Analysing hospital variation in health outcome at the level of EQ-5D dimensions[*][†]

Nils Gutacker[‡][1], Chris Bojke[1], Silvio Daidone[1], Nancy Devlin[2], and Andrew Street[1]

[1]Centre for Health Economics, University of York, UK
[2]Office of Health Economics, London, UK

Abstract

Background: The English Department of Health has introduced routine collection of patient-reported health outcome data for selected surgical procedures (hip and knee replacement, hernia repair, varicose vein surgery) to facilitate patient choice and increase provider accountability. The EQ-5D has been chosen as the preferred generic instrument and the current risk-adjustment methodology is based on the EQ-5D index score to measure variation across hospital providers. There are two problems with this. First, using a population value set to generate the index score may not be appropriate for purposes of provider performance assessment because it introduces an exogenous source of variation and assumes identical preferences for health dimensions among patients. Second, the multimodal distribution of the index score creates statistical problems that are not yet resolved. Analysing variation for each dimension of the EQ-5D dimensions (mobility, self care, usual activities, pain/discomfort, anxiety/depression) seems therefore more appropriate and promising.

Aims: For hip replacement surgery, we aim to explore a) the impact of treatment on each EQ-5D dimension b) the extent to which treatment impact varies across providers c) the extent to which treatment impact across EQ-5D dimensions is correlated within providers.

Methods: We employ multilevel ordered probit models that recognise the hierarchical nature of the data (measurement points nested in patients, which themselves are nested in hospital providers) and the response distributions. The treatment impact is modelled as a random coefficient that varies at hospital-level. We obtain provider-specific Empirical Bayes (EB) estimates of this coefficient. We estimate separate models for each of the five EQ-5D dimensions and analyse correlations of the EB estimates across dimensions.

Data: Since April 2009, all providers of NHS-funded care are required to collect patient-reported outcome measures (PROMs) for the aforementioned elective procedures using the EQ-5D descriptive system. We combine information on pre- and post-operative EQ-5D outcomes with Hospital Episode Statistics for the financial year 2009/10. For reasons of practicality, we limit our research to one condition (hip replacement surgery). The overall sample consists of 25k patients with complete pre- and post-operative responses.

Findings: This research is still at an early stage. Our preliminary analysis suggests that hospital treatment is indeed associated with improvements in health and that variability in treatment impact is generally more pronounced on the dimensions 'mobility', 'usual activity' and 'pain & discomfort' than on others. The bivariate correlation between the provider EB estimates is substantial, suggesting a) that certain providers are better in improving health across multiple EQ-5D dimensions than others and b) multivariate models are appropriate and should be further investigated.

1   Introduction

Recent years have seen a growing trend to measure and publish hospital data on achieved health outcomes. Risk-adjusted mortality, re-admission or adverse events rates are now widely used by institutions such as the *Centers for Medicare & Medicaid Services* (USA) and *Dr. Foster* (UK) to generate league tables of hospital quality performance and highlight variation in outcomes across providers (Marshall et al., 2003). However, these measures reveal inherently little about the health of the vast majority of patients and there is a risk that variation in quality remains undetected. In order to allow for fair assessment of hospital performance and identify best practice it is necessary to move away from a focus on measures of hospital failure towards more comprehensive measures of patients' health outcome (Kind and Williams, 2004; Appleby and Devlin, 2004).

In this paper, we set out to measure variation in hospital quality using a new, routinely collected dataset on patient-reported outcome measures (PROMs). Since April 2009, all providers of publicly-funded inpatient care in the English National Health Service (NHS) have been required to collect such measures for four elective procedures: unilateral hip and knee replacements, varicose vein surgery, and groin hernia repairs (Department of Health, 2008). Eligible patients are invited to report their health status before and 3 or 6 months after surgery using a generic measure of health related quality of life, the EQ-5D, and condition-specific instruments. The EQ-5D consists of five questions that address impairments in overall health through self-assessed limitations on mobility, self care and usual activities as well as experienced pain and discomfort, and anxiety and depression (Brooks, 1996). The answers to each question are given on a three-point item scale (no/some/extreme problems) and can be aggregated to a (quasi-) continuous measures of patient health using utility weights obtained from members of the general public (Dolan, 1997). The condition-specific instruments follow the same logic but generally comprise more questions and response items, use somewhat arbitrary sets of weights and do not allow for comparisons across conditions.

So far, 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[1] 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. The NHS Information Centre has developed a preliminary risk-adjustment methodology that is currently being applied to the PROMs data (Coles, 2010). We build on these efforts and propose two refinements:

First, we argue that the data should be analysed at the level of PROM item responses instead of aggregated health measures. Collapsing EQ-5D health profile data into a single value by means of weighting comes at the cost of information loss, introduces exogenous variation that can bias statistical inference and raises normative concerns (Parkin et al., 2010). Furthermore, the idiosyncratic distribution of EQ-5D utility scores poses unresolved statistical problems for multivariate regression (Hernández Alava et al., 2010; Basu and Manca, in press).

---

[1]To our knowledge, the English NHS is the first health care system to make collection of patient-reported outcome 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 endorsed by the regulator for routine performance measurement.

Second, patients' health outcomes are likely to be influenced by both observed factors (e.g. age and gender) and unobserved factors (e.g. reporting heterogeneity, unobserved medical conditions) that are outside of the hospitals' control. There is a risk that such factors disguise variation in quality and lead to incorrect conclusions about provider performance. We argue that the risk-adjustment methodology should recognise and exploit the hierarchy of the data to distinguish random noise and patient heterogeneity from systematic variation in quality among providers.

In this study, we explore hospital performance with respect to self-reported health outcomes for hip replacement patients at the level of the individual, disaggregated PROM responses. We focus our attention on the EQ-5D and develop multilevel risk-adjustment model for each of the five functional dimensions. Our approach combines elements from the literature on determinants of self-reported health (e.g. Contoyannis et al., 2004) and the literature on cost-effectiveness in multi-centre trials (e.g. Manca et al., 2007) 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 providers. The Empirical Bayes (EB) estimates of this coefficient are then interpreted as capturing relative hospital quality. The presented methodology accounts for observed and unobserved patient and provider heterogeneity and the ordinal nature of the EQ-5D responses. It is readily applicable to other conditions for which PROMs data are collected and can also, in principle, be extended to the condition-specific instruments.

2 Conceptual framework

2.1 An agency model of health care provision

Health care describes the activity of improving patients' health or changing its trajectory by means of medical, surgical or preventive intervention. The underlying process can be considered as a production function, where the patient's initial health H_0 is transformed into a health status H_1 through the application of treatment. The effectiveness of the treatment depends on several factors, including the nature of the condition, N, the characteristics of the patient, X, and the quality Q with which the treatment is carried out, so that

$$H_1 = f(H_0, N, X, Q) \qquad (1)$$

Patients rarely possess the necessary medical knowledge and resources to perform the treatment themselves. Instead, they delegate the task to a specialised agent, here a team of hospital staff. The agent combines capital, labour and medical technology to provide the treatment and determines its quality through the specific amounts and combination of inputs chosen. However, the team's decisions may be constrained by exogenous factors in the production environment, thus limiting the level of quality achievable. Examples include the requirement to provide teaching or offer a diverse range of services, the team's skill-mix, or the availability of equipment, and the availability and proximity to capital resources, such as the operating theatre and ward. Denoting the agent's effort to provide the optimal treatment as e and production constraints as Z, quality can be expressed as

$$Q = q(e, Z) \qquad (2)$$

and substituting (2) into (1), we obtain

$$H_1 = g(H_0, N, X, e, Z) \tag{3}$$

Patients will always request the highest level of effort possible, thereby maximising the expected level of post-treatment health[2]. However, agents do not necessarily share this objective and may choose to exert less effort than feasible due to cost considerations, the disutility of effort, an unawareness of best practice[3], or co-ordination problems (Holmstrom and Milgrom, 1991). Because effort is inherently difficult to observe and post-treatment health is determined by factors outside of the teams' control, patients can neither contract or choose a provider on the basis of expected outcome of care nor verify the agents' performance ex-post. This gives rise to the well-known principal-agent problem (Lafonte and Tirole, 1993).

2.2   Measuring the providers' contribution to patient health

The economic and epidemiological literature emphasises comparative analysis (i.e. benchmarking) in a multiple regression framework as a means of identifying agents' relative performance (Shleifer, 1985; Iezzoni, 2003). By comparing providers against each other based on their realised achievement one can gain insights into the underlying production process and attribute variation in achievement to variation in observed, exogenous input factors, e.g. case-mix or size of operation. Any remaining, unobserved provider heterogeneity can then be analysed to make inferences about the relative level of effort exerted.

Assessing hospital performance with respect to the effort put into advancing quality requires information on patients' health before and after treatment. Clinical measures, such as blood pressure or joint movement, describe the patients' health in physiological terms but do not capture other relevant aspects, particularly quality of life (Black and Jenkinson, 2009). Only patients' themselves can give a full account of their perceived health and are therefore increasingly recognised by regulatory bodies as the preferred source of information on the effectiveness of care (NICE, 2008; FDA, 2009). To reduce the level of complexity and minimise cognitive burden, PROM instruments focus the evaluation on a restricted number of ordered health dimensions. A patient's overall health status H can then be characterised as a function of these responses, H^k, so that

$$H = h(H^1, \ldots, H^K) \tag{4}$$

where $k = 1, \ldots, K$ is the health dimension considered and h is an aggregation function to be defined.

Means of aggregating sub-dimensions of health into an overall score are available for a wide range of different PROM instruments. However, whether one should use such aggregation functions to summarise patients' health depends on the specific research question. Indeed, we argue that for the purpose of performance measurement, informing

---

[2]The highest level of effort possible is determined by the agent's participation constraint, i.e. the costs of care cannot exceed the reimbursement.

[3]Health care providers have a responsibility to stay informed about recent developments in medical technology and best practice. However, search and transaction costs in the form of time and effort may prevent them from obtaining this information. Thus, we regard unawareness as an (un-)conscious choice made by the provider.

patient choice and identifying best practice, the costs of aggregation usually 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 hospital providers affect each of them. Certain providers may perform well on one dimension but fall short on another dimension. 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, no aggregation function is neutral and observed variation in health outcome may be driven by the choice of function, not genuine heterogeneity in performance (Parkin et al., 2010). In some circumstances, one may be willing to accept the value judgement implied or explicitly expressed by the aggregation function. For example, the convention of using the preferences of the general public to aggregate EQ-5D profiles has a clearly articulated rationale in the context of cost effectiveness analysis and decisions concerning the allocation of taxpayer funding (Siegel et al., 1997)[4]. However, it should be understood that *measuring* and *valuing* health are two genuinely different activities. The use of aggregate PROMs data to inform patient choice raises normative concerns because it imposes a common valuation of health dimensions. In fact, reporting relative provider performance with respect to, say changes in EQ-5D utility, would only be justified if all (prospective) patients were to share the same relative values; an assumption unlikely to hold true in reality[5]. If patients are heterogeneous with respect to their relative valuation of health dimension, analysing variation on the level of health dimensions may be more appropriate as it allows patients to apply their own values and interpret performance data accordingly. Indeed, the PROMs programme has as its central concern measuring patients' views about their health outcomes, not societal valuations. Recognising this, we focus on how to use PROMs to measure changes in health status across patients and providers.

In addition to these normative considerations, there are statistical concerns arising from the construction of the aggregation function. For a substantive number of PROM instruments, aggregation functions are not *defined* by the developers of the instrument but *elicited* from a group of relevant stakeholders such as patients, doctors or the general public. This elicitation exercise involves making statistical inferences from a sample of participants to the population and from a set of health states to the entire spectrum covered by the PROM instrument. As a consequence, the elicited aggregation function itself is subject to uncertainty which may bias the results of performance assessments conducted in a multivariate regression framework (Parkin et al., 2010).

Third, most PROM instruments impose ordinal scales onto the health dimension under consideration. The reported health status is the result of a censoring process in which patients classify their continuous, underlying health to a limited set of ordered categories. The use of statistical methods that do not acknowledge the ordinal nature of the responses may result in logical inconsistencies, where outcomes are predicted that cannot possibly be derived from the questionnaire (Hernández Alava et al., 2010). Our analytical methods are designed to recognise this.

---

[4] Although this has been disputed - see Brazier et al. (2005).

[5] For example, patients may differ with respect to their pain threshold and therefore give different importance to the provider's ability to reduce pain. However, we have only limited anecdotal evidence from interviews with orthopaedic surgeons to support this claim. There seems to be scope for more systematic evaluation of patient heterogeneity with respect to preference for EQ-5D dimensions.

3 Econometric approach

The objective of the empirical analysis is to obtain measures of the relative systematic impact of hospital providers on patients' post-treatment health outcomes. To address the concerns raised in the previous section, we estimate hierarchical ordered probit models, separately for each of the five EQ-5D dimensions. All models are identical in structure and consist of the same set of covariates. We therefore drop the subscript on dimension and describe the model in general terms.

Let y_{ijt}^* denote the health status (with respect to e.g. anxiety) of patient $i = 1, \ldots, n_j$ in hospital $j = 1, \ldots, J$ 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 a three-point scale ($m = 1, 2, 3$ with $1 =$ no problems, $2 =$ some problems, $3 =$ severe problems)[6]. 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)

$$
y_{ijt} = \begin{cases} 1, & if \ y_{ijt}^* \leq \kappa_1 \\ 2, & if \ \kappa_1 < y_{ijt}^* \leq \kappa_2 \\ 3, & if \ y_{ijt}^* > \kappa_2 \end{cases} \tag{5}
$$

where the threshold parameters κ are unobserved and must be estimated from the data[7].

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 time-invariant factors, such as patient characteristics (e.g. age and gender) and baseline level of latent health. Our interest lies in the 'added' latent health that results from hospital treatment and the degree to which variation in this component can be systematically associated with the provider of care. We make the assumption that, conditional on a set of risk-adjustment factors, patients' health would follow the same trajectory if untreated. This allows us to interpret the difference in latent health between pre- and post-treatment as a relative treatment effect.

Our data are characterised by a hierarchical structure, with measurement points clustered in patients, which themselves are clustered in hospital providers. Given the non-linear nature of our model, these hierarchical 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 provider effect. This, bar the provider effect, is the approach adopted by Contoyannis et al. (2004) in their study of dynamics of self-reported health in the British Household Panel Survey. Alternatively, one can treat both pre- and post-treatment latent health as left-hand side variables and estimate panel models with unobserved patient heterogeneity. We adopt the second approach because it allows us a) to explicitly account for reporting heterogeneity and unobserved determinants of latent health, b) to utilise information contained in both observations to estimate cut-off parameters, and c) to acknowledge random noise in reported pre-treatment health.

---

[6]The mobility dimension is unusual in describing level 3 as 'confined to bed', a matter to which we shall return in the results section

[7]In econometric modelling of ordinal responses, these are usually arranged in increasing order, i.e. from worst to best. We decided to break with this practice and order responses according to their EQ-5D values to retain the familiar interpretation. As a consequence, smaller values of y_{ijt}^* indicate better health status and coefficient estimates have to be interpreted accordingly.

Latent functional status is then described by the outcome equation

$$y_{ijt}^{*} = \alpha_{ij} + \zeta_j + x_{ij}^{'}\beta + T^{'}\nu_j + T*(x_{ij}^{'}\delta + z_j^{'}\theta) + \epsilon_{ijt} \tag{6}$$

with

$$\nu_j = \mu + \gamma_j \tag{7}$$

The vector x_{ij} is a set of patient-level control variables and z_j is a vector of provider characteristics, both of which are, in our case, time-invariant and assumed to be strictly exogenous[8]. Treatment is modelled as a dummy variable T, which takes a value of 1 if $t = 1$ (post-treatment) and 0 otherwise. We interact this indicator with patient-level controls to allow for differential effects of patient characteristics on health status at baseline and on the effect of treatment. Likewise, we interact T with z_j to ensure that hospital characteristics do not affect patients' health before admission.

Unexplained variation is decomposed into four variance components: i) a patient-specific intercept $\alpha_{ij} \sim \mathcal{N}\left(0, \sigma_\alpha^2\right)$ that captures unobserved, time-invariant patient heterogeneity[9], ii) a provider-specific, time-invariant intercept $\zeta_j \sim \mathcal{N}\left(0, \sigma_\zeta^2\right)$ that addresses hospital clustering, iii) a random coefficient $\gamma_j \sim \mathcal{N}\left(0, \sigma_\gamma^2\right)$ that varies between hospitals and describes hospital quality effort, and iv) a serially uncorrelated error term $\epsilon_{ijt} \sim \mathcal{N}\left(0, 1\right)$ that leads to the well-known probit specification.

Our interest lies in estimates of the relative quality of each hospital, γ_j, captured by the provider-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 (Skrondal and Rabe-Hesketh, 2009)[10]. Hospital performance can then be described in two ways: First, we can rank providers according to their impact on latent health status y_{ij1}^*. This can be directly inferred from $\widehat{\gamma}_j$, where smaller values indicate better performance. Second, we can compute the probability of reporting a specific post-treatment outcome, conditional on the estimated quality effort of the provider. For the average patient treated in a hospital of average characteristics, this is given by

$$Prob\left(y_{j1} = m | \overline{x}, \overline{z}, \widehat{\gamma_j}, \widehat{\alpha_i} = \widehat{\zeta}_j = 0\right) = \Phi\left(\kappa_m - S_{j1}\right) - \Phi\left(\kappa_{m-1} - S_{j1}\right) \tag{8}$$

where

$$S_{j1} = \widehat{\mu} + \overline{x}^{'}\widehat{\beta} + \overline{x}^{'}\widehat{\delta} + \overline{z}^{'}\widehat{\theta} + \widehat{\gamma_j} \tag{9}$$

and $\kappa_0 = -\infty$, $\kappa_3 = +\infty$.

---

[8]There exists no formal test to verify the assumption of exogeneity (Greene and Hensher, 2010, p.278). However, we have experimented with fixed effects on the provider-level and found coefficients to be virtually identical. We conclude that the random effects approach is, at least, not more biased than a fixed effects approach. Note that patient fixed effects are ruled out by the low number of observations on this level and the resulting incidental parameter bias.

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

[10]This method is known as Empirical Bayes prediction and is "undoubtedly the most widely used method for assigning values to random effects" (Skrondal and Rabe-Hesketh, 2009, p.665). It combines prior information about the distribution of γ_j with the observed data to obtain posterior means. However, in contrast to a fully Bayesian approach, the prior is not independent of the data but based on the variance estimates obtained from the ML estimation, hence 'empirical'.

Both methods produce identical rankings of relative hospital performance. However, only the second method relates the result back to the original scale of the questionnaire and allows investigation of differences between providers with respect to the probability of achieving a favourable health outcome. Note that, due to the non-linear nature of the model, these probabilities depend on the severity of the patient. Accordingly, one would expect differences between providers to be more pronounced for more severe patients than for their healthier counterparts.

All models are estimated by maximum likelihood, where the integrals for the random effects are approximated by adaptive quadrature. 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.

4 Data

Our study builds on patient-level EQ-5D data for hip replacement patients collected as part of the PROM survey during the period April 2009 to March 2010. All providers of NHS-funded care are required to participate in the survey (Department of Health, 2008). This includes all NHS-operated hospitals and, where applicable, independent sector treatment centres. Patients aged 15 or over that undergo elective, unilateral hip replacement surgery are invited to take part in the survey (see NHS Information Centre (2010a, pp.22-28) for inclusion criteria). 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.

The PROM survey takes the form of a questionnaire and comprises generic and condition-specific instruments as well as questions regarding the patients' medical background. The EQ-5D is the preferred generic instrument and consists of two parts: the EQ-5D descriptive system and the EQ-5D visual analogue scale (EQ-VAS). We focus our attention on the descriptive system which describes impairments in overall health through self-assessed limitations on mobility, self care and usual activities as well as pain and discomfort, and anxiety and depression. For each of the five dimensions, patients can indicate whether they have *no* problems, *some/moderate* problems and either *extreme problems* (in the case of the pain/discomfort and anxiety/depression dimensions), are *unable to* (self care and usual activities) or are *confined to bed* (mobility). Responses on each dimension are translated into numeric values ranging from 1 to 3, with 1 corresponding to no problems. A patient's health profile can then be described as a series of numerical values, e.g. 11221 representing a patient that has some problems performing usual activities and experiences moderate pain or discomfort but reports no problems on any other health dimension.

We link the PROMs data to the Hospital Episode Statistics (HES) inpatient database, which contains detailed information on all inpatient care provided in English hospitals. The unit of observations in HES is the episode of care under the supervision of one consultant (FCE). In order to obtain the full level of patient information documented across the inpatient stay, we link all associated FCEs and create provider spells (Castelli et al., 2008). We retain only spells that can be matched to the PROM dataset and for which we observe a full EQ-5D health profile at baseline and follow-up.

The depth of information contained in HES allows us to account for a wide range of

clinical characteristics and patient demographics. For example, based on documented procedure codes (OPCS-4.5) we construct indicator variables for the type of hip replacement and whether it was a primary or revision surgery. Further, we generate counts of non-duplicate, secondary diagnosis (ICD-10) and procedure codes within a spell as broad controls for co-morbidities and complications. We account for patient demographics by constructing variables based on patient's age (in years) and age squared as well as creating an indicator variable for male gender. To characterise a patient's approximate socio-economic background, we record the income deprivation profile for the patient's neighbourhood of residence as measured by the Index of Multiple Deprivation (Neighbourhood Renewal Unit, 2004).

At provider level, we account for a set of production constraints that are hypothesised to affect hospitals' ability to provide high quality care and, hence, influence patients' health outcomes in favourable ways. Larger hospitals may be able to capitalise on their size and attract high-skilled surgeons or cross-subsidise expensive surgical equipment. There is no consensus in the literature on the existence of a scale effect (Katz et al., 2001; Solomon et al., 2002) and we generate a measure of size based on the total count of FCEs provided by the hospital. Regulatory bodies may force hospitals to offer a wide range of services or provide a substantive degree of emergency care. As a consequence, these providers may not be able to profit from specialisation and stream-lined production processes (Royal College of Surgeons, 2007). To address economies of scope, we create an index of specialisation that reflects the dispersion of Healthcare Resource Group (HRGs) treated within the hospital (Daidone and D'Amico, 2009). This measure resembles a Gini index and is bound between zero (no specialisation; equal number of patients in each HRG) and one (all patients of hospital j fall into one HRG). Further, to measure demand uncertainty, we calculate the percentage of hip replacement patient that receive elective care. Finally, we categorise hospitals into teaching and non-teaching facilities based on the classification system adopted by the National Patient Safety Agency (2011).

5 Results

5.1 Descriptive statistics and transition matrix

Our sample consists of 24,568 patients treated in 153 NHS and private providers. We present descriptive statistics of patient and provider characteristics in Table 1.

Table 1 about here

Elective hip replacement surgery is performed predominantly on elderly patients. This reflects the degenerative nature of the underlying condition, with osteoarthritis and rheumatoid arthritis being the most common reasons for surgical intervention (Singh, 2011). Total replacement of the hip joint is by far the most popular procedure and involves replacement of both the head of the femur and the socket of the pelvis. The majority of patients in our sample are female and admitted for primary replacement of the hip. The observed revision rates of 7.4% are below the national average of 10% reported for England and Wales (NJR, 2010). This discrepancy may in part be explained by differences in the inclusion criteria adopted, the focus on patients treated in English hospitals or the substantial level of non-response in the PROM survey (NHS Information Centre, 2010b, p.10).

We observe large variation across hospitals with respect to scale and scope of operation. The smallest provider in our sample provides care to about 2,000 patients annually, whereas the largest hospital treats more than 300,000 patients. Similarly, we find that hospitals vary substantially with respect to the scope of operation, with some providers focussing only on a handful of distinct activities (here HRGs) and treating exclusively elective patients, whereas others perform nearly 70% of hip replacement surgery on patients admitted through A&E. This heterogeneity can, in part, be explained by the distinct types of providers in our sample, with independent sector treatment centres ($J = 9$) being generally smaller ($mean = 18,741$), providing no emergency care and offering a less diverse set of services than their NHS-operated counterparts.

Table 2 present 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 month after surgery. Accordingly, patients in the lower triangle report improvements in health statuses, whereas those in the upper triangle report deteriorations. Patients on the diagonal provide the same evaluation before and after treatment.

Table 2 about here

Several interesting observations can be made from these data. First, the number of patients benefiting from treatment varies greatly by the health dimension under consideration. For example, the dimension most positively affected by treatment is pain and discomfort, where 72% of the patients in our sample report improvements as indicated by a transition to a more favourable category. This is consistent with clinical expectations and the general understanding that pain reduction (and improvements in physical function) is the most important outcome for (elderly) hip replacement patients (Fitzgerald et al., 2004). In contrast, only about 30% of patients report category improvements on the anxiety dimension.

Second, the idiosyncratic labelling of the mobility question is clearly reflected in the distribution of pre- and post-treatment scores. Only about 100 of nearly 25,000 patients report to be 'confined to bed' at baseline, further reducing to 19 after six months. One would expect that, due to the limited granularity of the mobility question, patients responding as having 'some problems' are more heterogeneous with respect to underlying health than is the case on other health dimensions.

Finally, 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 where about 44% and 57% of patients fall into this category. Interestingly, we find that more than 6% of patients do not report at least 'some problems' with respect to their ability to walk about. While these patients may primarily seek medical attention to alleviate joint-related pain[11], we would have expected to observe some self-reported limitations on mobility for *all* patients in the sample.

5.2 Regression results

Table 3 reports parameter estimates and associated standard errors for each of the five models.

---

[11]127 patients report no problems on mobility or discomfort at baseline. Out of these, 79 patients report 'no problems' on any of the EQ-5D dimensions. This raises the question why these patients have qualified for surgery.

We find several variables to be consistently associated with self-reported health at baseline, including age (-), male gender (+), number of comorbidities (-) and the socio-economic status of the patient's neighbourhood of residence (+). Patients admitted for primary surgery tend to report worse health status than those returning for revision surgery, but this effect is only statistically significant for mobility as well as pain and discomfort. The type of hip replacement procedure to be carried out is not associated with better or worse pre-treatment health status.

The mean effect of treatment on post-treatment latent health is positive and significant for all dimensions other than anxiety. We observe some noteworthy heterogeneity in treatment effect that is associated with observed characteristics of the patient. For example, the number of comorbidities and the indicator for revision surgery are negatively associated with the treatment effect, indicating that treatment is less beneficial for multimorbid or revision patients. Similar, patients living in more deprived areas experience, on average, less improvements in latent health than those residing in higher income areas. Again, we do not observe any statistically significant relationship between the type of hip replacement surgery performed and the effect of treatment.

With regard to provider characteristics, we find a statistically significant effect only for the share of elective patients treated. Hospitals that provide no or only very limited emergency care realise better outcomes than providers facing more uncertain demand. We hypothesise that this is the result of a more foreseeable production process that allows for more patient-tailored care. But, it might also be indicative for cream-skimming, i.e. selection on characteristics that we do not observe of patients most likely to benefit from treatment, by private providers that do not provide emergency care. Given the implicit correlation between the ratio of elective care and private ownership, our results are broadly consistent with the findings of Chard et al. (2011) who report better health outcomes for hip replacement patients treated in independent sector treatment centres compared to NHS hospitals. Similar to Solomon et al. (2002) we do not find evidence of scale effects or superior performance of teaching hospitals.

All variance components are statistically significant as confirmed by likelihood ratio tests. The variance of the provider-level intercept and random coefficient are both substantially smaller than the patient-level and random error component; the latter being fixed at 1. This suggests that the majority of unexplained variation in latent health occurs on the first two levels of the hierarchy and is not systematically associated with the provider of care. The existence of a provider-specific intercept implies that some hospitals treat systematically healthier patients than others. We interpret the statistically significance of the random coefficient on treatment as evidence of variation in hospital quality. The covariance between provider-level intercept and random coefficient is negative in all models, albeit not always significant. This suggests that hospitals providing care to a healthier population realise less improvements in latent health.

5.3   Relative hospital performance

We now turn to the results of the hospital performance assessment. Figures 1a - 1e 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. Hospitals located to the left side of each graph are performing better than those to the right.

10

Figures 1a - 1e about here

The graphical presentation of random coefficients in a caterpillar plot is informative in several ways: First, we find that provider heterogeneity, as represented by the slope of the curve, is more pronounced on dimensions such as mobility or usual activities than on, for example, self care. This is a reflection of the differences in estimated variance components that carry over to the Empirical Bayes estimates. Second, we find that only a small number of hospitals have a statistically significantly different treatment impact compared to the national average, here standardised to zero. However, note that the confidence intervals are only appropriate for comparisons against zero, but are too wide for comparison of any two providers (Goldstein and Healy, 1995). Third, confidence intervals on the mobility dimension are wider than on any other dimension of the EQ-5D. This could be a reflection of the lesser amount of information contained in the data, with only two outcome categories being reasonably well populated.

The shortcoming of this type of analysis of provider-specific random coefficients is its focus on latent health. While it is possible to assess *statistical* significance, one cannot make statements about *clinical* or *patient-perceived* significance. Conversely, hospital-specific probabilities of reporting a given post-treatment health status provide some insights into the likely impact of provider heterogeneity with respect to the scale and outcome of interest. In some cases, we find differences between providers to be quite remarkable. For example, the expected probabilities of reporting 'no problems' on mobility six month after surgery range from about 61% to 77%. In contrast, expected probabilities for the same outcome on the self care dimension are significantly less dispersed and consistently above 90% for all providers. We believe that this display of hospital performance is intuitive and, hence, more accessible to patients and the medical community.

Finally, to explore whether 'good' performance on one dimension is associated with good overall performance, we have calculated the rank correlation of provider-specific random coefficient for each of the five health dimensions. The resulting correlation matrix is presented in Table 4.

Table 4 about here

The correlation between performance estimates is substantial across all dimensions, suggesting that good (and bad) performance is systematic and relates to all EQ-5D dimensions. At the same time, we find correlation to be more pronounced for mobility, usual activities and pain/discomfort than for anxiety/depression or self care. These dimensions are seen as clinically most relevant for this patient group and have shown to be more affected by performance heterogeneity. This would suggest that one performance indicator may be sufficient to inform patient choice, independent of the relative valuation put on each of these three dimensions.

6   Conclusion and discussion

The routine collection of patient-reported outcome data has been long overdue and the PROMs exercise promises to be an important component of efforts to improve health care provision in the English NHS. In this paper, we set out to measure variability in hospital quality and gain a deeper understanding of determinants of patient-reported health outcomes. To achieve these goals, a number of methodological issues need to be

addressed. Some of these issues are well-recognised, and efforts have be made to resolve them, most notably to ensure risk-adjustment (Coles, 2010; Chard et al., 2011). Our paper focuses on three methodological issues that have received less attention to date.

First, rather than focusing on an overall EQ-5D utility score, we argue that it is both more accurate and more informative to assess each of EQ-5D dimensions in its own right. Our approach does not require assumptions to be made about how to aggregate across health dimensions and offers insight about which dimensions are particularly affected by provider heterogeneity. We set out an analytical strategy to explore patient-level and provider-level variation in categorical responses within and across dimensions of the EQ-5D.

Second, policy interest is in assessing the change in health status 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 conduct multilevel analysis with measurement points clustered in patients, which themselves are nested in providers. 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.

Third, in recognition of the expectation that the PROMs data are to be used to inform patient choice, 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 hospital providers. We find more heterogeneity in performance across the mobility, usual activities and pain and anxiety dimensions. Prospective patients who place greater weight on one particular dimension rather than another may use this information to select a provider that has a differentially greater impact on this dimension than its peers do.

Several issues remain that we have not addressed in this study. For example, we have commented on the issue of missing values only in passing. 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 only about 50% of eligible hip replacement patients participate in the baseline survey, with a further 8% dropping out of the subsequent survey. Falsely assuming that this substantial amount of missing values are generated at random could lead to biased inferences from a non-representative population (Little and Rubin, 1987).

Another interesting question is whether the correlation between performance estimates across health dimensions is truly a reflection of hospital excellence or whether other factors play into this. For example, the different EQ-5D questions may, at least partly, be measuring the same underlying construct. A technically challenging but potentially fruitful way to explore this matter is the seemingly unrelated regression framework suggested by Zellner (1962), with error terms and provider effects correlated across EQ-5D dimensions.

Word count: 7,397 including footnotes and references.

References

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

Basu, A. and Manca, A. (in press). Regression estimators for generic health-related Quality of Life and Quality-Adjusted Life Years. Medical Decision Making.

Black, N. and Jenkinson, C. (2009). How can patients' views of their care enhance quality improvements?, *BMJ* **339**: 202–205.

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

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

Castelli, A., Laudicella, M. and Street, A. (2008). Measuring NHS output growth, *CHE Research Paper 43*, Centre for Health Economics, University of York.

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

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

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

Daidone, S. and D'Amico, F. (2009). Technical efficiency, specialization and ownership form: evidences from a pooling of Italian hospitals, *Journal of Productivity Analysis* **32**: 203–16.

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

Dolan, P. (1997). Modeling valuations for EuroQol health states, *Medical Care* **35**(11): 1095–108.

FDA (2009). *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*, US Food and Drug Administration.

Fitzgerald, J. D., Orav, E. J., Lee, T. H., Marcantonio, E. R., Poss, R., Goldman, L. and Mangione, C. M. (2004). Patient quality of life during the 12 months following joint replacement surgery, *Arthritis Care & Research* **51**: 100–109.

Garellick, G., Kaerrholm, J., Rogmark, C. and Herberts, P. (2009). Swedish hip arthroplasty register, *Annual Report 2009 (shortened)*, Sahlgrenska University Hospital. Sahlgrenska University Hospital.

Goldstein, H. and Healy, M. J. R. (1995). The graphical presentation of a collection of means, *Journal of the Royal Statistical Society. Series A* **158**: 175–177.

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

Hernández Alava, M., Wailoo, A. J. and Ara, R. (2010). Tails from the Peak District: Adjusted censored mixture models of EQ-5D health state utility values, *Discussion Paper 10/08*, Health Economics and Decision Science, ScHARR, University of Sheffield.

Holmstrom, B. and Milgrom, P. (1991). Multi-task principle-agent problems: Incentive contracts, asset ownership and job design, *Journal of Law, Economics and Organization* **7**: 24–52.

Iezzoni, L. (2003). *Risk adjustment for measuring health care outcomes*, Health Administration Press, Chicago.

Katz, J. N., Losina, E., Barrett, J., Phillips, C. B., Mahomad, N. N., Lew, R. A., Guadagnoli, E., Harris, W. H., Poss, R. and Baron, J. A. (2001). Association between hospital and surgeon procedure volume and outcomes of total hip replacement in the United States Medicare population, *The Journal of Bone & Joint Surgery* **83**: 1622–1629.

Kind, P. and Williams, A. (2004). Measuring success in health care - the time has come to do it properly!, *Health Policy matters* **9**: 1–8.
URL: *http://www.york.ac.uk/media/healthsciences/documents/research/HPM9final.pdf*

Lafonte, J.-J. and Tirole, J. (1993). *A Theory of Incentives in Procurement and Regulation*, MIT Press, Cambridge, Massachusetts.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

Manca, A., Lambert, P. C., Sculpher, M. and Rice, N. (2007). Cost-effectiveness analysis using data from multinational trials: The use of bivariate hierarchical modeling, *Medical Decision Making* **27**(4): 471–490.

Marshall, M. N., Shekelle, P. G., Davies, H. T. O. and Smith, P. C. (2003). Public reporting on quality in the United States and the United Kingdom, *Health Affairs* **22**: 134–148.

McKelvey, R. D. and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variable, *The Journal of Mathematical Sociology* **4**: 103–120.

National Patient Safety Agency (2011). Organisation patient safety incident reports.
URL: *http://www.nrls.npsa.nhs.uk/EasySiteWeb/getresource.axd?AssetID=62923*

Neighbourhood Renewal Unit (2004). The English indices of deprivation.

NHS Information Centre (2010a). A guide to PROMs methodology, *Provisional monthly Patient Reported Outcome Measures (PROMs) in England*, NHS Information Centre.

NHS Information Centre (2010b). Pre- and post- operative data April 2009 to April 2010 - experimental statistics, *Provisional monthly Patient Reported Outcome Measures (PROMs) in England*, NHS Information Centre.

NICE (2008). *Guide to the methods of technology appraisal*, National Institute for Health and Clinical Excellence.

NJR (2010). *Annual Report 7*, National Joint Registry for England and Wales.

Parkin, D., Rice, N. and Devlin, N. (2010). Statistical analysis of EQ-5D profiles: Does the use of value sets bias inference?, *Medical Decision Making* **30**: 556–565.

Royal College of Surgeons (2007). *Separating emergency and elective surgical care: Recommendations for practice*, The Royal College of Surgeons of England.

Shleifer, A. (1985). A theory of yardstick competition, *The RAND Journal of Economics* **16**: 319–27.

Siegel, J. E., Torrance, G., Russell, L., Luce, B., Weinstein, M. and Gold, M. (1997). Guidelines for pharmacoeconomic studies: Recommendations from the panel on cost-effectiveness in health and medicine, *PharmacoEconomics* **11**: 159–168.

Singh, J. A. (2011). Epidemiology of knee and hip arthroplasty: A systematic review, *The Open Orthopaedics Journal* **5**: 80–85.

Skrondal, A. and Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models, *Journal of the Royal Statistical Society. Series A* **172**: 659–87.

Smith, P. C. (2002). Developing composite indicators for assessing health system efficiency, *in* OECD (ed.), *Measuring up - Improving health system performance in OECD countries*, OECD Publications Service, chapter 14, pp. 295–316.

Solomon, D. H., Losina, E., Baron, J. A., Fossel, A. H., Guadagnoli, E., Lingard, E. A., Miner, A., Phillips, C. B. and Katz, J. N. (2002). Contribution of hospital characteristics to the volume–outcome relationship: Dislocation and infection following total hip replacement surgery, *Arthritis & Rheumatism* **46**: 2436–2444.

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.

Tables and Figures

Variable	Description	Mean	SD	Min	Max
male	= 1, if patient is male	0.406	0.491	0	1
age	Patient's age in years	68.415	10.462	15	96
diagtot	Number of coded co-morbidities	2.237	2.066	0	24
opertot	Number of coded secondary procedures	1.356	0.925	0	18
revision	= 1, if revision surgery	0.074	0.261	0	1
thr	= 1, if total hip replacement (base category)	0.877	0.328	0	1
tpr	= 1, if total prosthetic replacement	0.009	0.096	0	1
hpr	= 1, if hybrid prosthetic replacement	0.124	0.329	0	1
imd04i	Index of Multiple Deprivation, income domain	0.124	0.096	0.010	0.830

(a) Patient-level variables (N=24,568)

Variable	Description	Mean	SD	Min	Max
size	Overall number of patients treated (in thousands)	106.693	61.186	2.271	318.558
spec_index	Specialisation index	0.415	0.155	0.233	0.954
teaching	= 1, if teaching hospital	0.150	0.359	0	1
ratio_eleccare	Share of elective vs. emergency patients	0.634	0.138	0.326	1

(b) Provider-level variables (J=153)

Table 1: Descriptive statistics
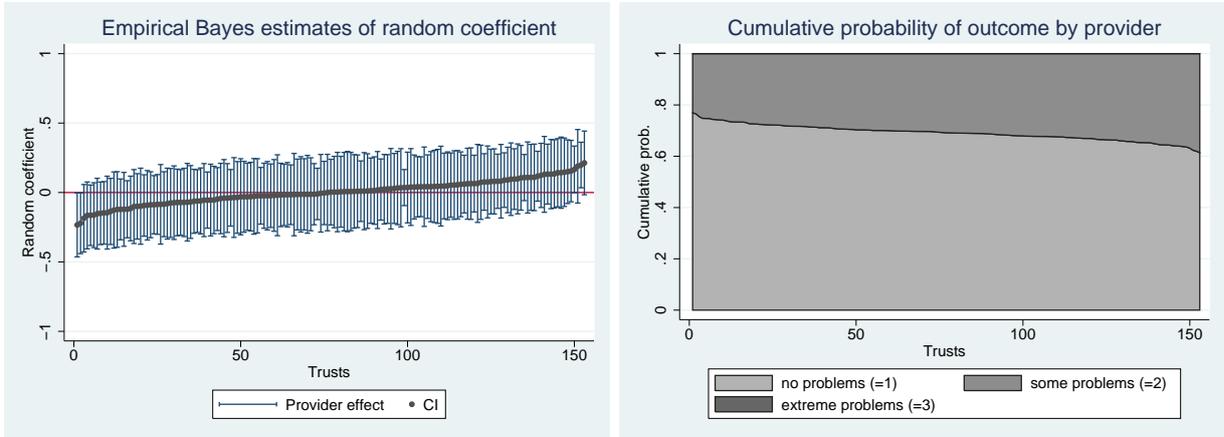
(a) Mobility
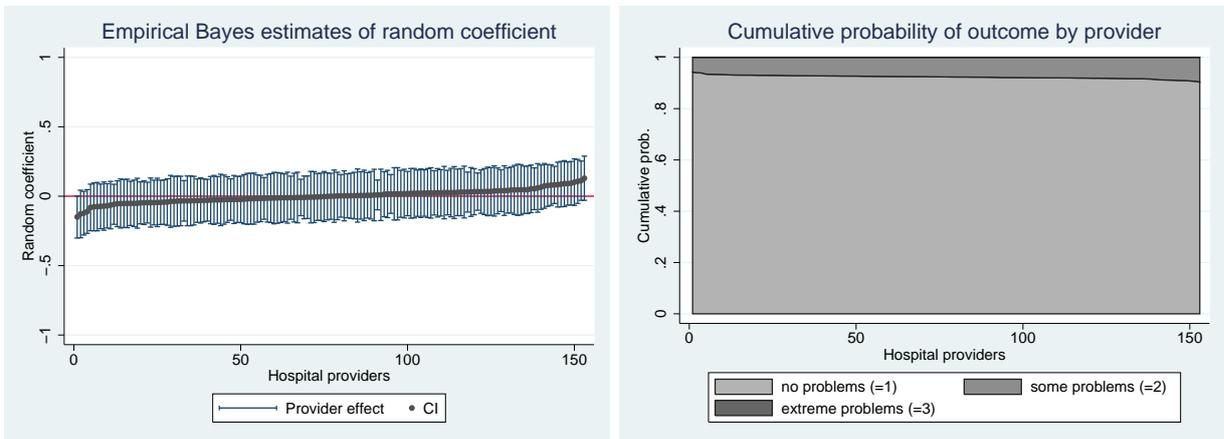
(b) Self-Care

(c) Usual Activities

Figure 1: Performance estimates on the latent health and probability scale (cont'd)

(d) Pain & Discomfort

(e) Anxiety & Depression

Figure 1: Performance estimates on the latent health and probability scale

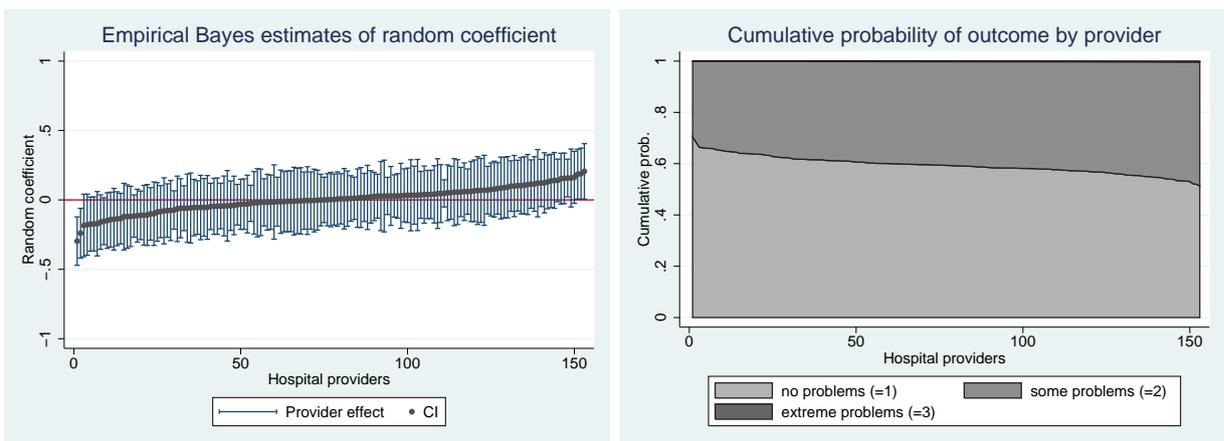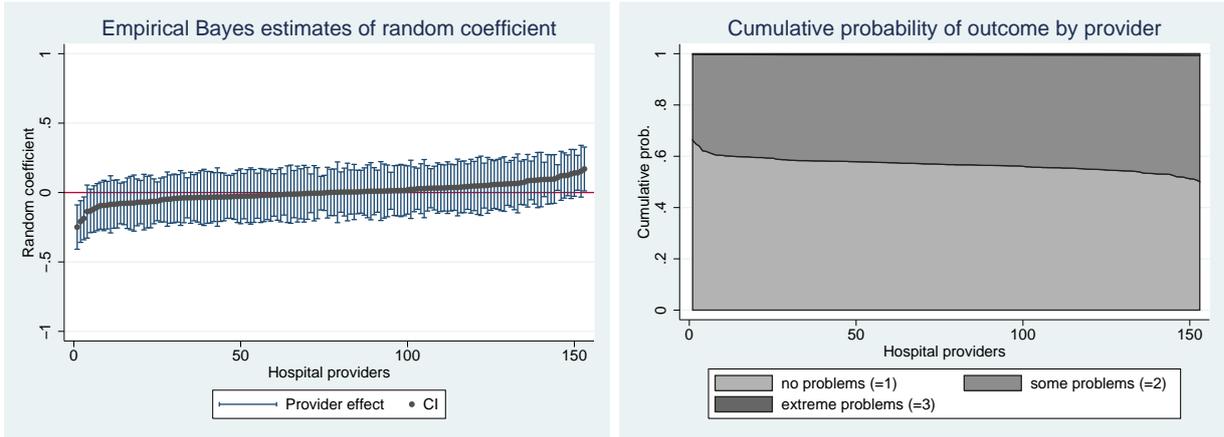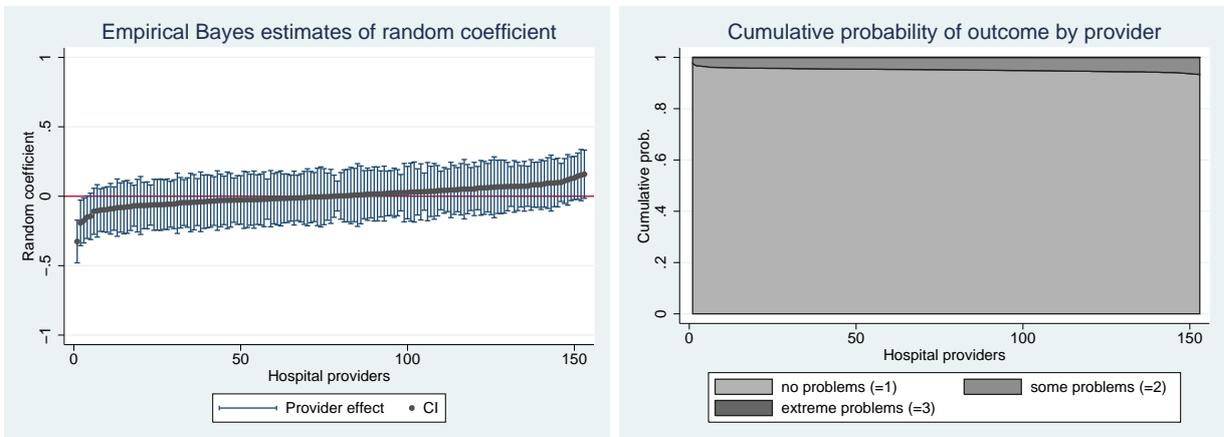|		post-treatment			
		no (=1)	some (=2)	extreme (=3)	Total
pre-treatment	no (=1)	1,258	291	0	1,549
	some (=2)	12,056	10,955	14	23,025
	extreme (=3)	22	92	5	119
	Total	13,336	11,338	19	24,693

(a) Mobility

		post-treatment			
		no (=1)	some (=2)	extreme (=3)	Total
pre-treatment	no (=1)	9,853	1,013	20	10,886
	some (=2)	8,694	4,714	79	13,487
	extreme (=3)	89	171	60	320
	Total	18,636	5,898	159	24,693

(b) Self-Care

		post-treatment			
		no (=1)	some (=2)	extreme (=3)	Total
pre-treatment	no (=1)	1,136	314	27	1,477
	some (=2)	9,670	8,159	490	18,319
	extreme (=3)	1,539	2,731	627	4,897
	Total	12,345	11,204	1,144	24,693

(c) Usual Activities

		post-treatment			
		no (=1)	some (=2)	extreme (=3)	Total
pre-treatment	no (=1)	175	53	3	231
	some (=2)	8,199	5,647	270	14,116
	extreme (=3)	4,360	5,286	700	10,346
	Total	12,734	10,986	973	24,693

(d) Pain & Discomfort

		post-treatment			
		no (=1)	some (=2)	extreme (=3)	Total
pre-treatment	no (=1)	13,014	1,055	60	14,129
	some (=2)	6,301	2,787	230	9,321
	extreme (=3)	547	516	180	1,243
	Total	19,865	4,358	470	24,693

(e) Anxiety & Depression

Table 2: Transition matrices for all EQ-5D dimensions

Table 3: Regression results

Variable	Mobility		Self-Care		Usual Activities		Pain & Discomfort		Anxiety & Depression	
	Beta	SE	Beta	SE	Beta	SE	Beta	SE	Beta	SE
Patient-level variables										
male	-0.161	0.029 ***	0.011	0.023	-0.085	0.019 ***	-0.302	0.018 ***	-0.472	0.024 ***
age	-0.051	0.011 ***	-0.068	0.008 ***	-0.053	0.007 ***	-0.030	0.007 ***	-0.048	0.008 ***
age2	0.000	0.000 ***	0.001	0.000 ***	0.000	0.000 ***	0.000	0.000 ***	0.000	0.000 ***
diagtot	0.080	0.008 ***	0.076	0.006 ***	0.059	0.005 ***	0.068	0.005 ***	0.069	0.006 ***
opertot	0.057	0.018 **	0.046	0.013 **	0.039	0.011 **	0.004	0.011	0.031	0.013 *
revision	-0.225	0.058 ***	-0.043	0.046	-0.038	0.038	-0.259	0.036 ***	0.080	0.047
imd04i	0.871	0.163 ***	1.142	0.121 ***	0.461	0.101 ***	1.189	0.097 ***	1.071	0.121 ***
tpr	0.295	0.167	0.158	0.117	0.121	0.098	0.003	0.092	0.118	0.117
hpr	-0.077	0.045	0.007	0.036	-0.036	0.030	-0.005	0.028	-0.009	0.036
treatment	-1.476	0.434 **	-1.312	0.341 ***	-0.826	0.297 **	-2.320	0.292 ***	-0.365	0.326
treatment*male	-0.073	0.034 *	-0.021	0.028	-0.160	0.024 ***	0.089	0.023 ***	0.177	0.029 ***
treatment*age	-0.015	0.013	0.000	0.010	-0.027	0.009 **	0.010	0.009	-0.011	0.010
treatment*age2	0.000	0.000	0.000	0.000	0.000	0.000 ***	0.000	0.000	0.000	0.000
treatment*diagtot	0.056	0.009 ***	0.043	0.007 ***	0.056	0.006 ***	0.019	0.006 **	0.036	0.007 ***
treatment*opertot	-0.049	0.021 *	-0.010	0.016	-0.016	0.014	-0.027	0.014 *	-0.022	0.016
treatment*revision	0.818	0.067 ***	0.586	0.055 ***	0.593	0.048 ***	0.700	0.047 ***	0.354	0.054 ***
treatment*imd04i	0.656	0.184 ***	0.604	0.145 ***	1.011	0.126 ***	0.027	0.124	0.664	0.143 ***
treatment*tpr	-0.291	0.188	-0.009	0.137	-0.042	0.122	-0.039	0.120	-0.011	0.136
treatment*hpr	-0.011	0.052	-0.035	0.044	-0.068	0.037	-0.069	0.036	0.021	0.044
size	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
spec_index	0.154	0.171	-0.186	0.162	0.026	0.152	0.141	0.131	0.057	0.149
teaching	-0.007	0.054	0.025	0.052	-0.021	0.051	0.011	0.043	0.079	0.047
ratio_eleccare	-0.656	0.196 **	0.006	0.194	-0.496	0.171 **	-0.436	0.155 **	-0.506	0.177 **
κ_1	-2.960	0.371 ***	-1.961	0.270 ***	-3.000	0.229 ***	-3.120	0.219 ***	-1.353	0.267 ***
κ_2	2.164	0.369 ***	1.324	0.270 ***	-0.368	0.228	-0.696	0.218 **	0.721	0.267 **
σ^2_α	0.497	0.035 ***	1.048	0.039 ***	0.450	-0.018 ***	0.297	0.015 ***	1.142	0.038 ***
σ^2_ζ	0.029	0.007 ***	0.034	0.006 ***	0.019	0.004 ***	0.018	0.003 ***	0.022	0.005 ***
σ^2_γ	0.023	0.008 ***	0.010	0.005 **	0.018	0.005 ***	0.013	0.004 ***	0.013	0.004 ***
$cov(\sigma^2_\zeta, \sigma^2_\gamma)$	-0.013	0.007	-0.006	0.004	-0.005	0.003	-0.009	0.003 **	-0.014	0.005
logL	-22,141		-30,739		-36,838		-37,335		-31,948	

Note: Variance components cannot take values smaller than zero. We determine statistical significance through likelihood-ratio tests. In some cases, this involves testing two constraints at the same time. For example, by constraining σ^2_ζ to be zero, we must constrain $cov(\sigma^2_\zeta, \sigma^2_\gamma)$ to be zero as well.

	Mobility	Self-Care	Usual Activities	Pain & Discomfort	Anxiety & Depression
Mobility	1				
Self-Care	0.359	1			
Usual Activities	0.508	0.403	1		
Pain & Discomfort	0.484	0.380	0.465	1	
Anxiety & Depression	0.035	0.211	0.169	0.301	1

Table 4: Spearman's rank correlation matrix