The joint evaluation of hospital cost and quality performance: New evidence from the English PROM initiative*†

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

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

1 Introduction

Public sector organisations often serve different stakeholders and pursue multiple objectives. Assessing the performance of organisations in the context of multiple objectives is complicated by the lack of agreement on the relative importance of each objective and potential trade-offs between them (Hauck and Street, 2006). There is a risk that, in analysing and assessing achievements in isolation, important trade-offs may be overlooked and organisations may be unduly rewarded or punished.

In health care, two common competing objectives are the requirements to contain resource use and provide high quality care. A large body of literature analyses variation in the performance of hospitals on each objective in isolation and provides evidence of significant heterogeneity (Hollingsworth, 2008). Substantially fewer studies have examined the mean relationship between quality of care and costs, with quality typically being measured in terms of mortality or re-admission rates (e.g. Fleming, 1991; Morey et al., 1992; Carey and Burgess, 1999; Picone et al., 2003; Deily and McKay, 2006; McKay and Deily, 2008; Schreyögg and Stargardt, 2010; Gutacker et al., 2011). Their findings have shed some light on the important question whether higher resource inputs are, on average, associated with better quality of care and whether efforts to reduce costs by hospitals may have adverse effects on quality. But, in each of these studies the effect of quality on costs (and vice versa) is averaged across the study sample, thereby precluding consideration of the relationship between performances on both dimensions among individual hospital.

We are only aware of two studies which assess joint hospital performance with respect to cost containment and their ability to provide high quality care. Timbie et al. (2008) use data on 11,259 AMI patients treated in 69 Massachusetts hospitals to model variation in performance with respect to treatment costs and in-hospital survival. They use the expressed value of a statistical life to translate costs and outcomes into incremental net benefits and identify hospitals that provide cost-effective care. Hvenegaard et al. (2011) investigate hospital performance with respect to treatment costs and two measures of quality: 30-day mortality and wound complications. In contrast to Timbie et al. (2008), they do not aggregate costs and outcomes into a single metric but identify the providers' position in the two-dimensional performance space. Using data on 3,754 patients admitted for vascular surgery in six Danish hospitals, Hvenegaard et al. identify two hospitals as begin outliers with respect to cost/wound complication performance but none with respect to costs and mortality. Both studies have in common that they rely on measures of treatment 'failure', that is mortality and wound complications. Consequently, these measures do

*We would like to thank Steven Bloomer and David Parkin for their valuable inputs and comments. 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 programme or the Department of Health.

†Work in progress. Do not quote or cite without the authors' permission.

‡Corresponding author. E-Mail: nils.gutacker@york.ac.uk

not allow for consideration of the quality of care received for the majority of patients that do not experience such events.

The purpose of this paper is to explore the joint performance of National Health Service (NHS) hospitals in England with respect to their risk-adjusted resource use and health outcomes for four surgical procedures. We are interested in several facets of performance, including i) the extent to which resource use and patient-reported outcomes (PROs) vary systematically across NHS providers, ii) whether performance is correlated across dimensions of resource use and patient health outcomes, and iii) whether there are subsets of providers who excel or underperform with respect to their case-mix and caseload adjusted two-dimensional (resource use and outcome) benchmark.

Our study combines data from the English NHS Patient Reported Outcomes Measures (PROM) survey with Hospital Episode Statistics (HES) and Reference Cost (RC) databases for the financial year 2009/10. Since April 2009, all providers of NHS-funded care are required to invite all elective patients undergoing unilateral hip or knee replacement, varicose vein surgery, or groin hernia repair to report on their health status before and up to six months after surgery. Health status is assessed using condition-specific PROMs as well as the generic EQ-5D. We link this information to detailed inpatient records about patient demographics and severity. Resource use is measured as inpatient costs and length of stay. The sample size varies by condition from 6k to 24k patients treated in 136 to 148 NHS hospitals.

We use multilevel seemingly unrelated regression (SUR) models with correlated random effects at patient and provider level to a) allow for unobserved influences on resource use and health outcomes and b) distinguish systematic hospital variation in performance from random noise (Hauck and Street, 2006). Our statistical approach acknowledges that decisions about resource allocation and quality of care are likely to be interdependent. The methodology has been successfully applied in the context of cost-effectiveness analysis based on multi-centre randomised controlled trial data (Nixon and Thompson, 2005; Manca et al., 2007). To our knowledge, though, the approach has not been applied to observational data of resource use and health outcomes. Application to such data requires adjustment for case-mix, severity and relevant hospital production constraints (e.g. volume) in recognition of the non-random allocation of patients to providers.

We obtain Empirical Bayes estimates of provider effects and calculate provider-specific incremental cost/outcome ratios relative to the level of resource use and health outcomes that could be expected given their case-mix and production environment. We identify hospitals that achieve systematically better outcomes at lower costs and those that underperform on both dimensions.

2 Assessing the performance of health care providers

The delivery of health care is characterised by multiple agency relationships, where different principals (e.g. patients, purchasers and funders of care) delegate tasks to agents (i.e. health care providers such as hospitals) in return for a reward (Arrow, 1963; Smith and Street, 2006). The purpose of performance measurement is to assess the degree to which agents fulfil these objectives. When no absolute targets can be formulated, performance is assessed relative to the achievements of other organisations; a process known as benchmarking. Examples for such objectives are the cost of providing a hip replacement or the rate of inpatient mortality following bypass surgery. While the direction of performance is clear, i.e. lower cost and mortality indicate better performance, no absolute target can be defined at which an objective would be considered fulfilled. Instead, a hospital provider will be considered to perform well when observed achievements are better than a reference value derived from other agents (e.g. average cost of care or mortality rate).

Performance assessment of health care providers is typically carried out retrospectively and within a non-experimental, routine care setting. In order to allow for fair comparison across providers, achievements must be adjusted for factors outside the control of the hospital (risk-adjustment). Furthermore, random variation must be disentangled from the systematic, unobserved contribution of the provider so that performance can be attributed.

2.1 Performance assessment of a single objective

We begin our exposition with the simple case of one objective. Let y_j be the observed achievement of hospital $j = 1, \ldots, J$. Achievements are expected to be generated by the following process

$$y_j = f(E_j, Z_j) + \epsilon_j \tag{1}$$

where hospitals make a contribution to the observed achievement through their production decisions or, more generally, the effort, E_j, exerted. The vector Z_j denotes a set of exogenous factors that vary across providers, affect achievement but are not under the provider's control. Examples include production constraints, access to factor markets, the level of technology available, or systematic differences in the population of patients treated (Jacobs et al., 2006). Of course, variation in observed achievement may be random, i.e. due to measurement error or chance. This random variation is represented by the error term ϵ_j. Achievement is linked to the different effects through the function f that remains to be specified. When performance cannot be assessed against an absolute target, as is the case for cost containment or health outcomes, the average level of effort, $\bar{E}$, forms a natural benchmark and $(E_j - \bar{E}) \neq 0$ is a test of hospital-specific performance.

Both the provider's contribution and random variation are unobserved in practice. In order to distinguish the two, repeated measures of performance are required[1]. These multiple observations can be made over time, e.g. in the form of monthly assessment. Alternatively, when objectives are defined with respect to characteristics of individual products or outputs, repeated observations of the production process can be used.

One example of an achievement that is reflected in product characteristics is the cost of care for individual patients. Regulators or purchasers of care require hospitals to produce care efficiently and contain costs. Lower costs are therefore indicative of better performance. Some providers may utilise more resources because their patients present with more severe conditions or are more difficult to treat due to existing co-morbidities. By observing the production of care repeatedly, i.e. for multiple patients, one can distinguish the systematic level of costs of care from other sources such as patient characteristics, production constraints or random variation. In comparing these systematic levels across providers, a regulator or purchaser of care can determine the relative performance of individual hospitals compared to their peers (Dormont and Milcent, 2004; Laudicella et al., 2010).

The model described in equation 1 extends readily to the multilevel case, where achievements are assessed for each patient and patients are naturally clustered within providers. Let y_{ij} be the achievement of hospital $j = 1, \ldots, J$ as reflected in patient $i = 1, \ldots, I_j$, so that

$$y_{ij} = f(E_j, X_{ij}, Z_j) + \epsilon_{ij} \tag{2}$$

where X_{ij} is a set of exogenous factors determining achievements that vary across patients (e.g. demographics, comorbidities) and all other variables retain their original interpretation.

2.2 Multiple objectives

When hospitals pursue multiple, correlated objectives, isolated views on performance are likely to provide an incomplete picture. We extend our previous model to the two-objective case with y_{ijk} being the achievement with respect to cost containment ($k = c$) or health outcome ($k = h$). These objectives may be correlated through common underlying (observed or unobserved) factors, so that

$$y_{ijk} = f(\dot{E}_{jk}, \dot{X}_{ijk}, \dot{Z}_{jk}) + \dot{\epsilon}_{ijk} \tag{3}$$

with

[1] Alternatively, one can separate both effects by making strong, often non-testable parametric assumptions about the distribution of the unobserved random variables. This technique is known as stochastic frontier analysis (SFA) and has been widely applied in the efficiency literature.

$$\dot{X}_{ijk} = x_{ijk} + X_{ij}$$
$$\dot{Z}_{jk} = z_{jk} + Z_j$$
$$\dot{\epsilon}_{ijk} = \epsilon_{ijk} + \vartheta_{ij}$$
$$\dot{E}_{jk} = e_{jk} + E_j$$

where X_{ij}, Z_j and ϑ_{ij}, E_j are characteristics at patient and provider level that are associated with both resource use and health outcomes. In contrast, x_{ijk}, z_{jk}, ϵ_{ijk}, and e_{jk} are factors specific to each objective.

If common factors exist, objectives will be correlated. The sign (and strength) of the correlation depends on the circumstances and objectives. For example, more severe patients may require more resources (higher costs) and benefit less from treatment (lower health outcomes) than their healthier counterparts. In this case, the correlation between the two objectives of cost containment and excellent health outcomes would be negative. Conversely, employing more experienced surgeons or providing better post-operative care may lead to better health outcomes at higher costs, i.e. the correlation is positive. If both effects operate at the same time, the sign of the overall correlation is ambiguous and depends on the absolute effect of each factor on achievements and the relative strength of correlation.

Performance can be expressed as a point in the two-dimensional performance space, $\frac{(\dot{E}_{j,c}-\bar{E}_c)}{(\dot{E}_{j,h}-\bar{E}_h)}$, where the origin is formed by the standardised benchmark, i.e. $\bar{E}_c = \bar{E}_h = 0$[2][3]. This is depicted in Figure 1.

Figure 1 about here

Joint hospital performance can be classified in three different groups: positive outliers, negative outliers, and inliers. Hospitals located in the south-east quadrant are identified as being positive outliers. These hospitals provide care at substantially lower resource use and achieve higher health outcomes than expected given their case-mix and caseload. Conversely, hospitals located in the north-west quadrant are considered negative outliers, as they provide low quality care and utilise more resources than expected/required. Hospitals located in the north-east or south-west quadrants are considered inliers. These inliers could be further distinguished into those that provide care cost-effectively, and those that do not. However, this would require information on the valuation that the principal attaches to post-operative health outcomes, which is often not available[4].

3 Methods

We use random effects multilevel models to distinguish between systematic hospital-specific effects, the impact of differing patient case-mix and random variation in both cost and health outcomes. The equations are estimated in a Seemingly Unrelated Regression (SUR) framework to allow for correlation between resource use and health outcomes due to common omitted variables (Zellner, 1962). The Empirical Bayes estimates (EBE) produced for hospital-specific effects in cost and health are the main focus of the analysis.

Data are analysed at patient level with patients ($i = 1, \ldots, I_j$) clustered within hospitals ($j = 1, \ldots, J$) with achievements, y_{ijk}, associated with the outcome and cost functions ($k \in [c, h]$).

[2]This expression is comparable in nature to *'incremental cost-effectiveness ratios'* (ICERs) often calculated in cost-effectiveness research and as such is amenable to similar presentational techniques and interpretations, e.g. cost-effectiveness plane plots. The primary difference between ICERs as calculated for the assessment of the cost-effectiveness of new medical technologies and our measures of performance are that our comparator is the risk-adjusted benchmark, not another technology or placebo.

[3]This approach can readily be extended to multiple objectives with $k > 2$. In this case, provider performance estimates are represented by vector coordinates in k-dimensional space.

[4]Note that we do not observe patients' health status repeatedly after treatment, i.e. we have no information on the patients' health profiles over time. This preludes calculating quality-adjusted life years (QALYs) for which a valuation is available.

We assume a linear additive relationship between achievements, y_{ijk}, and a hospital specific intercept, γ_{jk}, a constant term, α_{0k}, a set of patient level explanatory variables, X_{ijk} and a set of provider level explanatory variables, Z_{jk}, both of which may differ across equations, such that

$$\begin{bmatrix} y_{ij,c} \\ y_{ij,h} \end{bmatrix} = \begin{bmatrix} \alpha_{0,c} \\ \alpha_{0,h} \end{bmatrix} + \begin{bmatrix} X_{ij,c} & 0 \\ 0 & X_{ij,h} \end{bmatrix} \begin{bmatrix} \beta_c \\ \beta_2 \end{bmatrix} + \begin{bmatrix} Z_{j,c} & 0 \\ 0 & Z_{j,h} \end{bmatrix} \begin{bmatrix} \theta_c \\ \theta_h \end{bmatrix} + \begin{bmatrix} \bar{X}_{j,c} & 0 \\ 0 & \bar{X}_{j,h} \end{bmatrix} \begin{bmatrix} \kappa_c \\ \kappa_h \end{bmatrix} + \begin{bmatrix} \gamma_{j,c} \\ \gamma_{j,h} \end{bmatrix} + \begin{bmatrix} \epsilon_{ij,c} \\ \epsilon_{ij,h} \end{bmatrix} \tag{4}$$

where β_c,β_h,θ_c and θ_h are vectors of parameters to be estimated which relate patient and provider characteristics to health outcomes and costs. The standard assumption regarding the use of random effects in such models is that they are uncorrelated with the included explanatory variables. If this assumption does not hold, i.e. hospital effects are correlated with systematically different patient case-mixes, coefficient estimates may be biased. To address this potential problem we adopt the approach proposed by Mundlak (1978) and include an additional set of covariates in the regression model[5], which are set to the values of the cluster means, $\bar{X}_{j,c}$ and $\bar{X}_{j,h}$. The purpose of the inclusion of the cluster means is to model the linear correlation between covariates and random effects explicitly, thereby resulting in estimates of β and θ that are equivalent to those obtained from the fixed effects SUR estimator.

By estimating the model as a system of equations in a SUR framework rather than each equation in isolation it is possible to measure directly the covariance and correlation in the error terms at patient and provider level. The random effects at patient level are assumed to be a draw from a Bivariate Normal Distribution such that

$$\begin{pmatrix} \epsilon_{ij,c} \\ \epsilon_{ij,h} \end{pmatrix} \sim BVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\epsilon,c} & \rho_1\sigma_{\epsilon,h}\sigma_{\epsilon,c} \\ \rho_1\sigma_{\epsilon,c}\sigma_{\epsilon,h} & \sigma^2_{\epsilon,h} \end{bmatrix} \right) \tag{5}$$

and at provider level

$$\begin{pmatrix} \gamma_{j,c} \\ \gamma_{j,h} \end{pmatrix} \sim BVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\gamma,c} & \rho_2\sigma_{\gamma,h}\sigma_{\gamma,c} \\ \rho_2\sigma_{\gamma,c}\sigma_{\gamma,h} & \sigma^2_{\gamma,h} \end{bmatrix} \right) \tag{6}$$

with $\rho_1 \in [-1, 1]$ being a measure of correlation at patient level and $\rho_2 \in [-1, 1]$ measuring correlation at provider level. Correlation between error terms, which would occur with $\rho_1 \neq 0$ and/or $\rho_2 \neq 0$, indicates common omitted variables across cost and outcome equations at the relevant level. For example, if there were an omitted patient characteristic which made healthy outcomes less likely whilst at the same time increasing costs, we would expect ρ_1 to be negative.

The correlation coefficient, ρ_2, is of greater interest and reflects a population-averaged relationship between the estimates of $\hat{\gamma}_{j,c}$ and $\hat{\gamma}_{j,h}$. These capture the provider-specific effects after allowing for the hospital's patient case-mix and relevant production constraints and form our estimate of the provider-specific deviation from the benchmark, i.e. $(E_{jk} - \bar{E}_k)$. Values of $\hat{\gamma}_{j,c}$ above [below] zero indicate that, all other things being equal, the hospital has above [below] average costs, increasing in the magnitude of the coefficient. Similarly, values of $\hat{\gamma}_{j,h}$ above [below] zero indicate that, all other things being equal, the hospital has above [below] average health outcomes. A positive correlation coefficient ρ_2 would indicate that higher costs / longer length of stay are generally associated with higher health gains (and vice versa) i.e. a trade-off between resource use and quality. A negative correlation would indicate that efforts to contain costs go along with efforts to provide high quality care. A zero correlation would indicate no relationship.

All models are estimated in MLwiN[6] using Iterative Generalised Least Squares (IGLS), equivalent of full maximum likelihood. We obtain EBE of $\hat{\gamma}_{j,c}$ and $\hat{\gamma}_{j,h}$ with accompanying 95% credible intervals in post-estimation, with estimated variance components and parameters substituted in for the unknown population parameters (Skrondal and Rabe-Hesketh, 2009). Providers are considered

[5]However, because the coefficients on these covariates are not of intrinsic interest, we do not report them in our regression results.

[6]Data management and visualisations are done in Stata 12. The link between Stata and MLwiN is provided by the user-written program `runmlwin` (Leckie and Charlton, 2011).

to be situated within one quadrant of the performance space if both credible intervals do not contain zero.

4 Data

Our study combines data from the English Hospital Episode Statistics (HES) inpatient database with the Patient Reported Outcome Measures (PROM) survey and the Reference Cost (RC) databases for the financial year 2009/10. The Hospital Episode Statistics database includes detailed information on all NHS-funded inpatient care provided by public and private hospitals in England. We extract data on all elective patients, aged 18 or over, who underwent unilateral hip or knee replacement, varicose vein surgery or groin hernia repair. For the purpose of this study, all data are recorded at the level of inpatient spells[7] and we retain only complete records, i.e. where all relevant information is recorded.

4.1 Health outcomes

One important outcome of health care is the patient's health status after receiving care. We measure health outcomes after surgery using condition-specific PROMs as well as a generic measure, the EQ-5D. Data are derived as part of the PROM survey, where all eligible patients undergoing one of the aforementioned procedures are invited to participate (Department of Health, 2008a). Patients that have consented to participate are sent a paper questionnaire three or six months after surgery via mail. The length of the follow-up period depends on the surgical procedure. Table 1 provides an overview of the PROMs that are currently collected.

Table 1 about here

The EQ-5D descriptive system[8] is a widely used generic measure of health-related quality of life (Brooks, 1996; Kind et al., 2005) and is collected for all four surgical procedures in the PROMs programme. Patients are asked to indicate whether they experience *no*, *some*, or *extreme* problems with respect to five health dimensions (mobility, self-care, usual activities, pain & discomfort, anxiety & depression). The answers form the patient's health profile, which can be aggregated to an index score using a UK-specific set of weights. These weights are derived from the general public and reflect societal preferences (Dolan, 1997). The resulting index scores range from -0.542 to 1, where one is defined as perfect health and zero is defined as equivalent to being dead. Values smaller than zero are interpreted as health states that are considered worse than being dead.

The condition-specific instruments are the Oxford Hip and Knee Scores (OHS, OKS) and the Aberdeen Varicose Vein Questionnaire (AVVQ). The OHS and OKS each comprise 12 questions that reflect limitations in health-related quality of life brought about by the specific joint (Dawson et al., 1996, 1998). Responses to each question are recorded on an ordinal scale ranging from 0 to 4, where four indicates no problems. An overall score is calculated by weighting questions equally and summing across answers. Accordingly, the overall scores range from 0 (worst) to 48 (best). The AVVQ contains 13 questions which can also be aggregated into an index score (Garratt et al., 1993). This index score lies between 0 and 100 with higher numbers indicating worse health states. In order to facilitate interpretation and comparison of estimation results, we recode the AVVQ scores so that 0 (worst) and 100 (best).

No condition-specific PROM is collected for groin hernia repair.

[7] One of the idiosyncrasies of the Hospital Episode Statistics database is the recording of inpatient activity at consultant level. When patients are treated by multiple consultants, more than one "Finished Consultant Episode" (FCE) is generated. This is typically the case when patients require multi-disciplinary care and are transferred between specialities within the same hospital. In order to record all relevant information on admission, discharge or comorbidities and to keep in line with international literature on hospital activity, we link all FCEs arising from admission to discharge to inpatient spells (Lakhani et al., 2005).

[8] The PROM survey also collects data by means of the EQ-5D instrument's visual analogue scale, the EQ-VAS, but we do not consider this further.

4.2 Resource use

Resource use is measured as either inpatient costs or length of inpatient stay (LoS). We derive information on hospital costs from the Reference Cost database. This is an annual compilation of cost data that forms the basis for the calculation of reimbursement tariffs and whose completion is compulsory for all NHS-operated hospitals. LoS is derived directly from HES.

Reference costs are measured using a top-down costing methodology (Department of Health, 2010). Total hospital costs are assigned to increasingly more granular levels of a hierarchy of costing centres; beginning at treatment services, to specialities and finally to individual Healthcare Resource Groups (HRG, the English equivalent of DRGs). Costs at HRG-level are reported separately for departments and are further broken down by admission type (day case, elective and emergency care) and length of stay, where HRG-specific trim points are used to differentiate short, usual and long inpatient spells.

Costs are often preferred to LoS as a proxy of resource use because they take into account different valuation of input factors, e.g. differences in wages or price variation in intermediate goods. However, the true costs of production are often difficult to assess and the reported estimates of costs may thus be prone to measurement error or aggregation bias. This is especially true when costs are not collected at the most granular level and subsequently added up (i.e. bottom-up or patient-level costing), but broken down according to accounting rules (Carey and Burgess, 2000). LoS, while being a blunter measure of resource use, is less likely to be affected by these limitations. We therefore explore the sensitivity of our results to the choice of resource use measure.

All costs are adjusted for the Market Forces Factor (MFF) specific to the provider. This adjustment takes into account unavoidable variation in input prices across the country as defined by the Department of Health (Department of Health, 2008b).

Because private providers do not submit RC reports, we limit our study to public, NHS-operated hospital providers.

4.3 Covariates

We derive a generic set of risk-adjustment variables that reflect patient severity and are used to model both resource use and health outcomes for all four conditions. The primary measure of severity is the patient's self-reported health status before surgery, i.e. the pre-operative PROM. These data are collected during the last outpatient appointment preceding the surgery or at the day of admission and are recorded in the same format as the post-operative outcomes. We extract information on age, gender, and number and type of comorbidities (ICD-10) from the HES inpatient dataset. The latter are used to construct the (weighted) Charlson index of severity (Charlson et al., 1987) as well as the number of additional diagnoses coded. We also record whether the surgery was a revision (based on OPCS 4.5 procedure codes) or whether the patient was treated by multiple consultants during the hospital stay. The patient's socio-economic status is approximated by the income deprivation profile of the neighbourhood in which the patient resides (i.e. the Index of Multiple Deprivation, IMD)(Neighbourhood Renewal Unit, 2004).

We also construct indicator variables for the five most common HRGs to which patients are allocated and group all other observations in the category 'other'. HRGs are, by design, homogeneous with respect to the expected level of care and resource consumption as well as the clinical procedures involved. They are therefore expected to explain a majority of variation in observed costs or LoS. In contrast, HRGs are not designed or validated to categorise risk-profiles with respect to health outcomes. We therefore include HRG dummies only in the resource use equations, not in outcome equations.

Finally, we calculate the volume of patients treated for each of the four conditions by the respective provider. Volume has been identified as one potential driver of resource use and health outcomes. Given the excess demand faced by most English NHS hospitals, we expect volume to be outside of the providers' control, at least in the short run, and therefore adjust all performance estimates accordingly.

5 Results

5.1 Descriptive statistics

Table 2 reports descriptive statistics for patients undergoing each of the four procedures. Our overall sample consists of 66,974 patients with knee replacement being the most frequently observed procedure and varicose vein surgery the least frequently observed. Patients are clustered in 136 to 148 providers and the average number of patients treated by each provider ranges from 96 (varicose vein surgery) to 240 (knee replacement surgery).

Table 2 about here

After adjusting for MFF, mean costs are similar for hip and knee replacements (around £6k) and considerably lower for groin hernia repair and varicose vein surgery (<£1.5k). Average LoS is also much lower for the latter two procedures, with such patients rarely requiring an overnight stay. Because LoS is measured as full inpatient days, we observe only very limited variation in this measure of resource use for groin hernia and varicose vein surgery.

The EQ-5D scores suggest that average post-operative health status is lower following hip (0.75) and knee (0.69) replacement than it is for hernia repair (0.87) and varicose vein surgery (0.86). But for the latter two procedures, pre-operative scores are quite high too (>0.77). In contrast, pre-operative scores are relatively low for those about to undergo hip (0.34) and knee (0.40) replacement. The change in mean health status as a result of surgery is, then, quite considerably larger for hip and knee replacement than it is for hernia repair or varicose vein surgery. The condition-specific measures, where available, tell a similar story.

Summary statistics for the variables used as risk-adjusters reveal different patterns in the age and gender profiles of the patients undergoing each procedure. They also differ in terms of the complexity measures, with those having hip or knee replacement having higher Charlson scores and more diagnoses than the other patients. There are no obvious differences across procedures in terms of the socio-economic deprivation profile of the neighbourhood in which patients reside.

The HRG variables are specific to each procedure. Upwards of 75% of patients having each type of procedure are allocated a single HRG which forms the reference category in each regression.

5.2 Estimation results

5.2.1 Unilateral hip replacement

Estimation results for those who had a hip replacement are presented in Table 3. There are four sets of results, reflecting the different combinations of health outcome instrument and resource use measures in the SUR model. The estimated effects of the explanatory variables on either EQ-5D or OHS are not sensitive to the choice of resource use measure (cost or LoS). In contrast, the estimated impact on cost or LoS does appear slightly sensitive to the choice of outcome measure, this being because of how 'initial health' is measured in these estimations.

Table 3 about here

In general, significant predictors of post-treatment health are not sensitive to the choice of PRO instrument. As would be expected, poorer outcomes are related to a higher weighted Charlson score, more diagnoses, and lower socio-economic status. Better outcomes are observed for those with higher initial health status and for men. The impact of age on post-treatment health status is not highly significant and direction is sensitive to the choice of instrument.

As regards resource use measures, older age is a significant predictor of longer LoS but is only weakly related to higher cost. Both LoS and costs are higher for those with more diagnoses and if a revision surgery is performed. As expected, there are also significant differences in resource use related to the HRG to which patients are allocated. The direction and size of these differences are consistent with the tariff payments for these HRGs. Resource use is lower for patients presenting in a better state of initial health, with the effect being stronger for LoS than costs.

Some of the variables are significant predictors of higher LoS but not of costs, namely male gender, weighted Charlson score, socio-economic status and whether the patient has been under the care of more than one consultant.

After controlling for patient characteristics, there remains significant unexplained variation among patients in health outcomes and resource use, however these are measured. This is indicated by the significance of $\sigma^2_{\epsilon,h}$ for outcomes and $\sigma^2_{\epsilon,c}$ for resource use. The correlation between unobserved factors at patient level is negative, i.e $\rho_1 < 0$. This suggests that there exist common, unobserved patient characteristics that have a positive effect on health outcomes and a negative effect on resources used (and vice versa). This is not the case when jointly analysing variation in EQ-5D and costs.

At provider level, we find more variation among providers in outcomes when these are measured using EQ-5D than the OHS - see $\sigma^2_{\gamma,h}$. There is significant variation among providers in resource use, whether defined as LoS or cost - see $\sigma^2_{\gamma,r}$. There is no evidence at provider level of correlation between unobserved factors driving health outcomes and resource use, however these are measured.

Figure 2 presents scatter plots of the performance estimates in the two-dimensional performance space for all four resource use and outcome measures combinations. Each point represents a hospital and we only present credible intervals for those providers that are identified as positive or negative outliers.

Figure 2 about here

Only one provider is identified as being a positive outlier when resources are measured in terms of costs and health outcome is measured using the OHS. No outliers are identified in any of the other models. We believe that this is due to two factors: First, there is substantially more variation in costs than LoS across providers. Accordingly, more hospitals would be identified as cost performance outliers. Second, the estimates of provider performance with respect to outcomes, $\hat{\gamma}_{j,h}$, are estimated with considerable uncertainty. This is reflected in the large confidence intervals around the provider effect estimates. Hence, only one provider can be identified as providing above average quality of care.

5.2.2 Unilateral knee replacement

Table 4 presents the regression results for knee replacement surgery. The predictors of outcome for those undergoing knee replacement surgeries work in a similar direction to the predictors of outcome following hip replacement surgery. Patients report lower outcomes if they have a higher Charlson score, more diagnoses, and lower socio-economic status. Men and those with higher initial health report better outcomes. In contrast to the hip replacement results, age is positively associated with better outcomes following knee replacement.

Table 4 and Figure 3 about here

The direction and significance of the variables predicting LoS and costs for patients undergoing knee replacement are consistent with those for patients having hip replacement. The only difference in results between these two sets of patients is that initial health, as measured using the EQ-5D, does not appear a significant predictor of the costs of knee replacement.

In general, the variance parameters at both patient and provider level tell a similar story to those for the hip replacement sample. The exception is the lack of correlation at patient level between costs and both the EQ-5D and OKS.

We identify three hospitals that are positive outliers, i.e. achieve health outcomes better than expected while using less resources than expected (see Figure 3). Two hospitals are positive outliers with respect to joint cost-OHS performance, whereas one (different) hospital is identified as a positive outlier in the LoS-OHS performance space. No negative outliers can be identified.

5.2.3 Varicose vein surgery

Men and those who present in a higher state of initial health report significantly higher levels of post-treatment health status, whether this is measured using the EQ-5D or AVVQ (see Table 6). In contrast, lower post-treatment health status is associated with a higher Charlson score (though this is not highly significant) and more diagnoses. Older patients and those residing in more deprived neighbourhoods report significantly lower post-treatment EQ-5D scores but the effect is not significant for the AVVQ.

Table 6 and Figure 4 about here

With respect to resource use, we find that older patients, those with more diagnoses, and those who are cared for by multiple consultants have significantly longer LoS but not higher costs. Variation in resource use is captured by the HRG to which each patient is allocated, the sign and direction corresponding to the HRG tariffs.

There is significant unexplained variation among patients in both outcomes and resource use, independent of how these are measured. However, unlike for hip replacements, there is no evidence of correlation between outcomes and resource use at patient level.

Over and above that explained by patient characteristics, there remains significant variation in resource use among providers $\sigma^2_{\epsilon,c}$. We find no variation among providers in terms of EQ-5D and the variation when outcomes are measured using the AVVQ is not highly significant. There is no statistically significant correlation at provider level between outcomes and resource use, with the exception of the AVVQ and costs. Also, we do not observe positive or negative outliers with respect to two-dimensional performance (see Figure 4).

5.2.4 Groin hernia repair

The effect of patient characteristics on health outcomes and resource use are largely consistent with the results for other procedures (see Table 5). Males and those reporting better pre-operative health report higher post-operative EQ-5D scores. Post-operative EQ-5D scores are lower for those with higher Charlson scores, more diagnoses and from less well-off neighbourhoods. Both costs and LoS are increasing in age, the Charlson score, the number of diagnoses and for those cared for by more than one consultant. Patients who report better pre-operative health generate lower costs and stay in hospital shorter.

Table 5 about here

There is significant unexplained variation in health and resource use among patients and significant correlation between these measures at patient level. There is also significant unexplained variation in costs among hospitals. Unusually, though, there is no discernible unexplained variation in outcomes among providers, which thereby rules out further analysis of multidimensional provider performance for this condition.

6 Conclusions and discussion

The objective of the paper is to investigate the relationship between hospital-level measures of health outcome and resource use after allowing for patient case-mix and provider caseload. We do this by constructing multi-level health outcome and resource use equations using data from individual patients treated in the English NHS. By formulating the regression models of resource use and outcomes as a system of equations estimated in a SUR framework we are able to gain additional insight into the correlation between unobserved factors that drive both resource use and outcomes.

In keeping with the wider literature we find that the observed patient characteristics conform to a priori expectations in the manner that they affect resource use and outcomes across all four conditions. As expected, we find significant unexplained inter-patient variation in resource consumption and outcomes after allowing for observed characteristics. We also find that for all four conditions there is significant unexplained variation in resource use among hospitals, whether this is measured by cost or LoS. There is also unexplained variation among hospitals in the health outcomes experienced by patients having hip and knee replacement. These results suggest room for improvement among hospitals in both their utilisation of resources and patient outcomes. In contrast, there is no substantial variation among hospitals in outcomes for hernia and varicose vein patients, rendering the information redundant for benchmarking hospital performance for these patients.

One of the novelties of the paper is the use of correlated error terms across equations at patient and provider level. As expected, at patient level we find a negative correlation between costs and outcomes. This suggests that, despite the detailed nature of the observed characteristics - including the initial level of health - there exist omitted factors which drive both resource use and health improvement in opposite directions. The relationship is generally stronger for the LoS specification and in conditions where patients stay longer than one day, i.e. where there is more observed variation in inpatient length of stay.

However, we also find no significant general correlation between resource use and outcomes at provider level for any of the conditions. Plots of the hospital-specific effects for both resource use and outcomes confirm this conclusion with the general mass of plots looking randomly distributed without any obvious systematic relationship. Very few hospitals are identified as positive outliers; none are identified as negative outliers. This would suggest that, overall, there is scope to improve technical efficiency in the provision of elective surgery: some hospitals could achieve better outcomes by learning from other, best practice providers, without increasing costs; similarly, other providers could reduce costs without compromising the quality of the outcomes. This said, there are a few instances of high cost / high quality hospitals that could potentially justify the extra average costs of their procedures by pointing out the expected improved outcomes - whether the gain in health outcomes is worth the additional costs, and therefore optimal for the NHS (i.e. below a NICE-style cost-per-QALY threshold), is difficult to tell given the limited data.

References

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care, *The American Economic Review* **53**: 941–973.

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

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

Carey, K. and Burgess, J. F. (2000). Hospital costing: Experience from the VHA, *Financial Accountability & Management* **16**: 289–308.

Charlson, M. E., Pompei, P., Ales, K. L. and MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation, *Journal of Chronic Diseases* **40**(5): 373–83.

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

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

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

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

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

Department of Health (2010). NHS costing manual 2009/10, The Stationary Office, London.

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

Dormont, B. and Milcent, C. (2004). The sources of hospital cost variability, *Health Economics* **13**(10): 927–39.

Fleming, S. T. (1991). The relationship between quality and cost, *Inquiry* **28**: 29–38.

Garratt, A. M., Macdonald, L. M., Ruta, D. A., Russell, I. T., Buckingham, J. K. and Krukowski, Z. H. (1993). Towards measurement of outcome for patients with varicose veins, *Quality in Health Care* **2**: 5–10.

Gutacker, N., Bojke, C., Daidone, S., Devlin, N., Parkin, D. and Street, A. (2011). Truly inefficient or providing better quality of care? Analysing the relationship between risk-adjusted hospital costs and patients health outcomes, *CHE Research Paper 68*, Centre for Health Economics, University of York.

Hauck, K. and Street, A. (2006). Performance assessment in the context of multiple objectives: A multivariate multilevel analysis, *Journal of Health Economics* **25**: 1029–1048.

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery, *Health Economics* **17**(10): 1107–28.

Hvenegaard, A., Arendt, J. N., Street, A. and Gyrd-Hansen, D. (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.

Jacobs, R., Smith, P. C. and Street, A. (2006). *Measuring Efficiency in Health Care*, Cambridge University Press, Cambridge.

Kind, P., Brooks, R. and Rabin, R. (2005). *EQ-5D concepts and methods: a developmental history*, Springer, Dordrecht.

Lakhani, A., Coles, J., Eayres, D., Spence, C. and Rachet, B. (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.

Laudicella, M., Olsen, K. R. and Street, A. (2010). Examining cost variation across hospital departments–a two-stage multi-level approach using patient-level data, *Social Science & Medicine* **71**(10): 1872–81.

Leckie, G. and Charlton, C. (2011). runmlwin: Stata module for fitting multilevel models in the MLwiN software package, *Technical report*, Centre for Multilevel Modelling, University of Bristol.

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**: 471–490.

McKay, N. L. and Deily, M. E. (2008). Cost inefficiency and hospital health outcomes, *Health Economics* **17**: 833–848.

Morey, R. M., Fine, D. J., Loree, S. W., Retzlaff-Roberts, D. L. and Tsubakitani, S. (1992). The trade-off between hospital cost and quality of care, *Medical Care* **30**: 677–698.

Mundlak, Y. (1978). On the pooling of time series and cross section data, *Econometrica* **46**: 69–85.

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

Nixon, R. M. and Thompson, S. G. (2005). Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations, *Health Economics* **14**: 1217–1229.

Picone, G. A., Sloan, F. A., Chou, S.-Y. and Taylor, D. H. (2003). Does higher hospital cost imply higher quality of care?, *The Review of Economics and Statistics* **85**: 51–62.

Schreyögg, J. and Stargardt, T. (2010). The trade-off between costs and outcomes: The case of acute myocardial infarction, *Health Services Research* **45**: 1585–1601.

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. and Street, A. (2006). Concepts and challenges in measuring the performance of health care organizations, *in* A. M. Jones (ed.), *The Elgar Companion to Health Economics*, Edward Elgar Publishing, chapter 30, pp. 317–325.

Timbie, J. W., Newhouse, J. P., Rosenthal, M. B. and Normand, S.-L. T. (2008). A cost-effectiveness framework for profiling the value of hospital care, *Medical Decision Making* **28**: 419–434.

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.

Figures

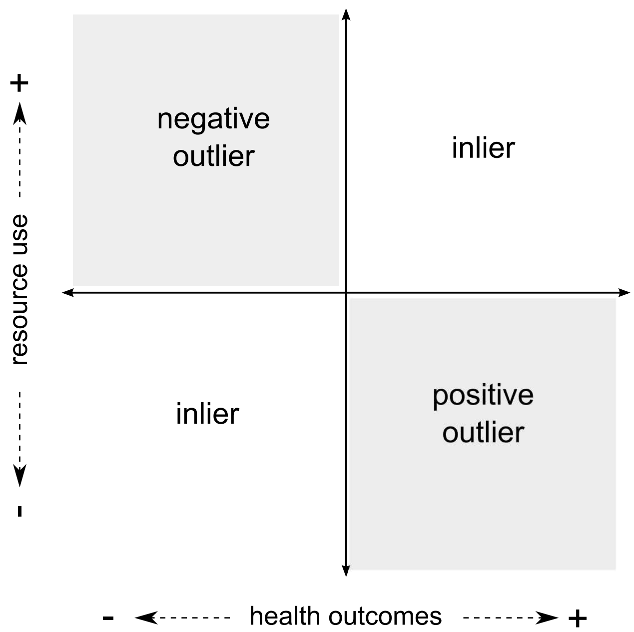

Figure 1: In- and outliers in the two-dimensional performance space

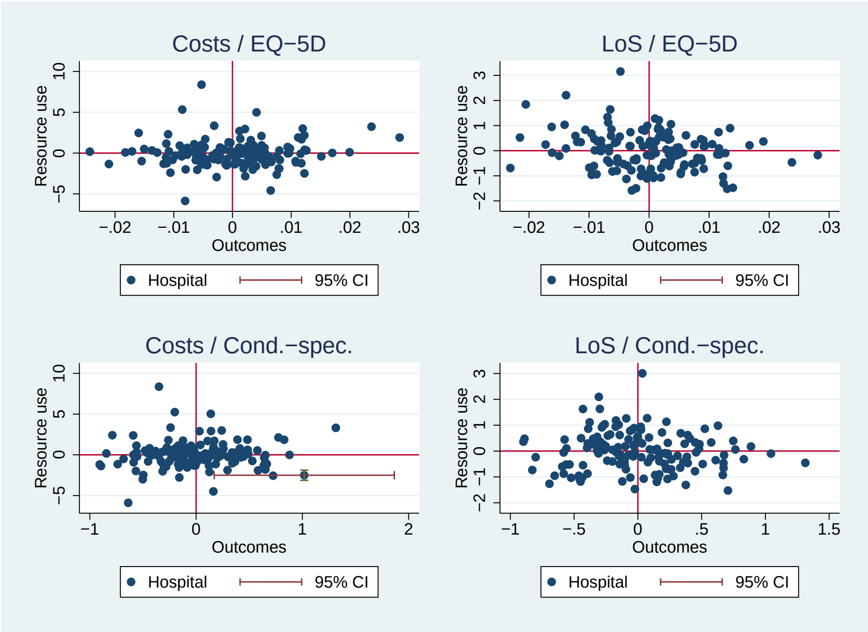

Figure 2: Provider effects - Hip replacement

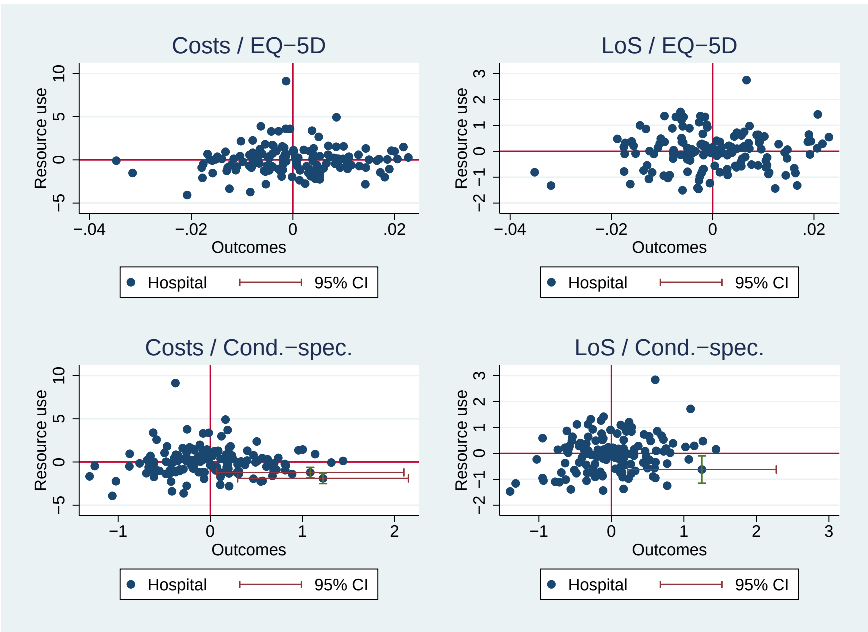

Figure 3: Provider effects - Knee replacement

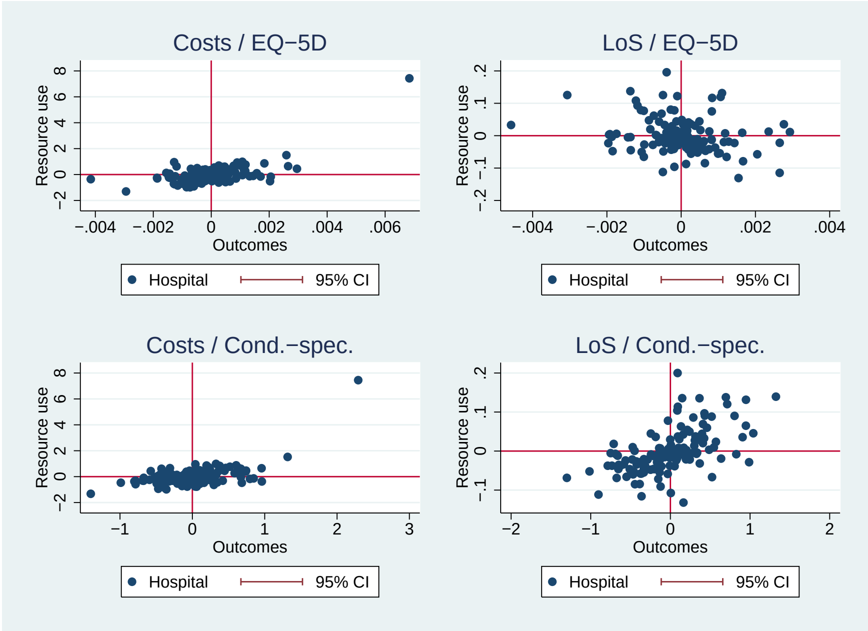

Figure 4: Provider effects - Varicose vein surgery

Tables

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

Table 1: PROM instruments by procedure

		Hip replacement		Knee replacement		Groin hernia repair		Varicose vein surgery	
Variable	Description	Mean / %	SD	Mean / %	SD	Mean / %	SD	Mean / %	SD
Resource use									
costs	Cost of care in 1k GBP, adjusted for MFF	6,104.98	2,053.49	6,058.26	2,114.47	1,435.54	680.37	1,222.94	768.64
los	Length of hospital stay	6.00	4.40	5.82	3.55	0.44	0.98	0.15	0.67
Health outcome									
post_EQ5D	Post-operative EQ-5D index score	0.75	0.26	0.69	0.27	0.87	0.19	0.86	0.21
post_cond_spec	Post-operative condition-specific measure	37.68	9.55	33.22	10.23	-	-	89.06	9.92
Explanatory factors									
pre_EQ5D	Pre-operative EQ-5D index score	0.34	0.32	0.40	0.31	0.79	0.20	0.77	0.21
pre_cond_spec	Pre-operative condition-specific measure	17.99	8.39	18.64	7.70			81.11	10.16
age	Age of patient in years	68.29	10.61	69.15	9.14	61.56	14.65	52.28	14.06
male	Indicator for male gender	0.41	0.49	0.44	0.50	0.93	0.25	0.36	0.48
wcharlsum	Weighted Charlson index	0.33	0.67	0.37	0.68	0.18	0.51	0.09	0.32
add_diagnoses	Number of additional diagnoses	1.99	1.85	2.04	1.86	0.92	1.33	0.46	0.94
revision	Indicator for revision surgery	0.08	0.27	0.05	0.22				
SES_income	Indice of Multiple Deprivation (IMD) - income domain	0.13	0.10	0.14	0.10	0.13	0.10	0.14	0.11
multiepi	Indicator for multiple consultants during hospital stay	0.01	0.12	0.01	0.11	0.00	0.04	0.00	0.03
hrg_1	Indicator for most common HRG*	0.82	0.39	0.85	0.36	0.82	0.39	0.75	0.43
hrg_2	Indicator for second most common HRG	0.05	0.21	0.05	0.22	0.14	0.35	0.11	0.31
hrg_3	Indicator for third most common HRG	0.04	0.19	0.05	0.21	0.02	0.14	0.06	0.24
hrg_4	Indicator for fourth most common HRG	0.04	0.19	0.03	0.17	0.00	0.06	0.05	0.22
hrg_5	Indicator for fifth most common HRG	0.03	0.16	0.00	0.06	0.00	0.05	0.01	0.10
hrg_other	Indicator for any other HRG	0.04	0.19	0.01	0.12	0.01	0.11	0.02	0.13
volume	Number of patients treated by the provider	221.67	132.72	239.57	134.48	156.60	93.36	96.34	75.04
N		21,275		23,916		15,687		6,096	
J		144		145		148		136	

Table 2: Descriptive statistics

	LoS / EQ-5D		Costs / EQ-5D		LoS / OHS		Costs / OHS	
Variable	Beta	SE	Beta	SE	Beta	SE	Beta	SE
Health outcome								
constant	0.625	0.101 ***	0.624	0.101 ***	30.588	4.067 ***	30.468	4.064 ***
age	0.000	0.000 *	0.000	0.000 *	-0.016	0.006 **	-0.016	0.006 **
male	0.018	0.003 ***	0.018	0.003 ***	0.752	0.123 ***	0.752	0.123 ***
wcharlsum	-0.025	0.003 ***	-0.025	0.003 ***	-0.660	0.092 ***	-0.660	0.092 ***
add_diagnoses	-0.015	0.001 ***	-0.015	0.001 ***	-0.560	0.036 ***	-0.560	0.036 ***
revision	-0.106	0.006 ***	-0.106	0.006 ***	-5.703	0.228 ***	-5.703	0.228 ***
SES_income	-0.213	0.018 ***	-0.213	0.018 ***	-9.649	0.668 ***	-9.649	0.668 ***
multiepi	-0.001	0.016	-0.001	0.016	-0.029	0.579	-0.029	0.579
volume	0.000	0.000 *	0.000	0.000	0.001	0.001	0.001	0.001
init_health	0.223	0.005 ***	0.223	0.005 ***	0.339	0.007 ***	0.339	0.007 ***
Resource use								
constant	7.908	3.351 *	18.457	5.263 ***	9.791	3.441 **	18.898	5.326 ***
age	0.070	0.003 ***	0.002	0.001 *	0.069	0.003 ***	0.002	0.001 *
male	-0.482	0.053 ***	0.017	0.016	-0.450	0.053 ***	0.019	0.016
wcharlsum	0.369	0.041 ***	-0.003	0.012	0.361	0.041 ***	-0.004	0.012
add_diagnoses	0.477	0.017 ***	0.061	0.005 ***	0.475	0.017 ***	0.061	0.005 ***
revision	6.212	0.239 ***	1.446	0.073 ***	6.224	0.239 ***	1.446	0.073 ***
SES_income	0.782	0.289 **	-0.080	0.088	0.709	0.289 *	-0.085	0.089
multiepi	4.051	0.251 ***	0.064	0.077	4.055	0.251 ***	0.064	0.077
hrg_2	0.275	0.128 *	0.593	0.039 ***	0.295	0.128 *	0.594	0.039 ***
hrg_3	-3.997	0.273 ***	-0.373	0.084 ***	-3.958	0.273 ***	-0.370	0.084 ***
hrg_4	3.260	0.144 ***	2.127	0.044 ***	3.250	0.143 ***	2.125	0.044 ***
hrg_5	-4.602	0.289 ***	-0.892	0.089 ***	-4.561	0.288 ***	-0.887	0.089 ***
hrg_other	1.753	0.158 ***	1.091	0.048 ***	1.750	0.157 ***	1.090	0.048 ***
volume	-0.002	0.001	0.000	0.001	-0.001	0.001	0.000	0.001
init_health	-0.953	0.082 ***	-0.070	0.025 **	-0.040	0.003 ***	-0.003	0.001 **
Patient-level								
$\sigma^2_{\epsilon,h}$	0.056	0.001 ***	0.056	0.001 ***	74.519	0.725 ***	74.519	0.725 ***
$cov(\sigma_{\epsilon,h}, \sigma_{\epsilon,c})$	-0.077	0.006 ***	-0.003	0.002	-3.174	0.223 ***	-0.264	0.068 ***
$\sigma^2_{\epsilon,c}$	13.969	0.136 ***	1.308	0.013 ***	13.955	0.136 ***	1.308	0.013 ***
Provider-level								
$\sigma^2_{\gamma,h}$	0.000	0.000 **	0.000	0.000 **	0.406	0.111 ***	0.405	0.111 ***
$cov(\sigma_{\gamma,h}, \sigma_{\gamma,c})$	-0.002	0.002	0.001	0.004	-0.037	0.074	0.031	0.138
$\sigma^2_{\gamma,c}$	0.688	0.097 ***	2.685	0.319 ***	0.670	0.095 ***	2.683	0.318 ***
N	21,275		21,275		21,275		21,275	
J	144		144		144		144	
LogL	-57,910		-33,048		-134,348		-109,514	

Table 3: Regression results - Hip replacement

	LoS / EQ-5D		Costs / EQ-5D		LoS / OKS		Costs / OKS	
Variable	Beta	SE	Beta	SE	Beta	SE	Beta	SE
Health outcome								
constant	0.211	0.149	0.211	0.149	7.088	6.050	6.924	6.064
age	0.003	0.000 ***	0.003	0.000 ***	0.073	0.007 ***	0.073	0.007 ***
male	0.008	0.003 **	0.008	0.003 **	0.301	0.122 *	0.301	0.122 *
wcharlsum	-0.020	0.002 ***	-0.020	0.002 ***	-0.561	0.089 ***	-0.561	0.089 ***
add_diagnoses	-0.013	0.001 ***	-0.013	0.001 ***	-0.414	0.036 ***	-0.414	0.036 ***
revision	-0.118	0.008 ***	-0.118	0.008 ***	-5.654	0.279 ***	-5.654	0.279 ***
SES_income	-0.182	0.017 ***	-0.182	0.017 ***	-8.413	0.619 ***	-8.413	0.619 ***
multiepi	-0.001	0.017	-0.001	0.017	0.839	0.631	0.839	0.631
volume	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.001
init_health	0.265	0.005 ***	0.265	0.005 ***	0.480	0.008 ***	0.480	0.008 ***
Resource use								
constant	9.582	4.634 *	19.518	9.159 *	11.722	4.634 *	22.094	9.328 *
age	0.067	0.002 ***	0.002	0.001 **	0.067	0.002 ***	0.002	0.001 **
male	-0.280	0.040 ***	0.022	0.014	-0.207	0.041 ***	0.030	0.015 *
wcharlsum	0.241	0.030 ***	-0.010	0.011	0.230	0.030 ***	-0.012	0.011
add_diagnoses	0.374	0.012 ***	0.033	0.005 ***	0.369	0.012 ***	0.032	0.005 ***
revision	1.250	0.100 ***	0.234	0.036 ***	1.208	0.100 ***	0.228	0.036 ***
SES_income	0.821	0.205 ***	0.018	0.074	0.702	0.205 ***	0.000	0.074
multiepi	3.616	0.210 ***	0.036	0.076	3.603	0.210 ***	0.036	0.076
hrg_2	0.698	0.092 ***	0.447	0.033 ***	0.687	0.092 ***	0.446	0.033 ***
hrg_3	3.750	0.097 ***	2.567	0.035 ***	3.753	0.097 ***	2.568	0.035 ***
hrg_4	-0.467	0.120 ***	-2.426	0.044 ***	-0.448	0.120 ***	-2.423	0.044 ***
hrg_5	4.724	0.311 ***	-0.083	0.113	4.647	0.310 ***	-0.084	0.113
hrg_other	1.088	0.184 ***	-0.582	0.067 ***	1.122	0.184 ***	-0.580	0.067 ***
volume	-0.001	0.001	-0.001	0.001	-0.001	0.001 *	-0.001	0.001
init_health	-0.850	0.065 ***	-0.041	0.023	-0.041	0.003 ***	-0.003	0.001 ***
Patient-level								
$\sigma^2_{\epsilon,h}$	0.060	0.001 ***	0.060	0.001 ***	82.379	0.755 ***	82.376	0.755 ***
$cov(\sigma_{\epsilon,h}, \sigma_{\epsilon,c})$	-0.075	0.005 ***	-0.004	0.002 *	-3.005	0.178 ***	-0.064	0.064
$\sigma^2_{\epsilon,c}$	9.094	0.083 ***	1.189	0.011 ***	9.073	0.083 ***	1.189	0.011 ***
Provider-level								
$\sigma^2_{\gamma,h}$	0.000	0.000 ***	0.000	0.000 ***	0.549	0.129 ***	0.553	0.130 ***
$cov(\sigma_{\gamma,h}, \sigma_{\gamma,c})$	0.000	0.002	0.001	0.004	0.071	0.071	0.022	0.148
$\sigma^2_{\gamma,c}$	0.586	0.078 ***	2.715	0.321 ***	0.565	0.076 ***	2.702	0.320 ***
N	23,916		23,916		23,916		23,916	
J	145		145		145		145	
LogL	-60,742.3		-36,783.5		-147,060		-123,151	

Table 4: Regression results - Knee replacement

	LoS / EQ-5D		Costs / EQ-5D	
Variable	Beta	SE	Beta	SE
Health outcome				
constant	0.741	0.075 ***	0.741	0.075 ***
age	0.000	0.000	0.000	0.000
male	0.032	0.005 ***	0.032	0.005 ***
wcharlsum	-0.021	0.003 ***	-0.021	0.003 ***
add_diagnoses	-0.010	0.001 ***	-0.010	0.001 ***
revision	-	-	-	-
SES_income	-0.126	0.015 ***	-0.126	0.015 ***
multiepi	0.019	0.031	0.019	0.031
volume	0.000	0.000	0.000	0.000
init_health	0.373	0.007 ***	0.373	0.007 ***
Resource use				
constant	0.123	0.851	-2.046	2.051
age	0.007	0.001 ***	0.002	0.000 ***
male	-0.028	0.028	0.009	0.015
wcharlsum	0.055	0.015 ***	0.022	0.008 **
add_diagnoses	0.159	0.007 ***	0.036	0.004 ***
revision	-	-	-	-
SES_income	0.125	0.076	0.059	0.040
multiepi	3.827	0.159 ***	0.271	0.084 **
hrg_2	0.164	0.025 ***	0.090	0.013 ***
hrg_3	0.394	0.055 ***	0.287	0.029 ***
hrg_4	0.230	0.124	0.206	0.065 **
hrg_5	0.168	0.131	0.015	0.069
hrg_other	1.082	0.067 ***	0.695	0.035 ***
volume	0.000	0.000 *	-0.001	0.001 *
init_health	-0.200	0.037 ***	-0.090	0.019 ***
Patient-level				
$\sigma^2_{\epsilon,h}$	0.030	0.000 ***	0.030	0.000 ***
$cov(\sigma_{\epsilon,h}, \sigma_{\epsilon,c})$	-0.008	0.001 ***	-0.003	0.001 ***
$\sigma^2_{\epsilon,c}$	0.775	0.009 ***	0.213	0.002 ***
Provider-level				
$\sigma^2_{\gamma,h}$	-	-	-	-
$cov(\sigma_{\gamma,h}, \sigma_{\gamma,c})$	-	-	-	-
$\sigma^2_{\gamma,c}$	0.028	0.004 ***	0.256	0.030 ***
N	15,687		15,687	
J	148		148	
LogL	-14,973.4		-5,097.48	

Table 5: Regression results - Groin hernia repair

	LoS / EQ-5D		Costs / EQ-5D		LoS / AVVQ		Costs / AVVQ	
Variable	Beta	SE	Beta	SE	Beta	SE	Beta	SE
Health outcome								
constant	0.618	0.079 ***	0.615	0.078 ***	52.145	6.156 ***	50.813	6.091 ***
age	-0.001	0.000 ***	-0.001	0.000 ***	-0.012	0.008	-0.012	0.008
male	0.015	0.005 **	0.015	0.005 **	1.291	0.223 ***	1.291	0.223 ***
wcharlsum	-0.021	0.007 **	-0.021	0.007 **	-0.930	0.338 **	-0.930	0.338 **
add_diagnoses	-0.017	0.003 ***	-0.017	0.003 ***	-0.179	0.123	-0.180	0.123
revision	-	-	-	-	-	-	-	-
SES_income	-0.146	0.023 ***	-0.146	0.023 ***	-1.911	1.041	-1.911	1.042
multiepi	0.075	0.067	0.075	0.067	-0.499	3.098	-0.499	3.099
volume	0.000	0.000	0.000	0.000	-0.001	0.003	-0.001	0.002
init_health	0.447	0.011 ***	0.447	0.011 ***	0.528	0.011 ***	0.528	0.011 ***
Resource use								
constant	0.375	0.366	1.957	1.389	0.243	0.547	2.721	2.254
age	0.003	0.001 ***	0.000	0.000	0.003	0.001 ***	0.000	0.000
male	0.013	0.017	0.029	0.011 *	0.018	0.017	0.028	0.011 *
wcharlsum	0.028	0.028	0.031	0.019	0.030	0.028	0.032	0.019
add_diagnoses	0.028	0.010 **	0.018	0.007 **	0.028	0.010 **	0.018	0.007 **
revision	-	-	-	-	-	-	-	-
SES_income	0.159	0.079 *	0.017	0.054	0.158	0.079 *	0.023	0.054
multiepi	6.371	0.235 ***	0.039	0.159	6.368	0.235 ***	0.038	0.159
hrg_2	0.193	0.027 ***	0.211	0.018 ***	0.185	0.027 ***	0.211	0.018 ***
hrg_3	0.097	0.039 *	-0.071	0.026 **	0.097	0.039 *	-0.070	0.026 **
hrg_4	0.088	0.037 *	0.111	0.025 ***	0.085	0.037 *	0.110	0.025 ***
hrg_5	0.463	0.083 ***	0.296	0.056 ***	0.460	0.082 ***	0.295	0.056 ***
hrg_other	0.233	0.062 ***	0.237	0.042 ***	0.222	0.062 ***	0.236	0.042 ***
volume	0.000	0.000	-0.001	0.001	0.000	0.000	0.000	0.002
init_health	-0.046	0.039	-0.023	0.026	-0.002	0.001 *	0.000	0.001
Patient-level								
$\sigma^2_{\epsilon,h}$	0.031	0.001 ***	0.031	0.001 ***	65.392	1.194 ***	65.421	1.194 ***
$cov(\sigma_{\epsilon,h}, \sigma_{\epsilon,c})$	0.000	0.001	0.001	0.001	-0.027	0.064	-0.014	0.044
$\sigma^2_{\epsilon,c}$	0.376	0.007 ***	0.173	0.003 ***	0.376	0.007 ***	0.173	0.003 ***
Provider-level								
$\sigma^2_{\gamma,h}$	0.000	0.000	0.000	0.000	0.688	0.262 **	0.653	0.255 *
$cov(\sigma_{\gamma,h}, \sigma_{\gamma,c})$	0.000	0.000	0.001	0.002	0.028	0.017	0.222	0.111 *
$\sigma^2_{\gamma,c}$	0.007	0.002 ***	0.641	0.079 ***	0.008	0.002 ***	0.643	0.080 ***
N	6,096		6,096		6,096		6,096	
J	136		136		136		136	
LogL	-3,744		-1,648		-27,120		-25,026	

Table 6: Regression results - Varicose vein surgery