Measuring variations in hospital cost while accounting for differences in patient-reported health outcomes: a multilevel approach

Nils Gutacker*, Chris Bojke*, Silvio Daidone*, Nancy Devlin†, David Parkin‡ and Andrew Street*

Preliminary draft – do not quote without Authors' permission

* Centre for Health Economics, University of York.
† Office for Health Economics, London.
‡ NHS South East Coast, Horley.

Abstract

Background: Most studies of hospital efficiency account for provider heterogeneity with respect to case-mix and production constraints. However, these studies have not convincingly addressed the issue of variations in health outcome as a potential explanation for observed costs.

Aims: To identify cost variations across hospital providers that remain after accounting for 'justifiable' sources of heterogeneity, namely differences in case-mix, production constraints and health outcome.

Data: Since April 2009, all providers of NHS-funded care are required to collect patient-reported outcome measures (PROMs) for four surgical procedures (hip and knee replacement, hernia repair, varicose vein surgery) using generic and disease-specific instruments. We combine information on the average health gain experienced by patients at each hospital with Hospital Episode Statistics and Reference Costs data for the financial year 2008/09.

Methods: Multilevel regression models with patients clustered in hospitals, estimated separately for each surgical procedure. We compare estimated provider effects across different model specifications to explore how to incorporate health outcome information.

Results: Preliminary results suggest that some of the variation in costs is indeed due to differences in health outcomes, and the impact for some hospitals can be substantial. Results vary by surgical procedure and by how health outcomes are measured.

Conclusions: PROM data provide insights into why costs of care vary across hospitals and our analysis shows that it is important to account for this information in comparative analyses. The remaining unexplained cost variability suggests scope for improvement in patient health from existing resource use or reductions in cost without harming outcomes from surgery.

Words: 7484; Pages: 17

Acknowledgements: We would like to thank David Nuttall (DH) for his input at various stages of the project. The project was funded by the Department of Health in England as part of a programme of policy research. The views expressed are those of the authors and may not reflect those of the funder.

1. Introduction

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

Comparative cost analysis in a multiple regression framework can help to address the question 'which variation in cost is justifiable' (Keeler, 1990). By benchmarking providers against each other on the basis of their observed costs, a regulator can gain insights into the cost structure and identify the resource implications of heterogeneity (Shleifer, 1985). Over the past three decades, several hundred studies have conducted comparative analyses of hospital costs (Hollingsworth, 2008). While these have contributed to a better understanding of provider heterogeneity with respect to patient case-mix and production constraints, they have not convincingly addressed the issue of variations in quality and, particularly, health outcome as a potential explanation for observed costs (Newhouse, 1994, Jacobs et al., 2006).

Since April 2009, all providers of NHS-funded care are required to collect patient-reported outcome measures (PROMs) for four elective procedures: unilateral hip and knee replacements, varicose vein surgery, and groin hernia repairs (Department of Health, 2008a). Standardised questionnaires, including both a generic instrument (the EQ-5D) and condition-specific instruments, are collected from patients before and 3 or 6 months after surgery. Although PROMs are applicable to all health care interventions where patients' current health is affected, they are particularly valuable for conditions where the likelihood of death is not the most important aspect of the condition or the treatment. PROMs are a major advance in the measurement of outcomes of care and allow more thorough comparison of hospital providers (Kind and Williams, 2004, Devlin and Appleby, 2010). They are a centrepiece in the government's strategy to "establish improvement in quality and healthcare outcomes as the primary purpose of all NHS-funded care" (Department of Health, 2010, p. 21).

Building on this initiative, this paper has two aims. First, we will quantify the amount of cost variation in provision that cannot be explained by case-mix, production constraints and differences in the quality of care provided. We interpret this unexplained hospital heterogeneity as variation in effort to contain cost. Second, we investigate whether or not the average level of patient reported health outcomes can explain some of the observed cost variation across providers and how far this depends on the choice of PROM instrument.

Our empirical approach is to estimate a multilevel model that recognises the clustering of patients within providers. We use these repeated observations of the hospital's production process to distinguish random noise from systematic cost variation attributable to effort. This approach differs from those typically employed in hospital efficiency studies in that it does not require us to specify a production possibility frontier; a task that has been frequently criticised in the past for its distributional assumptions and its sensitivity to modelling choices (Newhouse, 1994, Skinner, 1994). Furthermore, by focussing on single production lines with homogeneous products (e.g. hip replacement surgery) our analysis is less likely to violate the underlying assumption of a common production function across providers (Harper et al., 2001). We expect our results to be more robust to selection bias as we control for case-mix more thoroughly than otherwise possible in classical single-level regression models.

2. Conceptual framework

Social systems are often sufficiently complex to require a less-informed principal to delegate a task to a specialised agent in return for some reward[1]. In the context of the English health care system, one may think of the principal as the Department of Health and the agent(s) as NHS hospital trusts and independent treatment sector centres. The potential agency problems arising in such situations are well known (Lafonte and Tirole, 1993) and occur when principal and agent have different objectives or value them differently and the agent's effort is unobserved. These information asymmetries allow agents to misreport effort and pursue their own objectives.

One way of mitigating the problem of misreporting is to improve the information base, thereby returning to a situation more akin to symmetric information. Shleifer (1985) showed how comparative cost analysis can be used to make inferences about the underlying cost structure. When agents are heterogeneous with respect to their products and production processes, simple comparison does not suffice. Some of the variations in costs are likely to be the result of environmental factors and exogenous constraints on the production process that are outside the control of the agent and do not reflect choices about effort. One would thus expect that "*variation in cost is the norm rather than the exception*" (Jacobs and Dawson, 2003, p. 204). Any conclusions drawn from a naïve benchmark that does not account for such exogenous factors would therefore be biased and the principal risks misjudging the agents' relative performances.

In order to obtain unbiased estimates of the agents' efforts, Shleifer (1985, p. 324) proposes multiple regression of costs on legitimate "*characteristics that make firms differ, and correct[…] for this heterogeneity*". The natural framework for such regression analysis is the industry cost function that underlies all agents' production processes. In line with the literature on hospital costs (e.g. Street et al., 2010), we can specify this hospital cost function as

$$C = C(Y, q, r, w, Z, e) \tag{1}$$

where *Y* is a vector of outputs, *q* is a measure of quality of care provided, *r* and *w* are price vectors for capital and labour, *Z* is a vector of environmental factors that constrain the production process and *e* is the level of effort exerted. This '*behavioural*' formulation extends neo-classical cost functions in that it allows explicitly for deviations from the paradigm of cost minimisation and acknowledges several sources of provider heterogeneity. For expository purposes, we classify these sources of heterogeneity into three groups: patient heterogeneity, production constraints and the quality of care provided.

One obvious reason for variation in production costs is provider heterogeneity with respect to outputs. Hospitals do not produce one homogeneous good or service. Rather, hospitals provide medical services to patients of different severity and medical need. Even within patient groups receiving the same health care intervention, certain patients will require more attention and resources than others because they suffer from more severe conditions or differ with respect to other factors that determine treatment costs, e.g. age, gender or number and type of comorbidities. As a consequence, overall output of a hospital is better described as a mixture of different outputs, each of which is defined by the underlying severity of the patients. Unless patients are randomly allocated to hospitals, some providers may attract a more favourable case-mix than others, which in turns would allow them to reach a similar level of costs while exerting less effort. It is therefore crucial to correct for output heterogeneity in a comparative cost analysis in order to allow for fair comparison.

[1] Such agency relationships exist not only between institutions (e.g. regulators and hospitals) but as well within institutions (e.g. management and medical staff) (Harris, 1977). A better understanding of variations in effort amongst health care institutions is therefore crucial for policy makers and local managers alike.

A second reason why production costs differ across hospitals is because some providers face a more adverse production environment than others. For example, hospitals may differ in their access to factor markets and they may pay different prices for capital and labour inputs. Some of this variation in input prices is arguably not within the provider's control but determined by location or the existing infrastructure. If such exogenous constraints exist, it will be important to control for them in order to make fair comparisons (e.g. Street et al., 2010).

Production costs may also differ across hospitals because of variations in quality of care. Hospitals may be able to reduce the rate of hospital acquired infections by devising efficient quarantine strategies or affect the outcome of surgery by matching complicated cases to experienced surgeons. Assuming that such quality initiatives are costly, providers may have incentives to reduce quality below some standard and misreport the cost savings as resulting from high effort (Chalkley and Malcomson, 1998).

The ability of comparative cost studies to account for quality variation has been limited by its multi-dimensional nature and the inherent difficulties of measurement. Ideally, one would like to measure the effect of hospital treatment on each patient's outcome, i.e. the change in health status induced by health care. PROMs allow us to directly incorporate such measures of hospital quality in our analysis and determine the production costs of health improvement.

3. Econometric approach

3.1. Estimating provider cost functions

Most comparative cost analyses are based on hospital-level cost functions. The limitations of this approach are well known (Newhouse, 1994). In this study, we follow the growing literature on patient-level cost functions that recognise the inherent clustering of patients in hospital production lines (e.g. Dormont and Milcent, 2004, Olsen and Street, 2008, Laudicella et al., 2010). The rationale for this approach is simple: observed hospital output is the sum of all patient treatment. Each patient has specific medical needs that require the provider to alter their production process and tailor care to the individual (Harris, 1977, Bradford et al., 2001). At the same time, production constraints and provider decisions with respect to the general setup impact on all patients conditional on their medical needs. Examples include the cost of cleaning services or the price of labour. This implies that the cost of each inpatient case reflects both the individual contribution of case severity and the common contribution of cost driving factors at the level of the provider. By specifying the cost function at the level of the patient, we can incorporate both specific and general effects in our analysis and control more comprehensively for heterogeneity across patients.

We estimate multilevel models with provider-specific intercepts (Rice and Jones, 1997, Snijders and Bosker, 1999). Patients form the micro (i.e. level 1) observations and hospitals constitute macro (level 2) units. Our main aim is to identify the systematic cost variation at macro level that cannot be explained by case-mix, production constraints and the quality of care provided. We interpret this unobserved provider heterogeneity as variation in effort. Our approach draws from the *'value-added'* literature in education research where interest lies in the specific contribution of schools to pupils' educational outcomes (Goldstein, 1997).

We specify our empirical model as follows:

$$C_{ij} = \alpha_0 + \boldsymbol{X}'_{ij}\beta + \boldsymbol{Z}'_j\delta + \boldsymbol{q}'_j\theta + \gamma_j + \varepsilon_{ij} \qquad (2)$$

where C_{ij} is the cost of care[2] for patient $i = 1, \dots, n_j$ in hospital $j = 1, \dots, J$ adjusted for differences in input prices. The vector $\boldsymbol{X}_{ij}$ contains case-mix controls that vary at micro level and $\boldsymbol{Z}_j$ is a vector of production constraints at macro level. The vector $\boldsymbol{q}_j$ includes regressors of health gain and initial health status as obtained from the PROM survey. α_0 denotes the common intercept term. Unexplained variation is decomposed into two components: i) a random error term ε_{ij} that varies at micro level and is assumed to be distributed as $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon)$ and ii) a provider effect γ_j that captures unobserved heterogeneity at macro level. Given the linear specification of the presented model, the provider effects can be interpreted directly, i.e. represent the amount of cost deviation from the population average. Accordingly, if $\gamma_j < 0$ the provider exerts high effort which results in lower cost than on average expected. Conversely, if $\gamma_j > 0$ this should be interpreted as low effort and regulators might be required to intervene.

In order to assess the sensitivity of provider rankings and estimates of effort to the addition of PROM information, we estimate an alternative model where health outcome information is excluded, i.e. θ is restricted to be zero. We compare estimates of γ_j obtained from the 'full' and 'restricted' models to identify providers for which a naïve benchmark without quality controls provides misleading assessments of cost performance.

3.2. Modelling unobserved heterogeneity

The econometric literature emphasises two classes of models that can be applied in the case of unobserved cluster heterogeneity (Wooldridge, 2002). Fixed effect (FE) models are most common in panel data econometrics and treat the provider effect γ_j as parameters to be estimated from the data. Random effects (RE) models make the additional assumptions that all γ_j are identically distributed random variables and are uncorrelated with the explanatory variables.

Fixed effect estimators (e.g. *within* or *LSDV*) provide consistent estimates of the β parameters independently of the true underlying model. The price for this consistency is that FE estimators only utilise within-cluster information. In contrast, random effects estimators exploit both within- and between-cluster variation and are therefore generally more efficient. However, they are biased when the assumed exogeneity of explanatory variables conditional on the unobserved effect does not hold.

When confronted with clustered data, economists tend to favour the less restrictive fixed effects approach over random effects. Interest is usually confined to the unbiased estimation of β and unobserved heterogeneity is seen as a nuisance rather than of interest in itself. However, for the purpose of comparative cost analysis, we believe that a random effects approach is preferable for three pragmatic reasons.

Firstly, both FE and RE models produce estimates of β that are virtually identical. The Hausman test does not reject the null hypothesis of unbiasedness for three out of four conditions. For the fourth condition (hip replacement), we find that coefficients differ in the magnitude of 1 - 2 GBP; a difference that is economically insignificant. We conclude that bias is not a concern for our results and interpretations.

Secondly, random effects estimators allow for direct modelling of macro level effects such as production constraints and quality of care. In a fixed effects approach, these effects cannot be included because they would be perfectly collinear with the indicator variables or washed out as part of the within transformation. Some studies have employed *Estimated Dependent Variable* (EDV) models to circumvent the problem (Lewis and Linzer, 2005, Laudicella et al., 2010), where fixed

[2] We use the natural unit of costs instead of the logarithmic transformation. The large sample size allows us to address any moderate skew of the cost distribution through the Central Limit Theorem.

effects are obtained from a first-stage regression and subsequently regressed on macro-level covariates. However, this additional regression step makes the results less readily interpretable, adds complexity and modelling uncertainty, is less efficient and requires analysts to "use (or even invent) ad hoc methods to correct their second-step regressions" (Beck, 2005, p. 458). A random effects framework is better suited for the type of analysis that we propose and a common choice in multilevel studies.

Thirdly, in the random effects approach, the provider effects γ_j are typically not directly estimated from the data but predicted from the underlying distribution of the random variables (Skrondal and Rabe-Hesketh, 2009). This method is known as Empirical Bayes (EB) estimation and combines prior information about the parameter values with the information available from the data to obtain posterior means[3]. The resulting estimates of the provider effects (and their confidence intervals) are shrunken towards the mean of the prior distribution, where the amount of shrinkage is determined by the strength of information in the data. When information is sparse, i.e. the number of micro units within a macro unit is low, the posterior means resemble the mean of the prior more closely. Conversely, for macro units containing much information (i.e. large n_j), the results are primarily driven by the data and shrinkage is minimal. Fixed effects estimation does not allow for such shrinkage.

The advantages of Empirical Bayes estimation and shrunken provider effects have long been recognised in the literature on school effectiveness (Aitkin and Longford, 1986, Goldstein, 1997) and more recently in the performance assessment of health care providers (Bojke et al., 2011). Shrinkage is a form of precision-weighting and is therefore a valuable mechanism to account for uncertainty in estimates for hospitals treating a small number of patients. Indeed, shrunken estimates are shown to have lower mean squared prediction error than non-shrunken estimates obtained from fixed effects estimation and are best linear unbiased predictors in linear models with random effects (Efron and Morris, 1973). Acknowledging policy makers that use these results to inform their decisions, we believe that shrinkage is desirable. It concentrates the discussion on those providers for which we can draw conclusions about their cost performance based on sufficient data but does not require us to set arbitrary inclusion cut-offs with regard to cluster size.

4. Data

4.1. Hospital Episode statistics

Our study uses patient level data extracted from the Hospital Episode Statistics database (HES) for the financial year 2008/09 [4]. HES is the primary source of data for all inpatient care provided in NHS hospitals as well as activity in independent treatment sector centres funded through the NHS. The unit of observations in HES is the episode of care under the supervision of one consultant ("*finished consultant episode*" (FCE)). In order to retain 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 select only those spells in which eligible PROM procedures have been performed (see NHS Information Centre (2010, pp. 22-28) for inclusion criteria). Further, we restrict our analysis to NHS providers due to the poor quality of data submitted by the independent sector (Mason et al., 2010).

[3] Unlike a fully Bayesian approach, the prior is formed by the distribution of the random variables where the unknown variance is replaced by its estimate. This contrasts to the Bayesian convention where the prior reflect ex-ante knowledge about the distribution and should be formed before seeing the data.

[4] Unfortunately, we are experiencing a considerable delay in access to Hospital Episode Statistics (HES) for the financial year 2009/10. Based on prior work using previous years of HES data, we feel confident that the resulting bias is minimal. Indeed, we find that identically specified models using 2007/08 HES data produce estimates of provider performance that are highly correlated with 2008/09 results. We intend to provide updated results at HESG Summer 2011.

All patients are allocated to a Healthcare Resource Group (HRG v.4). By design, HRGs are expected to explain a substantial amount of variation in observed costs. The grouping algorithm used by the NHS Information Centre (NHS IC) assigns HRGs on the basis of FCEs rather than spells. We extract information on the HRG of the episode in which the (first) relevant PROM procedure has taken place and construct indicator variables for all HRGs, for which the 4-digit HRG (i.e. before split for complications and comorbidities) appears in at least 0.1% of the observations. Other observations are grouped in the category 'Other HRG'. The most frequent HRG is set as base category in the regressions.

The construction of any classification system necessarily requires a trade-off between parsimony and homogeneity of the resulting groups. As a consequence, HRGs are unlikely to capture all variation across providers. Hence, we include a set of variables that are based on diagnostic codes (ICD-10) and procedure codes (OPCS-4.5). These include the main reason and type of surgery (PROM-specific) and the weighted Charlson index as a measure of co-morbidity (Charlson et al., 1987). Further, we generate counts of non-duplicate, secondary diagnoses and procedure codes within a spell to control broadly for co-morbidities and complications.

We account for patient demographics by sorting patients into age quintiles and create an indicator variable for male gender. To characterise the inpatient stay itself, we construct indicator variables for transfers in and out of hospital, whether the patient is discharged home or not, multi-episode spells and in-hospital mortality.

We construct variables that capture the influence of observed characteristics of the provider and production environment that are likely to constrain the production process. Larger providers may be able to realise economies of scale and we generate a measure of size based on the count of patients treated by the provider. To address economies of scope, we create an index of specialisation that reflects the dispersion of HRGs treated within the hospital (Daidone and D'Amico, 2009). The index resembles a Gini index and is bound between zero (no specialisation) and one (all patients of hospital *j* fall into one HRG). Finally, hospital trusts are categorised into teaching and non-teaching facilities based on the classification system adopted by the National Patient Safety Agency (2009).

4.2. Reference cost

Hospital Episode Statistics do not include information on the cost of care. However, NHS trusts are required to provide information on their cost structure to the Department of Health for the annual compilation of the reference cost schedule and calculation of reimbursement prices. We utilise this data to construct patient-level cost data and assign them to each FCE.

The reference cost report is implemented using a top-down costing methodology. Here, total hospital costs are progressively cascaded down through a hierarchy of costing levels, starting at treatment services, to specialities and finally to individual HRGs. Costs at HRG-level are reported separately for departments and are further disaggregated according to admission type (day case, elective and emergency care) and length of stay, where HRG-specific trim points are used to differentiate between short, average and long inpatient spells. We map the reference cost to our sample according to the algorithm documented in Laudicella et al (2010).

As outlined before, HRGs are assigned for each FCE. As a consequence, multi-episode spells within our sample contain several bits of cost information. In absence of an agreed methodology on how to aggregate cost from FCE to spell level (Daidone and Street, 2011), we assign the cost of the FCE in which the (first) PROM procedure has taken place.

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

(Department of Health, 2008b). By applying this index to the costs reported in the reference cost schedule, we can wash out justifiable variation in input prices directly.

4.3. Patient-reported outcomes

Data from the PROMs programme are published by the NHS IC for all providers of NHS-funded care throughout the period April 2009 - March 2010. The data are obtained by surveying patients before and after their operation. For each hospital, we obtain data on the average health status pre-surgery, post-surgery, and the average change in health after treatment. The NHS IC also provides these averages adjusted for case mix. However, because we undertake our own case mix adjustments, we used the unadjusted data.

The PROMs survey includes both generic and condition-specific instruments for which data are reported separately. Table 1 summarises the PROM instruments used for each procedure that are reported by the NHS IC.

Table 1: PROM instruments by procedure

Procedure	Condition-specific PROM	Generic PROM	Months following surgery for post-op data collection
Unilateral knee-replacement surgery	Oxford Knee Score (OKS)	EQ-5D, EQ-VAS	6 months
Unilateral hip-replacement surgery	Oxford Hip Score (OHS)	EQ-5D, EQ-VAS	6 months
Varicose vein surgery	Aberdeen Varicose Vein Questionnaire	EQ-5D, EQ-VAS	3 months
Groin hernia repair	-	EQ-5D, EQ-VAS	3 months

The EQ-5D is a generic PROMs comprising a set of questions asking patients to indicate whether they have no, some or extreme problems on five dimensions (mobility; self care; usual activities; pain/discomfort; anxiety/depression). These responses are used to describe a patient's EQ-5D health profile. That health profile can be summarised using utility weights[5] obtained from members of the general public (Dolan 1997), anchored at 1 (full health) to 0 (dead), with scores < 0 indicating states considered worse than being dead. The patient also provides their own assessment of their overall health state on a visual analogue scale – the EQ-VAS – from 0 to 100 (worst to best possible health, respectively).

While there are good reasons to prefer generic instruments over condition-specific instruments, one should not *a priori* exclude the latter. We therefore test the sensitivity of our findings to the choice of instrument. Further, we check for non-linear effects of health outcome on costs using different transformations of the outcome score.

5. Results

5.1. Descriptive statistics

We present descriptive statistics in Table 2.

<Table 2>

Each of the four conditions is sufficiently populated to allow for precise estimation of case-mix effects at patient level. In contrast, the number of providers is comparably low (121 to 144 hospitals), re-emphasising the value of multilevel analysis as compared to traditional hospital-level

[5] Note that a foundation in utility theory is not strictly required for comparative cost analysis.

analysis. Furthermore, we observe large variations in cluster size within and across production lines. One would thus expect that provider effects are estimated with varying degrees of precision and that shrinkage can contribute to a more conservative assessment of hospitals' efforts.

The cost of care varies considerably across providers for each of the four analysed procedures. For example, for knee replacement surgery we observe average costs of care by provider that range from below £2,000 to more than £10,000. High cost cases are not confined to one or two providers. Rather, we observe that many hospitals report cost for patients in excess of two standard deviations above the national average. This indicates that these cases are truly high-cost cases and not artefacts of the way local accounting system operate or how costs are assigned to patients. We therefore retain all observations in our sample and do not trim 'outliers' based on observed costs.

The generic nature of EQ-5D and EQ-VAS allows for comparison of health outcomes across conditions. Patients undergoing hip or knee replacement surgery experience substantially larger increases in health status than those receiving groin hernia or varicose vein surgery. This is consistent with the less serious nature of the underlying conditions. We observe disagreement between EQ-VAS and EQ-5D on the direction of health change for the latter groups of patients. Whether this is a result of aggregation or a genuine difference between instruments cannot be explored with our dataset.

5.2. Regression results

5.2.1. Baseline estimates

Table 3 presents regression results from a model with EQ-5D outcome information. The reported Huber-White standard errors are robust to heteroscedasticity. All estimations are performed in SAS GLIMMIX using adaptive quadrature with identity link and Gaussian variance function.

<Table 3>

We find significant coefficients on the majority of HRG variables. This is encouraging, as it indicates that the current reimbursement system is able to distinguish between different types of patients and their expected costs. Several other patient characteristics explain costs over and above the allocated HRG. For example, we observe an age gradient and find that certain types of main diagnoses and procedures are significant predictors of treatment costs. Costs are higher for patients that receive more secondary procedures or suffer from a higher number of comorbidities as well as for patients that are transferred in or out of hospital or not discharged to their usual place of residence.

The results at provider-level are less clear cut. Hospitals with teaching activity face higher costs of production than non-teaching facilities but the effect is statistically significant only for hip replacement and varicose vein surgery. We do not find conclusive evidence that NHS hospitals realise positive economies of scale within production lines. This is somewhat surprising given the substantial differences in volume observed across providers for each of the four conditions. Similarly, hospitals that provide a homogeneous set of services do not seem to be able to economise on the streamlined production process. Rather, our results suggest that providers concentrating on a limited range of activity are more expensive than their counterparts with more diverse activities, although this effect is rarely statistically significant.

With respect to quality, we find that neither health outcome nor initial health status (both EQ-5D) can explain cost variation across providers for any of the procedures analysed. For three out of four surgical procedures, the coefficient on initial health status shows the expected negative sign. Patients that present with higher health status at admission require fewer resources than patients in worse conditions; a result that seems intuitively correct. The effect of health gain on costs is negative in all four models. This would indicate that some providers are not only better at providing

high quality of care leading to better health outcomes but are at the same time more efficient and produce at lower cost. However, no results are statistically significant at the 5% confidence level.

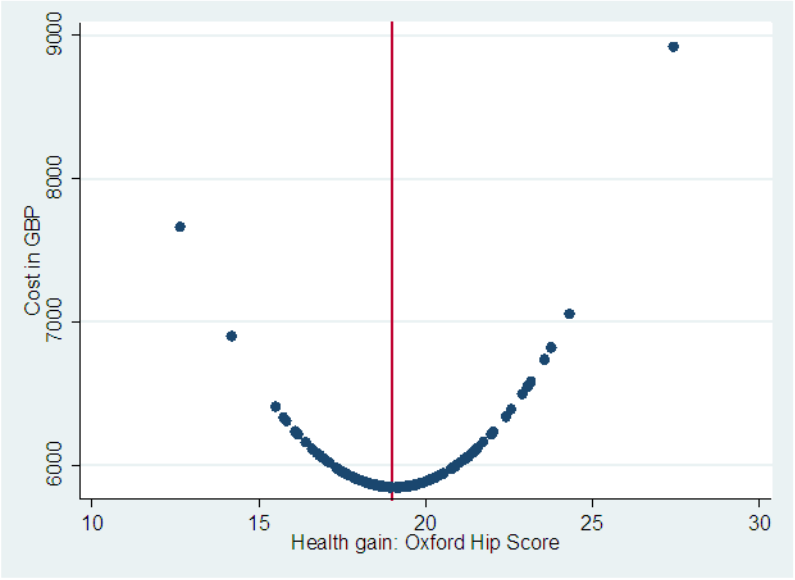
5.2.2. Sensitivity analysis – PROM information

In order to test the stability of our findings, we scrutinise two modelling choices: First, we re-estimate the various models using EQ-VAS and condition-specific PROMs. Second, we test for non-linear effects of health outcome on costs. We compare nested models using the Akaike Information Criteria (AIC). Results of this sensitivity analysis are summarised in Table 4.

<Table 4>

We find strong evidence of a non-linear effect of health gain on costs for hip replacement surgery when outcomes are measured via the condition-specific OHS. The linear and squared terms are both individually and jointly significant. The estimated relationship between costs and quality is U-shaped with initially negative marginal effects that turn positive when average health gain passes a saddle point at 19. This suggests that some providers could substantially improve health outcomes while reducing resource consumption. Figure 1 shows predicted costs across the observed range of outcome scores for a patient treated in a non-teaching hospital with average size, specialisation and initial health minus provider effect.

Figure 1: Relationship between predicted cost and health outcome for Oxford Hip Score



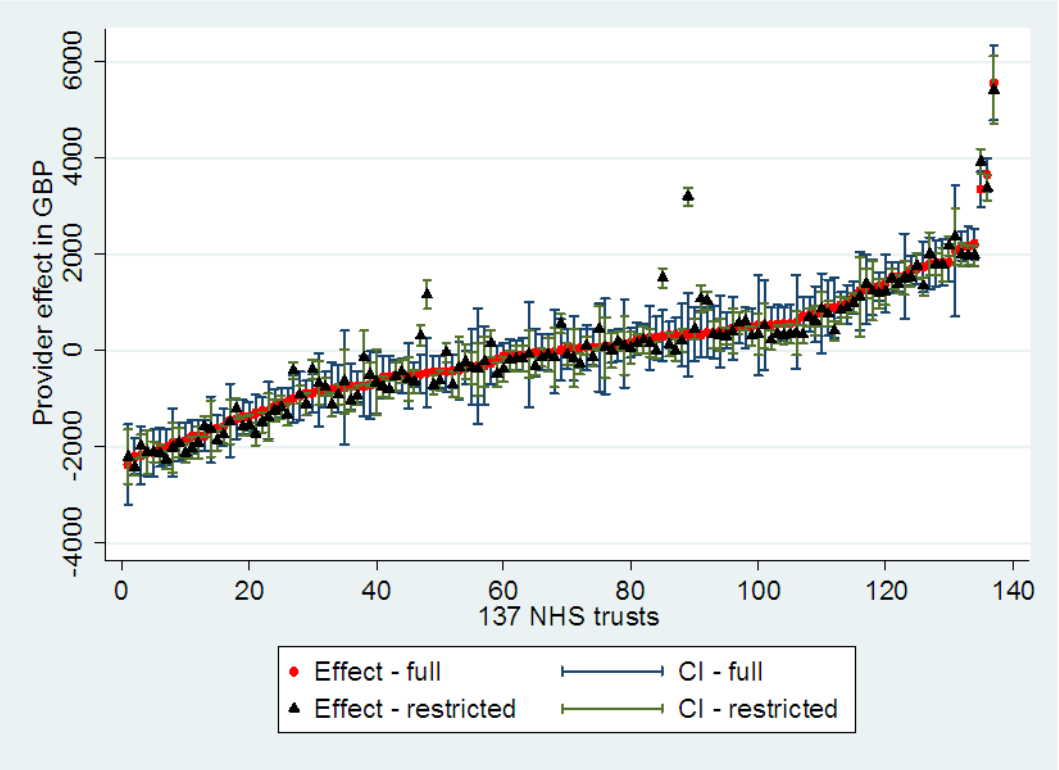
Non-linear effects are also found for knee replacement and varicose vein surgery with health outcomes measured with EQ-5D. However, these effects are only jointly significant and do not improve model fit. There is no apparent relationship between costs and outcomes for any of the other PROM conditions, instruments or specifications. This is re-emphasised by the AIC that favours models without PROM information.

Results have been checked for undue influence of outliers. We have re-estimated all models excluding one provider at a time and found results to remain stable.

5.3. Impact on provider effects

We now turn to the assessment of providers' efforts to contain cost. Figure 2 shows the Empirical Bayes estimates of the provider effects and their corresponding confidence intervals as obtained from models for hip replacement surgery without health outcome information ('restricted model') and with linear and squared OHS health outcome terms ('full model'). Hospitals to the left of the graph have lower costs than hospitals to the right.

Figure 2: Provider effects for hip surgery – OHS



We find substantial differences in provider effects after accounting for case-mix, production constraints and health outcomes. The 'best' hospital has production costs that lie about £2,350 below the national average, whereas the 'worst' hospital produces at more than £5,550 above the average. The spread between 'best' and 'worst' hospital is independent of the choice of PROM.

For the vast majority of hospitals, the additional quality information does not result in significantly different judgements with regard to their cost performance. However, four hospitals experience statistically significant changes in their estimated provider effects as indicated by non-overlapping confidence intervals. The differences between their effects with and without PROM (OHS) regressors range from £800 to about £2,880. Hence, without information on patient outcomes, these providers would be characterised as exerting less effort than they appear to when formulating their output process more thoroughly.

Following the same approach, we have compared 'full' and 'restricted models' for other conditions and instruments. However, we do not find any further statistically significant differences between provider effects. Costs vary between 'best' and 'worst' hospitals as follows:

- Unilateral knee replacement surgery: from -£3,800 to £4,250
- Groin hernia repair: from -£1,150 to £1,450
- Varicose vein surgery: from -£800 to £950

6. Discussion

The aim of this paper is to measure cost variation in the provision of four selected surgical procedures under special consideration of differences in the quality of care provided. Our work builds on a new policy initiative by the English Department of Health to collect patient-reported health outcomes using generic and disease-specific instruments. This study is a first attempt to incorporate health outcomes into comparative cost analysis and explore whether this new measure

of quality changes judgements about the relative performance of NHS hospitals. We make a case for multilevel modelling with precision-weighting and highlight the advantages of this technique over more conventional approaches to efficiency measurement.

Not surprisingly, our results suggest that systematic cost differences exist across hospitals in the provision of surgical procedures. Some of the variation in costs can be explained by different levels of health outcome and we find strong evidence of a non-linear relationship between cost and outcome for hip replacement surgery when outcomes are measured using the condition-specific Oxford Hip Score. For a small group of hospitals, such health outcome adjustment leads to a significant improvement in cost performance. However, for the majority of hospitals, adjustments are minimal and we cannot ascertain whether the effect is merely a statistical artefact or reflects true differences. The relationship between costs and health outcome is less evident for other conditions and instruments.

Several implications for policy makers and future research arise from our results. First, the impact of health outcome information on provider rankings and estimates of cost containment effort is, at best, minimal. However, this may be because our analysis of outcomes is restricted to information averaged at provider level. It will be interesting to see whether this finding still holds for analyses that utilise patient level outcome data. We shall explore this when the data become available to us.

Second, if the relationship between cost and quality is indeed non-linear, pay-for-performance and quality bonus programs have to acknowledge non-constant marginal costs and set different prices for different health outcomes. If the association between outcomes and cost is weak / non-existent (see e.g. groin hernia repair) then quality bonus payments of any form should be understood as incentive payments in excess of production costs. The way in which quality incentive schemes are designed might therefore be quite different for different conditions, i.e. for some conditions a purchaser or commissioner needs to reimburse the additional costs of production to allow the provider to break even whereas in the other cases non-financial incentives may suffice.

Third, at this early stage of the PROM initiative and on the basis of our preliminary analysis, we cannot single out a preferred PROM instrument that should be applied in future analyses of hospital cost performance. We therefore recommend using both generic and condition-specific instruments and conducting sensitivity analysis with regard to the choice of PROM instrument as we have done here.

7. References

Aitkin, M. & Longford, N. 1986. Statistical Modelling Issues in School Effectiveness Studies. *Journal of the Royal Statistical Society. Series A (General),* 149**,** 1-43.

Beck, N. 2005. Multilevel Analyses of Comparative Data: A Comment. *Political Analysis,* 13**,** 457-58.

Bojke, C., Castelli, A. & Nizalova, O. 2011. Exploring the concept of 'avoidable mortality' as a quality indicator for NHS hospital output: the case of circulatory diseases in England. *HESG Winter 2011*. York.

Bradford, W. D., Kleit, A. N., Krousel-Wood, M. A. & Re, R. N. 2001. Stochastic Frontier Estimation of Cost Models within the Hospital. *The Review of Economics and Statistics,* 83**,** 302-09.

Castelli, A., Laudicella, M. & Street, A. 2008. Measuring NHS Output Growth. CHE Research Paper: University of York.

Chalkley, M. & Malcomson, J. M. 1998. Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics,* 17**,** 1-19.

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

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

Daidone, S. & Street, A. 2011. Estimating the costs of specialised care. CHE Research Paper: University of York.

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. *Equity and excellence: Liberating the NHS*. The Stationary Office, London.

Devlin, N. & Appleby, J. 2010. Getting the most out of PROMs. London: The King's Fund.

Dormont, B. & Milcent, C. 2004. The sources of hospital cost variability. *Health Economics,* 13**,** 927-39.

Efron, B. & Morris, C. 1973. Stein's Estimation Rule and Its Competitors--An Empirical Bayes Approach. *Journal of the American Statistical Association,* 68**,** 117-30.

Goldstein, H. 1997. Methods in School Effectiveness Research. *School Effectiveness and School Improvement,* 8**,** 369-95.

Harper, J., Hauck, K. & Street, A. 2001. Analysis of costs and efficiency in general surgery specialties in the United Kingdom. *HEPAC Health Economics in Prevention and Care,* 2**,** 150-57.

Harris, J. E. 1977. The Internal Organization of Hospitals: Some Economic Implications. *The Bell Journal of Economics,* 8**,** 467-82.

Hollingsworth, B. 2008. The measurement of efficiency and productivity of health care delivery. *Health Economics,* 17**,** 1107-28.

Jacobs, R. & Dawson, D. 2003. Variation in unit costs of hospitals in the English National Health Service. *J Health Serv Res Policy,* 8**,** 202-08.

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

Keeler, E. B. 1990. What proportion of hospital cost differences is justifiable? *Journal of Health Economics,* 9**,** 359-65.

Kind, P. & Williams, A. 2004. Measuring success in health care - the time has come to do it properly! *Health Policy matters* [Online]. Available: http://www.york.ac.uk/media/healthsciences/documents/research/HPM9final.pdf.

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

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

Lewis, J. B. & Linzer, D. A. 2005. Estimating Regression Models in Which the Dependent Variable Is Based on Estimates. *Political Analysis,* 13**,** 345-64.

Mason, A., Street, A. & Verzulli, R. 2010. Private sector treatment centres are treating less complex patients than the NHS. *Journal of the Royal Society of Medicine,* 103**,** 322-31.

National Patient Safety Agency. 2009. *Organisation Patient Safety Incident Reports | Cluster types* [Online]. Available: http://www.nrls.npsa.nhs.uk/EasySiteWeb/getresource.axd?AssetID=62923&type=full&servicetype=Attachment [Accessed 27/01 2011].

Newhouse, J. P. 1994. Frontier estimation: How useful a tool for health economics? *Journal of Health Economics,* 13**,** 317-22.

Nhs Information Centre 2010. *Provisional Monthly Patient Reported Outcome Measures (PROMs) in England - A guide to PROMs methodology*.

Olsen, K. R. & Street, A. 2008. The analysis of efficiency among a small number of organisations: How inferences can be improved by exploiting patient-level data. *Health Economics,* 17**,** 671-81.

Rice, N. & Jones, A. 1997. Multilevel models and health economics. *Health Economics,* 6**,** 561-75.

Shleifer, A. 1985. A Theory of Yardstick Competition. *The RAND Journal of Economics,* 16**,** 319-27.

Skinner, J. 1994. What do stochastic frontier cost functions tell us about inefficiency? *Journal of Health Economics,* 13**,** 323-28.

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

Snijders, T. A. B. & Bosker, R. J. 1999. *Multilevel analysis - An introduction to basic and advanced multilevel modeling,* London, Sage.

Street, A., Scheller-Kreinsen, D., Geissler, A. & Busse, R. 2010. Determinants of hospital costs and performance variation: Methods, models and variables for the EuroDRG project. Working Papers in Health Policy and Management: TU Berlin.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data,* Cambridge, Massachusetts, MIT Press.

Table 2: Descriptive statistics

		Knees		Hips		Hernia		Veins	
Variable	**Description**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Patient characteristics									
costMFF	Cost of care, adjusted for MFF	5693.150	1820.49	6205.430	1834.50	1428.340	696.70	1176.520	525.71
patage	Patient age	69.036	9.77	68.805	11.39	58.692	17.38	49.907	14.36
male	=1 if male patient	0.423	0.49	0.391	0.49	0.915	0.28	0.368	0.48
trans_in	=1 if transfer from another provider	0.002	0.05	0.007	0.08	0.001	0.03	0.000	0.02
disdest_other	=1 if discharge to location other than 'home'	0.022	0.15	0.032	0.18	0.004	0.06	0.002	0.04
death	=1 if death during inpatient stay	0.002	0.04	0.003	0.05	0.000	0.02	0.000	0.01
trans_out	=1 if transfer to another provider	0.016	0.13	0.026	0.16	0.001	0.03	0.000	0.02
multiepi	=1 if multiple FCEs within spell	0.022	0.15	0.030	0.17	0.007	0.08	0.002	0.05
opertot	Total number of secondary procedures	1.446	0.97	1.457	1.06	1.364	0.72	1.457	0.82
diagtot	Total number of secondary diagnoses	2.222	2.06	2.250	2.18	1.013	1.52	0.476	0.97
wcharlson	Weighted Charlson index	0.363	0.68	0.351	0.76	0.194	0.57	0.084	0.32
PROM-specific variables (details excluded)	Number of HRGs	14		20		18		11	
	Number of indicators for main procedure	4		8		7		4	
	Number of indicators for main diagnosis	5		5		0		5	
Provider characteristics									
teaching_status	=1 if teaching hospital in 2008-09	0.15	0.36	0.14	0.35	0.17	0.38	0.17	0.38
volume	Number of patients receiving PROM procedure (=cluster size)	469	256.96	391	236.16	408	179.39	196	151.68
spec_index	Specialisation index	0.39	0.10	0.39	0.10	0.38	0.08	0.38	0.08
OKS_hg	Oxford Knee Score - Health gain	14.47	1.94	-	-	-	-	-	-
OKS_q1	Oxford Knee Score - Initial health status	18.40	1.79	-	-	-	-	-	-
OHS_hg	Oxford Hip Score - Health gain	-	-	19.48	1.99	-	-	-	-
OHS_q1	Oxford Hip Score - Initial health status	-	-	17.55	1.72	-	-	-	-
Aberdeen_hg	Aberdeen Varicose Vein Score - Health gain (*)	-	-	-	-	-	-	-8.61	2.40
Aberdeen_q1	Aberdeen Varicose Vein Score - Initial health status (*)	-	-	-	-	-	-	19.56	2.65
EQ5D_hg	EQ-5D (descriptive system only) - Health gain	0.29	0.07	0.41	0.07	0.08	0.03	0.10	0.05
EQ5D_q1	EQ-5D (descriptive system only) - Initial health status	0.39	0.07	0.34	0.07	0.79	0.03	0.77	0.05
EQVAS_hg	EQ-VAS - Health gain	3.07	3.85	8.69	4.97	-0.91	1.94	-0.35	3.04
EQVAS_q1	EQ-VAS - Initial health status	67.67	4.37	65.19	4.16	79.88	2.58	80.27	3.97
Number of observations and cluster size	Number of observations at patient-level	65222		53507		58737		23723	
	Number of observations at provider-level	139		137		144		121	
	Minimum cluster size	45		42		15		8	
	Average cluster size	469		391		408		196	
	Maximum cluster size	1412		1253		1039		984	

(*) Lower score / negative change is better

Table 3: Regression results

	Knees			Hips			Hernia			Veins		
Effect	**b**	**SE**	**p-val**	**b**	**SE**	**p-val**	**b**	**SE**	**p-val**	**b**	**SE**	**p-val**
Intercept	6773.77	1210.4	<.0001	5882.24	1202.5	<.0001	2497.04	930.9	0.008	1088.25	36.8	<.0001
age_cat2	-1.49	13.7	0.913	28.58	15.4	0.064	19.71	6.0	0.001	11.60	6.4	0.070
age_cat3	28.13	14.3	0.049	45.15	14.7	0.002	47.60	7.7	<.0001	19.42	7.8	0.013
age_cat4	30.28	13.4	0.024	54.66	20.2	0.007	83.33	10.6	<.0001	19.29	7.6	0.011
age_cat5	76.31	17.9	<.0001	112.86	20.6	<.0001	182.50	17.5	<.0001	41.23	10.5	<.0001
male	-14.53	9.7	0.136	-18.74	10.3	0.068	1.33	9.7	0.891	10.08	5.8	0.085
trans_in	681.55	285.4	0.017	898.48	217.8	<.0001	42.17	191.7	0.826	-222.56	209.4	0.288
disdest_other	192.40	58.6	0.001	223.19	59.3	0.000	184.39	66.8	0.006	-75.66	82.2	0.357
death	124.13	221.4	0.575	253.51	219.7	0.249	2010.59	770.6	0.009	240.28	38.3	<.0001
trans_out	233.43	58.6	<.0001	175.77	52.1	0.001	388.46	127.5	0.002	1227.11	908.8	0.177
multiepi	-55.02	36.3	0.129	-312.78	100.5	0.002	340.82	73.0	<.0001	281.45	119.0	0.018
opertot	77.46	17.1	<.0001	173.91	21.2	<.0001	84.80	15.5	<.0001	15.29	7.5	0.042
diagtot	28.84	4.6	<.0001	40.68	6.4	<.0001	52.45	5.6	<.0001	53.42	8.7	<.0001
wcharlson	-5.13	8.0	0.521	-17.02	12.0	0.157	70.21	15.0	<.0001	-32.82	13.2	0.013
	PROM-specific diagnoses, procedures and HRGs not reported											
teaching_status	291.40	307.6	0.344	846.48	384.5	0.028	123.39	79.9	0.123	361.23	93.2	0.000
volume	-0.01	0.4	0.989	-0.33	0.5	0.526	-0.09	0.2	0.552	-0.32	0.2	0.070
spec_index	1507.35	1046.5	0.150	2224.27	1087.2	0.041	94.54	414.8	0.820	390.87	437.1	0.371
eq5d_q1	-2295.91	1731.9	0.185	-1043.86	2111.1	0.621	-1519.71	1080.8	0.160	615.29	808.8	0.447
eq5d_hg	-2302.41	1730.0	0.183	-1351.53	1881.5	0.473	-2041.61	1257.2	0.104	-1129.67	714.3	0.114
sigma_u	1301.4			1308.2			378.4			356.0		
sigma_e	895.8			1088.2			490.5			375.4		
rho	0.68			0.59			0.37			0.47		
N	65222			53507			58737			23723		
J	139			137			144			121		

Table 4: Sensitivity analysis – PROMs

PROM	Health gain	Knees				Hips				Hernia				Veins			
		Beta	p-val	jointly sign.	ΔAIC	Beta	p-val	jointly sign.	ΔAIC	Beta	p-val	jointly sign.	ΔAIC	Beta	p-val	jointly sign.	ΔAIC
cond.-spec.	hg	-85.7	0.086	-	0.0	50.3	0.468	-	1.3					5.9	0.698	-	1.9
	hg hg^2	217.5 -10.8	0.285 0.123	no	1.0	-1677.2 44.0	<.0001 <.0001	yes	-6.4					5.3 -3.1	0.727 0.345	no	3.3
EQ-5D	hg	-2302.4	0.183	-	0.0	-1351.5	0.473	-	1.5	-2041.6	0.104	-	-0.4	-1129.7	0.114	-	-0.2
	hg hg^2	2491.2 -10081.0	0.229 0.015	yes	1.0	-5562.9 5293.9	0.560 0.657	no	3.4	-5102.0 17467.0	0.060 0.207	no	0.5	-1122.0 -7841.3	0.063 0.054	yes	0.4
EQ-VAS	hg	-57.5	0.079	-	-1.0	-18.0	0.539	-	1.6	-19.1	0.327	-	0.8	15.0	0.240	-	0.3
	hg hg^2	-87.4 4.2	0.070 0.316	no	0.0	-39.6 1.4	0.380 0.572	no	3.2	-13.8 5.0	0.472 0.218	no	1.7	15.7 0.6	0.211 0.786	no	2.2

Note: ΔAIC is reported with reference to a base model that includes only initial health status but not health gain.