuation of health status: whose values count?' *Health Economics*, 9: 109–126.

Deb, P. and J. F. Burgess (2003). *A Quasi-experimental Comparison of Econometric Models for Health Care Expenditures*. Department of Economics Working Papers 212. Hunter College.

Deily, M. E. and N. L. McKay (2006). 'Cost inefficiency and mortality rates in Florida hospitals'. *Health Economics*, 15: 419–431.

Department of Health (2002). *Reforming NHS Financial Flows - Introducing payment by results*. The Stationary Office, London.

– (2008a). *Guidance on the routine collection of Patient Reported Outcome Measures (PROMs)*. The Stationary Office, London.

– (2008b). *Report of the Advisory Committee on Resource Allocation*. The Stationary Office, London.

– (2010). *Payment by Results Guidance for 2009/10*. The Stationary Office, London.

– (2011a). *A simple guide to Payment by Results*. The Stationary Office, London.

– (2011b). *NHS reference costs 2009-2010 appendix: DBRC organisation-specific reference cost data*. The Stationary Office, London.

– (2012a). *Patient Reported Outcome Measures (PROMs) in England: The case-mix adjustment methodology*. The Stationary Office, London.

– (2012b). *Payment by Results Guidance for 2012-13*. The Stationary Office, London.

– (2015). *Handbook to the NHS constitution*. The Stationary Office, London.

Devlin, N. J. and J. Sussex (2011). *Incorporating multiple criteria in HTA: methods and processes*. Office for Health Economics, London.

Devlin, N., D. Parkin and J. Browne (2010). 'Patient-reported outcome measures in the NHS: new methods for analysing and reporting EQ-5D data'. *Health Economics*, 19: 886–905.

Dijs-Elsinga, J., W. Otten, M. M. Versluijs, H. J. Smeets, J. Kievit, R. Vree, W. J. van der Made and P. J. Marang-van de Mheen (2010). 'Choosing a Hospital for Surgery: The Importance of Information on Quality of Care'. *Medical Decision Making*, 30: 544–555.

Ding, V., R. Hubbard, C. Rutter and G. Simon (2013). 'Assessing the accuracy of profiling methods for identifying top providers: performance of mental health care providers'. *Health Services and Outcomes Research Methodology*, 13: 1–17.

Dixit, A. (2002). 'Incentives and Organizations in the Public Sector: An Interpretative Review'. *The Journal of Human Resources*, 37: 696–727.

Dolan, P. (1997). 'Modeling valuations for EuroQol health states'. *Medical Care*, 35: 1095–108.

Donabedian, A. (1966). 'Evaluating the Quality of Medical Care'. *The Milbank Memorial Fund Quarterly*, 44: 166–206.

– (1988). 'The Quality of Care: How Can It Be Assessed?' *Journal of the American Medical Association*, 260: 1743–1748.

Dormont, B. and C. Milcent (2004). 'The sources of hospital cost variability'. *Health Economics*, 13: 927–39.

Dowd, B., T. Swenson, R. Kane, S. Parashuram and R. Coulam (2014). 'Can data envelopment analysis provide a scalar index of 'value'?' *Health Economics*, 23: 1465–1480.

Dranove, D., D. Kessler, M. McClellan and M. Satterthwaite (2003). 'Is more information better? The effects of "report cards" on health care providers'. *Journal of Political Economy*, 111: 555–588.

Dranove, D. and A. Sfekas (2008). 'Start spreading the news: A structural estimate of the effects of New York hospital report cards'. *Journal of Health Economics*, 27: 1201–1207.

Drummond, M. F., M. J. Sculpher, G. W. Torrance, B. J. O'Brien and G. L. Stoddart (2005). *Methods for the Economic Evaluation of Health Care Programmes*. 3rd ed. Oxford: Oxford University Press.

Efron, B. and C. Morris (1973). 'Stein's Estimation Rule and Its Competitors – An Empirical Bayes Approach'. *Journal of the American Statistical Association*, 68: 117–130.

Elixhauser, A., C. Steiner, D. Harris and R. Coffey (1998). 'Comorbidity measures for use with administrative data'. *Medical Care*, 36: 8–27.

Emmert, M., F. Eijkenaar, H. Kemter, A. S. Esslinger and O. Schöffski (2012). 'Economic evaluation of pay-for-performance in health care: a systematic review'. *European Journal of Health Economics*, 13: 755–767.

Evans, R. (1974). *Strained mercy: the economics of Canadian medical care*. Toronto: Butterworths.

Faber, M., M. Bosch, H. Wollersheim, S. Leatherman and R. Grol (2009). 'Public reporting in health care: how do consumers use quality-of-care information?: A systematic review'. *Medical Care*, 47: 1–8.

Farrar, S., D. Yi, M. Sutton, M. Chalkley, J. Sussex and A. Scott (2009). 'Has payment by results affected the way that English hospitals provide care? Difference-in-differences analysis'. *British Medical Journal*, 339: b3047.

Fischer, C., H. F. Lingsma, P. J. Marang-van de Mheen, D. S. Kringos, N. S. Klazinga and E. W. Steyerberg (2014). 'Is the Readmission Rate a Valid Quality Indicator? A Review of the Evidence'. *PLoS One*, 9: e112282.

Fitzpatrick, R. (2009). 'Patient-reported outcome measures and performance measurement'. In: *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Ed. by P. C. Smith, E. Mossialos, I. Papanicolas and S. Leatherman. Cambridge: Cambridge University Press. Chap. 2.2, 63–86.

Fleming, S. T. (1991). ‘The relationship between quality and cost’. *Inquiry*, 28: 29–38.

Fung, C. H., Y.-W. Lim, S. Mattke, C. Damberg and P. G. Shekelle (2008). ‘Systematic Review: The Evidence That Publishing Patient Care Performance Data Improves Quality of Care’. *Annals of Internal Medicine*, 148: 111–123.

Garellick, G., J. Kärrholm, C. Rogmark and P. Herberts (2009). *Swedish Hip Arthroplasty Register - Annual Report 2009 (shortened version)*. Sahlgrenska University Hospital.

Garratt, A. M., L. M. Macdonald, D. A. Ruta, I. T. Russell, J. K. Buckingham and Z. H. Krukowski (1993). ‘Towards measurement of outcome for patients with varicose veins’. *Quality in Health Care*, 2: 5–10.

Garratt, A., L. Schmidt, A. Mackintosh and R. Fitzpatrick (2002). ‘Quality of life measurement: bibliographic study of patient assessed health outcome measures’. *British Medical Journal*, 324: 1417–1422.

Gaynor, M., K. Ho and R. J. Town (2015). ‘The Industrial Organization of Health-Care Markets’. *Journal of Economic Literature*, 53: 235–284.

Gaynor, M., H. Seider and W. Vogt (2005). ‘The Volume-Outcome Effect, Scale Economies, and Learning by Doing’. *American Economic Review*, 95: 243–247.

Gaynor, M., C. Propper and S. Seiler (2012). *Free to Choose? Reform and Demand Response in the English National Health Service*. Working Paper 18574. National Bureau of Economic Research.

Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gertler, P. J. and D. M. Waldman (1992). ‘Quality-adjusted Cost Functions and Policy Evaluation in the Nursing Home Industry’. *Journal of Polical Economy*, 100: 1232–1256.

Geweke, J., G. Gowrisankaran and R. J. Town (2003). ‘Bayesian Inference for Hospital Quality in a Selection Model’. *Econometrica*, 71: 1215–1238.

Gibbons, R. D. and D. Hedeker (1997). ‘Random Effects Probit and Logistic Regression Models for Three-Level Data’. *Biometrics*, 53: 1527–1537.

Glazer, J., T. McGuire and S.-L. T. Normand (2008). 'Mitigating the Problem of Unmeasured Outcomes in Quality Reports'. *The B.E. Journal of Economic Analysis & Policy*, 8: Article 7.

Goddard, M. and R. Jacobs (2009). 'Using composite indicators to measure performance in health care'. In: *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Ed. by P. Smith, E. Mossialos, I. Papanicolas and S. Leatherman. Cambridge: Cambridge University Press. Chap. 3.4, 339–368.

Goddard, M., R. Mannion and P. C. Smith (2000). 'Enhancing performance in health care: a theoretical perspective on agency and the role of information'. *Health Economics*, 9: 95–107.

Goldstein, H. (1986). 'Multilevel mixed linear model analysis using iterative generalized least squares'. *Biometrika*, 73: 43–56.

– (1997). 'Methods in School Effectiveness Research'. *School Effectiveness and School Improvement*, 8: 369–395.

Goldstein, H., W. Browne and J. Rasbash (2002). 'Partitioning variation in multilevel models'. *Understanding Statistics*, 1: 223–231.

Goldstein, H. and M. J. R. Healy (1995). 'The Graphical Presentation of a Collection of Means'. *Journal of the Royal Statistical Society. Series A*, 158: 175–177.

Goldstein, H. and D. J. Spiegelhalter (1996). 'League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance'. English. *Journal of the Royal Statistical Society: Series A*, 159: 385–443.

Gomes, M., N. Gutacker, C. Bojke and A. Street (2015). 'Addressing missing data in patient-reported outcome measures (PROMs): implications for comparing provider performance.' *Health Economics*, forthcoming. DOI: 10.1002/hec.3173.

Gowrisankaran, G. and R. J. Town (1999). 'Estimating the quality of care in hospitals using instrumental variables'. *Journal of Health Economics*, 18: 747–767.

Gravelle, H., R. Santos and L. Siciliani (2014). 'Does a hospital's quality depend on the quality of other hospitals? A spatial econometrics approach'. *Regional Science and Urban Economics*, 49: 203–216.

Gray, A. M., O. Rivero-Arias and P. M. Clarke (2006). 'Estimating the Association between SF-12 Responses and EQ-5D Utility Values by Response Mapping'. *Medical Decision Making*, 26: 18–29.

Greene, W. H. and D. A. Hensher (2010). *Modeling Ordered Choices*. Cambridge: Cambridge University Press.

Greiner, W., T. Weijnen, M. Nieuwenhuizen, S. Oppe, X. Badia, J. Busschbach, M. Buxton, P. Dolan, P. Kind, P. Krabbe, A. Ohinmaa, D. Parkin, M. Roset, H. Sintonen, A. Tsuchiya and F. de Charro (2003). 'A single European currency for EQ-5D health states'. *The European Journal of Health Economics*, 4: 222–231.

Grossman, M. (1972). 'On the Concept of Health Capital and the Demand for Health'. *Journal of Political Economy*, 80: 223–255.

Gutacker, N., K. Bloor and R. Cookson (2015). 'Comparing the performance of the Charlson/Deyo and Elixhauser comorbidity measures across five European countries and three conditions'. *European Journal of Public Health*, 25: 15–20.

Gutacker, N., C. Bojke, S. Daidone, N. Devlin and A. Street (2013). 'Hospital Variation in Patient-Reported Outcomes at the Level of EQ-5D Dimensions: Evidence from England'. *Medical Decision Making*, 33: 804–818.

Gutacker, N., A. Street, M. Gomes and C. Bojke (2015). 'Should English healthcare providers be penalised for failing to collect patient-reported outcome measures (PROMs)?' *Journal of the Royal Society of Medicine*, forthcoming. DOI: 10.1177/0141076815576700.

Häkkinen, U., G. Rosenqvist, M. Peltola, S. Kapiainen, H. Rättö, F. Cots, A. Geissler, Z. Or, L. Serdén and R. Sund (2014). 'Quality, cost, and their trade-off in treating AMI and stroke patients in European hospitals'. *Health Policy*, 117: 15–27.

Hall, B. and B. Hamilton (2004). 'New information technology systems and a Bayesian hierarchical bivariate probit model for profiling surgeon quality at a large hospital'. *The Quarterly Review of Economics and Finance*, 44: 410–429.

Harris, J. (1977). 'The Internal Organization of Hospitals: Some Economic Implications'. *The Bell Journal of Economics*, 8: 467–482.

Hauck, K., N. Rice and P. C. Smith (2003). 'The influence of health care organisations on health system performance'. *Journal of Health Services Research & Policy*, 8: 68–74.

Hauck, K. and A. Street (2006). 'Performance assessment in the context of multiple objectives: A multivariate multilevel analysis'. *Journal of Health Economics*, 25: 1029–1048.

Hauck, K. and X. Zhao (2011). 'How Dangerous is a Day in Hospital? A Model of Adverse Events and Length of Stay for Medical Inpatients'. *Medical Care*, 49: 1068–1075.

Hausman, J. (1978). 'Specification Tests in Econometrics'. *Econometrica*, 46: 1251–1271.

Health & Social Care Information Centre (2013a). *Emergency readmissions to hospital within 28 days of discharge: primary hip replacement surgery*. Compendium of Population Health Indicators. Accessed on 24/11/2014. URL: `https://indicators.ic.nhs.uk/download/NCHOD/Specification/Spec_33D_533ISP4CPP1_12_V1.pdf`.

– (2013b). *Finalised Patient Reported Outcome Measures (PROMs) in England: April 2011 to March 2012*. Accessed on 26/03/2015. URL: `http://www.hscic.gov.uk/catalogue/PUB11359/final-proms-eng-apr11-mar12-fin-report-v2.pdf`.

– (2015). *Hospital Episode Statistics, Admitted Patient Care, England - 2013-14: Procedures and interventions*. Accessed on 10/07/2015. URL: `http://www.hscic.gov.uk/catalogue/PUB16719/hosp-epis-stat-admi-proc-2013-14-tab.xlsx`.

Hedeker, D. and R. D. Gibbons (2006). *Longitudinal Data Analysis*. Hoboken, NJ: Wiley.

Hensher, D. A. and W. H. Greene (2003). 'The Mixed Logit model: The state of practice'. *Transportation*, 30: 133–176.

Hernández Alava, M., A. J. Wailoo and R. Ara (2012). 'Tails from the Peak District: Adjusted Limited Dependent Variable Mixture Models of EQ-5D Questionnaire Health State Utility Values'. *Value in Health*, 15: 550–561.

Hibbard, J., J. Stockard and M. Tusler (2005). 'Hospital Performance Reports: Impact On Quality, Market Share, And Reputation'. *Health Affairs*, 24: 1150–1160.

Hibbard, J. H., J. Stockard and M. Tusler (2003). 'Does Publicizing Hospital Performance Stimulate Quality Improvement Efforts?' *Health Affairs*, 22: 84–94.

Hildon, Z., J. Neuburger, D. Allwood, J. van der Meulen and N. Black (2012). 'Clinicians' and patients' views of metrics of change derived from patient reported outcome measures (PROMs) for comparing providers' performance of surgery'. *BMC Health Services Research*, 12: 171.

Hodgkin, D. (1996). 'Specialized service offerings and patients' choice of hospital: the case of cardiac catheterization'. *Journal of Health Economics*, 15: 305–332.

Hole, A. R. (2007a). 'A comparison of approaches to estimating confidence intervals for willingness to pay measures'. *Health Economics*, 16: 827–840.

– (2007b). 'Fitting mixed logit models by using maximum simulated likelihood'. *The Stata Journal*, 7: 388–401.

Holmström, B. and P. Milgrom (1991). 'Multi-task principle-agent problems: Incentive contracts, asset owernship and job design'. *Journal of Law, Economics and Organization*, 7: 24–52.

Holmström, B. (1979). 'Moral hazard and observability'. *Bell Journal of Economics*, 10: 74–91.

– (1982). 'Moral hazard in teams'. *Bell Journal of Economics*, 13: 324–340.

Hox, J. (2002). *Multilevel Analysis: Techniques and Application*. Mahwah, NJ: Lawrence Erlbaum.

Hsiao, C. (1986). *Analysis of panel data*. Cambridge: Cambridge University Press.

Hussey, O., S. Wertheimer and A. Mehrotra (2013). 'The Association Between Health Care Quality and Cost: A Systematic Review'. *Annals of Internal Medicine*, 158: 27–34.

Husted, H., G. Holm and S. Jacobsen (2008). 'Predictors of length of stay and patient satisfaction after hip and knee replacement surgery - Fast-track experience in 712 patients'. *Acta Orthopaedica*, 79: 168–173.

Hutchings, A., J. Neuburger, J. van der Meulen and N. Black (2014). 'Estimating recruitment rates for routine use of patient reported outcome measures and the impact on provider comparisons.' *BMC Health Services Research*, 14: 66.

Hvenegaard, A., J. N. Arendt, A. Street and D. Gyrd-Hansen (2011). 'Exploring the relationship between costs and quality: Does the joint evaluation of costs and quality alter the ranking of Danish hospital departments?' *European Journal of Health Economics*, 12: 541–551.

Iversen, T. and L. Siciliani (2011). 'Non-price rationing and waiting times'. In: *The Oxford Handbook of Health Economics*. Ed. by S. Glied and P. C. Smith. Oxford: Oxford University Press.

Jackson, T. (2001). 'Using computerised patient-level costing data for setting DRG weights: the Victorian (Australia) cost weight studies'. *Health Policy*, 56: 149–163.

Jacobs, R., P. C. Smith and A. Street (2006). *Measuring Efficiency in Health Care*. Cambridge: Cambridge University Press.

Jensen, P. H., E. Webster and J. Witt (2009). 'Hospital type and patient outcomes: an empirical examination using AMI readmission and mortality records'. *Health Economics*, 18: 1440–1460.

Jilke, S. (2015). 'Choice and equality: are vulnerable citizens worse off after liberalization reforms?' *Public Administration*, 93: 68–85.

Jones, H. and D. Spiegelhalter (2009). 'Accounting for regression-to-the-mean in tests for recent changes in institutional performance: Analysis and power.' *Statistics in Medicine*, 28: 1645–1667.

Judge, A., J. Chard, I. Learmonth and P. Dieppe (2006). 'The effects of surgical volumes and training centre status on outcomes following total joint replacement: analysis of the Hospital Episode Statistics for England'. *Journal of Public Health*, 28: 116–124.

Jung, K., R. Feldman and D. Scanlon (2011). 'Where would you go for your next hospitalization?' *Journal of Health Economics*, 30: 832–841.

Kalbfleisch, J. and R. Wolfe (2013). 'On Monitoring Outcomes of Medical Providers'. *Statistics in Biosciences*, 5: 286–302.

Karnon, J., O. Caffrey, C. Pham, R. Grieve, D. Ben-Tovim, P. Hakendorf and M. Crotty (2013). ‘Applying risk adjusted cost-effectiveness (RAC-E) analysis to hospitals: Estimating the costs and consequences of variation in clinical practice’. *Health Economics*, 22: 631–642.

Keeler, E. B. (1990). ‘What proportion of hospital cost differences is justifiable?’ *Journal of Health Economics*, 9: 359–65.

Keller, K. and R. Staelin (1987). ‘Effects of Quantity and Quality of Information on Decision Effectiveness’. *Journal of Consumer Research*, 14: 200–213.

Ketelaar, N., M. Faber, S. Flottorp, L. H. Rygh, K. H. O. Deane and M. P. Eccles (2011). *Public release of performance data in changing the behaviour of healthcare consumers, professionals or organisations*. Cochrane Database of Systematic Reviews. 11.

Kind, P., R. Brooks and R. Rabin (2005). *EQ-5D concepts and methods: a developmental history*. Dordrecht: Springer.

Kind, P. and A. Williams (2004). ‘Measuring success in health care - the time has come to do it properly!’ *Health Policy matters*, 9: 1–8.

Kruse, M. and J. Christensen (2013). ‘Is quality costly? Patient and hospital cost drivers in vascular surgery’. *Health Economics Review*, 3: 22.

Laffont, J.-J. and J. Tirole (1993). *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: MIT Press.

Lakhani, A., J. Coles, D. Eayres, C. Spence and B. Rachet (2005). ‘Creative use of existing clinical and health outcomes data to assess NHS performance in England: Part 1 - performance indicators closely linked to clinical care’. *British Medical Journal*, 330: 1426–1431.

Landrum, M., S. Bronskill and S.-L. T. Normand (2000). ‘Analytic Methods for Constructing Cross-Sectional Profiles of Health Care Providers’. *Health Services and Outcomes Research Methodology*, 1: 23–47.

Landrum, M., S.-L. T. Normand and R. Rosenheck (2003). ‘Selection of Related Multivariate Means’. *Journal of the American Statistical Association*, 98: 7–16.

Larsen, K., O. Sørensen, T. Hansen, P. Thomsen and K. Søballe (2008). 'Accelerated perioperative care and rehabilitation intervention for hip and knee replacement is effective: A randomized clinical trial involving 87 patients with 3 months of follow-up'. *Acta Orthopaedica*, 79: 149–159.

Laudicella, M., K. R. Olsen and A. Street (2010). 'Examining cost variation across hospital departments - A two-stage multi-level approach using patient-level data'. *Social Science & Medicine*, 71: 1872–1881.

Lawton, R., N. Taylor, R. Clay-Williams and J. Braithwaite (2014). 'Positive deviance: a different approach to achieving patient safety'. *BMJ Quality & Safety*, 23: 880–883.

Leckie, G. and C. Charlton (2013). 'runmlwin: Stata module for fitting multilevel models in the MLwiN software package'. *Journal of Statistical Software*, 52: 1–40.

Leckie, G. and H. Goldstein (2009). 'The limitations of using school league tables to inform school choice'. *Journal of the Royal Statistical Society: Series A*, 172: 835–851.

Lewis, J. B. and D. A. Linzer (2005). 'Estimating Regression Models in Which the Dependent Variable is Based on Estimates'. *Political Analysis*, 13: 345–364.

Lilford, R. and P. Pronovost (2010). 'Using hospital mortality rates to judge hospital performance: A bad idea that just won't go away'. *British Medical Journal*, 340: 955–957.

Lindeboom, M. and E. van Doorslaer (2004). 'Cut-point shift and index shift in self-reported health'. *Journal of Health Economics*, 23: 1083–1099.

Little, R. and D. Rubin (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Luft, H. S., S. S. Hunt and S. C. Maerki (1987). 'The Volume-Outcome Relationship: Practice-Makes-Perfect or Selective-Referral Patterns?' *Health Services Research*, 22: 157–182.

Ma, C.-t. A. (1994). 'Health Care Payment Systems: Cost and Quality incentives'. *Journal of Economics & Management Strategy*, 3: 93–112.

Mäkelä, K. T., U. Häkkinen, M. Peltola, M. Linna, H. Kröger and V. Remes (2011). 'The effect of hospital volume on length of stay, re-admissions, and complications of total hip arthroplasty'. *Acta Orthopaedica*, 82: 20–26.

Mann, R., J. Brazier and A. Tsuchiya (2009). 'A comparison of patient and general population weightings of EQ-5D dimensions'. *Health Economics*, 18: 363–372.

Marshall, M. N., P. G. Shekelle, H. T. O. Davies and P. C. Smith (2003). 'Public Reporting On Quality In The United States And The United Kingdom'. *Health Affairs*, 22: 134–148.

Marshall, M. N., P. S. Romano and H. T. O. Davies (2004). 'How do we maximize the impact of the public reporting of quality of care?' *International Journal for Quality in Health Care*, 16: 57–63.

Martin, S., N. Rice, R. Jacobs and P. Smith (2007). 'The market for elective surgery: Joint estimation of supply and demand'. *Journal of Health Economics*, 26: 263–285.

Martin, S. and P. C. Smith (2005). 'Multiple Public Service Performance Indicators: Toward an Integrated Statistical Approach'. *Journal of Public Administration Research and Theory*, 15: 599–613.

Maynard, A. (2012). 'The powers and pitfalls of payment for performance'. *Health Economics*, 21: 3–12.

Maynard, A. and K. Bloor (2010). 'Patient reported outcome measurement: learning to walk before we run'. *Journal of the Royal Society of Medicine*, 103: 129–132.

McFadden, D. (1974). 'Conditional logit analysis of qualitative choice behaviour'. In: *Frontiers in economics*. Ed. by P. Zarembka. Vol. 4. New York: Academic Press, 105–142.

McGrail, K., S. Bryan and J. Davis (2012). 'Let's All Go to the PROM: The Case for Routine Patient-Reported Outcome Measurement in Canadian Healthcare'. *HealthcarePapers*, 11: 8–18.

McGuire, A. (2001). 'Theoretical concepts in the economic evaluation of health care'. In: *Economic evaluation in health care - merging theory and practice*. Ed. by M. Drummond and A. McGuire. Oxford University Press.

McGuire, A., J. Henderson and G. Mooney (1988). *The economics of health care: an introductory text*. London: Routledge and Kegan Paul.

McKay, N. L. and M. E. Deily (2008). 'Cost inefficiency and hospital health outcomes'. *Health Economics*, 17: 833–848.

McKelvey, R. D. and W. Zavoina (1975). 'A statistical model for the analysis of ordinal level dependent variables'. *The Journal of Mathematical Sociology*, 4: 103–120.

Meacock, R., S. R. Kristensen and M. Sutton (2014). 'The cost-effectiveness of using financial incentivies to improve provider quality: a framework and application'. *Health Economics*, 23: 1–13.

Moja, L., A. Piatti, V. Pecoraro, C. Ricci, G. Virgili, G. Salanti, L. Germagnoli, A. Liberati and G. Banfi (2012). 'Timing Matters in Hip Fracture Surgery: Patients Operated within 48 Hours Have Better Outcomes. A Meta-Analysis and Meta-Regression of over 190,000 Patients'. *PLoS ONE*, 7: e46175.

Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. Springer Series in Statistics. New York: Springer.

Monitor and NHS England (2013). *National Tariff Payment System - Annex 4A: Additional information on currencies with national prices*.

Monstad, K., L. Engester and B. Espehaug (2006). *Patient preferences for choice of hospital*. Working Paper No 05/06. Health Economics Bergen.

Montez-Rath, M., C. L. Christiansen, S. L. Ettner, S. Loveland and A. K. Rosen (2006). 'Performance of statistical models to predict mental health and substance abuse cost'. *BMC Medical Research Methodology*, 6: 53.

Mooney, G. and M. Lange (1991). *Economic appraisal in pre-natal screening: reassessing benefits*. Discussion Paper 08/91. Health Economics Research Unit, University of Aberdeen.

Morey, R. M., D. J. Fine, S. W. Loree, D. L. Retzlaff-Roberts and S. Tsubakitani (1992). 'The Trade-off Between Hospital Cost and Quality of Care'. *Medical Care*, 30: 677–698.

Moscone, F., E. Tosetti and G. Vittadini (2012). 'Social interaction in patients' hospital choice: evidence from Italy'. *Journal of the Royal Statistical Society: Series A*, 175: 453–472.

Mukamel, D. B., J. Zwanziger and K. J. Tomaszewski (2001). 'HMO Penetration, Competition, and Risk-Adjusted Hospital Mortality'. *Health Services Research*, 36: 1019–1035.

Mundlak, Y. (1978). 'On the Pooling of Time Series and Cross Section Data'. *Econometrica*, 46: 69–85.

Murray, C. J., E. Özaltin, A. Tandon, J. A. Salomon, R. Sadana and S. Chatterji (2003). 'Empirical Evaluation of the Anchoring Vignette Approach in Health Surveys'. In: *Health Systems Performance Assessment - Debates, Methods and Empiricism*. Ed. by C. J. Murray and D. B. Evans. Geneva: World Health Organization. Chap. 30, 369–399.

National Clinical Audit Advisory Group (2011). *Detection and management of outliers*. The Stationary Office, London.

National Joint Registry (2011). *NJR PROMs questionnaires*. Accessed 12/1/2012. URL: `http://www.njrcentre.org.uk/njrcentre/tabid/199/Default.aspx`.

National Patient Safety Agency (2011). *Organisation Patient Safety Incident Reports*. URL: `http://www.nrls.npsa.nhs.uk/EasySiteWeb/getresource.axd?AssetID=62923`.

Naylor, C. and S. Gregory (2009). *Independent sector treatment centres*. Tech. rep. London: The King's Fund.

Neuburger, J., A. Hutchings, N. Black and J. van der Meulen (2013). 'Socioeconomic differences in patient-reported outcomes after a hip or knee replacement in the English National Health Service'. *Journal of Public Health*, 35: 115–124.

Neuburger, J., A. Hutchings, J. van der Meulen and N. Black (2013). 'Using patient-reported outcomes (PROs) to compare the provider of surgery: does the choice of measure matter?' *Medical Care*, 51: 517–523.

Newhouse, J. P. (1994). 'Frontier estimation: How useful a tool for health economics?' *Journal of Health Economics*, 13: 317–322.

NHS Information Centre (2010a). *A guide to PROMs methodology*. Provisional Monthly Patient Reported Outcome Measures (PROMs) in England. NHS Information Centre.

– (2010b). *PROMs score summary by provider - April 2009 to March 2010*. [Accessed: 01/02/2011]. URL: `http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937%7B%5C&%7DcategoryID=1488`.

NICE (2011). *Hip fracture: The manamanage of hip fracture in adults*. NICE Guidelines 124. London.

Noble, M., G. Wright, G. Smith and C. Dibben (2006). 'Measuring multiple deprivation at the small-area level'. *Environment and Planning A*, 38: 169–185.

Normand, S.-L. T., M. Glickman and C. Gatsonis (1997). 'Statistical Methods for Profiling Providers of Medical Care: Issues and Applications'. *Journal of the American Statistical Association*, 92: 803–814.

Nuttall, D., D. Parkin and N. Devlin (2015). 'Inter-provider comparison of patient-reported outcomes: developing an adjustment to account for differences in patient case mix'. *Health Economics*, 24: 41–54.

O'Hagan, A., J. Stevens and J. Montmartin (2000). 'Inference for the Cost-Effectiveness Acceptability Curve and Cost-Effectiveness Ratio'. *PharmacoEconomics*, 17: 339–349.

Olsen, K. R. and A. Street (2008). 'The analysis of efficiency among a small number of organisations: How inferences can be improved by exploiting patient-level data'. *Health Economics*, 17: 671–681.

Parkin, D., N. Rice and N. Devlin (2010). 'Statistical Analysis of EQ-5D Profiles: Does the Use of Value Sets Bias Inference?' *Medical Decision Making*, 30: 556–565.

Paton, F., D. Chambers, P. Wilson, A. Eastwood, D. Craig, D. Fox, D. Jayne and E. McGinnes (2014). 'Effectiveness and implementation of enhanced recovery after surgery programmes: a rapid evidence synthesis'. *BMJ Open*, 4: e005015.

Pauly, M. V. (1980). *Doctors and Their Workshops: Economic Models of Physician Behavior*. Chicago: University of Chicago Press.

Pedraja-Chaparro, F., J. Salinas-Jimenez and P. Smith (1999). 'On the Quality of the Data Envelopment Analysis Model'. *The Journal of the Operational Research Society*, 50: 636–644.

Petersen, L., L. Woodard, T. Urech, C. Daw and S. Sookanan (2006). 'Does pay-for-performance improve the quality of health care?' *Annals of Internal Medicine*, 145: 265–272.

Picone, G. A., F. A. Sloan, S.-Y. Chou and D. H. Taylor (2003). 'Does Higher Hospital Cost Imply Higher Quality of Care?' *The Review of Economics and Statistics*, 85: 51–62.

Pope, D. (2009). 'Reacting to rankings: Evidence from "America's best hospitals"'. *Journal of Health Economics*, 28: 1154–1165.

Porter, M. E. (2010). 'What Is Value in Health Care?' *The New England Journal of Medicine*, 363: 2477–2481.

Portrait, F., O. Galiën and B. van den Berg (2015). 'Measuring healthcare providers' performance within managed competition using multidimensional quality and cost indicators'. *Health Economics*, forthcoming. DOI: 10.1002/hec.3158.

Prendergast, C. (1999). 'The Provision of Incentives in Firms'. *Journal of Economic Literature*, 37: 7–63.

Propper, C., M. Sutton, C. Whitnall and F. Windmeijer (2008). 'Did 'Targets and Terror' Reduce Waiting Times in England for Hospital Care?' *The B.E. Journal of Economic Analysis & Policy*, 8: Article 5.

Propper, C. and D. Wilson (2003). 'The Use and Usefullness of Performance Measures in the Public Sector'. *Oxford Review of Economic Policy*, 19: 250–267.

– (2012). 'The use of performance measures in health care systems'. In: *The Elgar Companion to Health Economics*. Ed. by A. M. Jones. 2nd ed. Edward Elgar. Chap. 33, 350–358.

Propper, C., M. Damiani, G. Leckie and J. Dixon (2007). 'Impact of patients' socioeconomic status on the distance travelled for hospital admission in the English National Health Service'. *Journal of Health Services Research & Policy*, 12: 153–159.

Qian, X., L. B. Russell, E. Vailyeva and J. E. Miller (2011). "Quicker and sicker' under Medicare's prospective payment systemt for hospitals: New evidence on an old issue from a national longitudinal survey'. *Bulletin of Economic Research*, 63: 1–27.

Rabe-Hesketh, S., A. Skrondal and A. Pickles (2002). 'Reliable estimation of generalized linear mixed models using adaptive quadrature'. *Stata Journal*, 2: 1–21.

Racz, M. J. and J. Sedransk (2010). 'Bayesian and Frequentist Methods for Provider Profiling Using Risk-Adjusted Assessments of Medical Outcomes'. *Journal of the American Statistical Association*, 105: 48–58.

Raudenbush, S. and J. Willms (1995). 'The Estimation of School Effects'. *Journal of Educational and Behavioral Statistics*, 20: 307–335.

Rice, N. and A. Jones (1997). 'Multilevel models and health economics'. *Health Economics*, 6: 561–575.

Rice, N., S. Robone and P. Smith (2012). 'Vignettes and health systems responsiveness in cross-country comparative analysis'. *Journal of the Royal Statistical Society: Series A*, 175: 337–369.

Robinson, W. (1950). 'Ecological correlations and the behavior of individuals'. *American Sociological Review*, 15: 351–357.

Rogerson, W. (1994). 'Choice of treatment intensities by a nonprofit hospital under prospective pricing'. *Journal of Economics & Management Strategy*, 3: 7–51.

Roland, M. (2004). 'Linking physician pay to quality of care - a major experiment in the United Kingdom'. *New England Journal of Medicine*, 351: 1448–1454.

Romley, J. A. and D. P. Goldman (2011). 'How Costly is Hospital Quality? A Revealed-Preference Approach'. *The Journal of Industrial Economics*, 59: 578–608.

Rosenthal, M. B. (2007). 'Nonpayment for performance? Medicare's new reimbursement rule'. *New England Journal of Medicine*, 357: 1573.

Ruwaard, S. and R. Douven (2014). *Quality and hospital choice for cataract treatments: the winner takes most*. CPB Netherlands Bureau for Economic Policy Analysis.

Ryan, M., D. Scott, C. Reeves, A. Bate, E. van Teijlingen, E. Russell, M. Napper and C. Robb (2001). 'Eliciting public preferences for healthcare: a systematic review of techniques'. *Health Technology Assessment*, 5: 1–186.

Ryan, M. (1992). *The agency relationship in health care: identifying areas for future research*. Discussion Paper 02/92. Health Economics Research Unit, University of Aberdeen.

Samuelson, W. and R. Zeckhauser (1988). 'Status Quo Bias in Decision-Making'. *Journal of Risk and Uncertainty*, 1: 7–59.

Santos, R., H. Gravelle and C. Propper (2015). 'Does quality affect patients' choice of doctor? Evidence from England'. *The Economic Journal*, forthcoming. DOI: 10.1111/ecoj.12282.

Sappington, D. E. M. (1991). 'Incentives in Principal-Agent Relationships'. *Journal of Economic Perspectives*, 5: 45–66.

Schreyögg, J. and T. Stargardt (2010). 'The Trade-off between Costs and Outcomes: The Case of Acute Myocardial Infarction'. *Health Services Research*, 45: 1585–1601.

Searle, S., G. Casella and C. McCulloch (1992). *Variance components*. New York: Wiley.

Shleifer, A. (1985). 'A Theory of Yardstick Competition'. *RAND Journal of Economics*, 16: 319–327.

Siciliani, L., V. Moran and M. Borowitz (2014). 'Measuring and comparing health care waiting times in OECD countries'. *Health Policy*, 118: 292–303.

Siciliani, L., P. Sivey and A. Street (2013). 'Differences in length of stay for hip replacement between public hospital, specialised treatment centres and private providers: selection or efficiency?' *Health Economics*, 22: 234–242.

Siegel, J. E., G. Torrance, L. Russell, B. Luce, M. Weinstein and M. Gold (1997). 'Guidelines for Pharmacoeconomic Studies: Recommendations from the Panel on Cost-Effectiveness in Health and Medicine'. *PharmacoEconomics*, 11: 159–168.

Siggeirsdottir, K., O. Olafsson, H. Jonsson Jr., S. Iwarsson, V. Gudnason and B. Y. Jonsson (2005). 'Short hospital stay augmented with education and home-based rehabilitation improves function and quality of life after hip replacement - Randomized study of 50 patients with 6 months of follow-up'. *Acta Orthopaedica*, 76: 555–562.

Singh, J. A. (2011). 'Epidemiology of Knee and Hip Arthroplasty: A Systematic Review'. *The Open Orthopaedics Journal*, 5: 80–85.

Sivey, P. (2008). 'The effect of hospital quality on choice of hospital for elective heart operations in England'. PhD thesis. University of York.

– (2012). 'The effect of waiting time and distance on hospital choice for English cataract patients'. *Health Economics*, 21: 444–456.

Skinner, J. (1994). 'What do stochastic frontier cost functions tell us about inefficiency?' *Journal of Health Economics*, 13: 323–328.

Skrondal, A. and S. Rabe-Hesketh (2009). 'Prediction in multilevel generalized linear models'. *Journal of the Royal Statistical Society: Series A*, 172: 659–687.

Smith, P. C. and A. Street (2005). 'Measuring the efficiency of public services: the limits of analysis'. *Journal of the Royal Statistical Society. Series A*, 168: 401–417.

Smith, P. C. and A. Street (2013). 'On the uses of routine patient-reported health outcome data'. *Health Economics*, 22: 119–131.

Smith, P. C. (2002). 'Developing composite indicators for assessing health system efficiency'. In: *Measuring up - Improving health system performance in OECD countries*. Ed. by OECD. OECD Publications Service. Chap. 14, 295–316.

– (2015). 'Performance management: the clinician's tale'. *Health Economics, Policy and Law*, 10: 357–360.

Snijders, T. A. B. and R. J. Bosker (1999). *Multilevel analysis - An introduction to basic and advanced multilevel modeling*. London: Sage.

Spiegelhalter, D. J. (2005). 'Funnel plots for comparing institutional performance'. *Statistics in Medicine*, 24: 1185–1202.

Steyerberg, E. and H. Lingsma (2010). 'Complexities in quality of care information'. *Medical Decision Making*, 30: 529–530.

Street, A., N. Gutacker, C. Bojke, N. Devlin and S. Daidone (2014). 'Variation in outcome and costs among NHS providers for common surgical procedures:

econometric analysis of routinely collected data'. *Health Services and Delivery Research*, 2: 1–89.

Street, A., D. Scheller-Kreinsen, A. Geissler and R. Busse (2010). *Determinants of hospital costs and performance variation: Methods, models and variables for the EuroDRG project*. Working Papers in Health Policy and Management 03/2010. TU Berlin.

Sundararajan, V., T. Henderson, C. Perry, A. Muggivan, H. Quan and W. A. Ghali (2004). 'New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality'. *Journal of Clinical Epidemiology*, 57: 1288–1294.

Sutton, M., S. Nikolova, R. Boaden, H. Lester, R. McDonald and M. Roland (2012). 'Reduced Mortality with Hospital Pay for Performance in England'. *New England Journal of Medicine,* 367: 1821–1828.

Teixeira-Pinto, A. and S.-L. T. Normand (2008). 'Statistical methodology for classifying units on the basis of multiple-related measures'. *Statistics in Medicine,* 27: 1329–1350.

Terza, J., A. Basu and P. Rathouz (2008). 'Two-stage residual inclusion estimation: Addressing endogeniety in health econometric modeling'. *Journal of Health Economics*, 27: 531–543.

Thomas, J. (1996). 'Does risk-adjusted readmission rate provide valid information on hospital quality?' *Inquiry*, 33: 258–270.

Thomas, N., N. T. Longford and J. E. Rolph (1994). 'Empirical Bayes methods for estimating hospital-specific mortality rates'. *Statistics in Medicine,* 13: 889–903.

Thum, Y. (1997). 'Hierarchical Linear Models for Multivariate Outcomes'. *Journal of Educational and Behavioral Statistics,* 22: 77–108.

Timbie, J. W., J. P. Newhouse, M. B. Rosenthal and S.-L. T. Normand (2008). 'A Cost-Effectiveness Framework for Profiling the Value of Hospital Care'. *Medical Decision Making,* 28: 419–434.

Timbie, J. W. and S.-L. T. Normand (2008). 'A comparison of method for combining quality and efficiency performance measures: Profiling the value of hospital care following acute myocardial infarction'. *Statistics in Medicine*, 27: 1351–1370.

Timbie, J. W., D. Shahian, J. Newhouse, M. B. Rosenthal and S.-L. T. Normand (2009). 'Composite measures for hospital quality using quality-adjusted life years'. *Statistics in Medicine*, 28: 1238–1254.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. 1st ed. Cambridge: Cambridge University Press.

Van Hout, B., M. Al, G. Gordon and F. Rutten (1994). 'Costs, effects, and C/E-ratios alongside a clinical trial'. *Health Economics*, 3: 309–319.

Varagunam, M., A. Hutchings, J. Neuburger and N. Black (2014). 'Impact on hospital performance of introducing routine patient reported outcome measures in surgery'. *Journal of Health Services Research & Policy*, 19: 77–84.

Varkevisser, M., S. A. Geest and F. T. Schut (2010). 'Assessing hospital competition when prices don't matter to patients: the use of time-elasticities'. *International Journal of Health Care Finance and Economics*, 10: 43–60.

– (2012). 'Do patients choose hospitals with high quality ratings? Empirical evidence from the market for angioplasty in the Netherlands'. *Journal of Health Economics*, 31: 371–378.

Vrangbaek, K., R. Robertson, U. Winblad, H. Van de Bovenkamp and A. Dixon (2012). 'Choice policies in Northern European health systems'. *Health Economics, Policy and Law*, 7: 47–71.

Walker, S., M. Sculpher and M. Drummond (2011). 'The Methods of Cost-effectiveness Analysis to Inform Decisions about the Use of Health Care Interventions and Programs'. In: *The Oxford Handbook of Health Economics*. Ed. by S. Glied and P. C. Smith. Oxford University Press. Chap. 31, 733–758.

Ware, J. and C. Sherbourne (1992). 'The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection.' *Medical Care*, 30: 473–483.

Weech-Maldonado, R., D. Shea and V. Mor (2006). 'The Relationship Between Quality of Care and Costs in Nursing Homes'. *American Journal of Medical Quality*, 21: 40–48.

Wennberg, J. and A. Gittelsohn (1973). ‘Small Area Variation in Health Care Delivery: A population-based health information system can guide planning and regulatory decision-making’. *Science*, 182: 1102–1108.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Zellner, A. (1962). ‘An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias’. *Journal of the American Statistical Association*, 57: 348–368.