XX Encuentro de Economía Pública
Sevilla, 31 de enero y 1 de febrero de 2013

"Comparing hospital quality performance estimates based on different patient-reported outcome measures"

Ana Luisa Godoy Caballero
Universidad de Extremadura
(anagodoycaballero@unex.es)

Chris Bojke
Centre for Health Economics. University of York
(chris.bojke@york.ac.uk)

Nils Gutacker
Centre for Health Economics. University of York
(nils.gutacker@york.ac.uk)

ABSTRACT

Generic as well as disease-specific PROMs have been collected by hospital providers in the English National Health Service (NHS) since April 2009 for four elective procedures: hip and knee replacement, varicose vein surgery and hernia repair. These measures provide information about the self-assessment of patients' health status. The aim of this study is to compare the provision of health services in the NHS according to the different patient-reported outcome (PRO) measures and to analyse whether our judgement about hospital performance depends on the choice of PRO instrument. In order to do that we carry out a literature review searching for papers that make direct comparisons between the generic, EQ-5D or EQ-VAS, and specific measures OHS, OKS or AVVQ. In addition, we estimate fixed effects models for 20,509 patients treated in 153 hospitals who have completed the questionnaires before and after hip replacement. This methodology will allow for the analysis of hospital effects, i.e. the variation in the measures at provider level. The results show a high positive correlation (over 0.70) between the measures indicating that in general high/low scores in one of the measures are associated to high/low scores in the other measure used for comparisons. However, there are some hospitals judged outliers according to one measure but not the other one.

Keywords: patient outcomes, PROMs, hospital quality, multilevel modelling.

JEL codes: I11, H41

1 INTRODUCTION

Providers of secondary care in the English National Health Service have been required to collect patient reported outcome measures (PROMs) since April 2009. PROMs have been collected before and three or six months after surgery for four elective procedures: hip and knee replacement, varicose vein surgery and hernia repair.

The inclusion of PROMs in the analysis of hospital quality performance, i.e. the quality existing in the provision of health services by the different hospitals, will allow for the consideration of patients' own perspective of their health and-health related quality of life (Devlin and Appleby, 2010). Patients' point of view is important, given that, according to Dawson et al. (1998) "patients provide reliable and valid judgements of health status and of the benefits of treatment" (pp. 63).

The collection of PROMs marks a change in the way performance of secondary providers of care is assessed. Many analyses of provider performance have focused on the analysis of the activity or output of different health centres. However, more recently, the analysis of hospital performance has also included measures of patient outcomes rather than just hospital outputs, with a change in emphasis from the production of health care to the production of health itself (Devlin and Appleby, 2010). Outcome analysis had tended to be limited to measures of mortality or emergency readmissions (Thomas et al., 1994; Dimick et al., 2012; Selim et al., 2002; Chua et al., 2010), but the use of PROMs potentially allows for greater insight into the changes of the Health Related Quality of Life (HRQoL) that a patient may enjoy as a result of hospital activity (Gutacker et al., 2012).

Considering these measures, the aim of the study is to compare the performance of the NHS hospitals. The EQ-5D for technology assessment has been considered to be the preferred instrument by NICE (NICE, 2008). Therefore, by comparing different PROMs, we will be able to identify unusual performing hospitals that warrant further investigation according to the different instruments and answer the question of whether the measure of how we judge hospital performance depends on the choice of the PROMs instrument. More specifically, we will be able to see whether the estimate of individual hospital quality differs when based on a generic or a disease-specific PROM.

In order to carry out the analysis we have structured the paper as follows. In the next section we present the main characteristics of PROMs, as well as the advantages and disadvantages of generic and disease-specific measures. Furthermore, section two also presents a description of the different measures we use in the study.

In section three we present a literature review based on the comparative performance of the generic measures EQ-5D and EQ-VAS and the specific Oxford Hip Score (OHS), Oxford Knee Score (OKS) and Aberdeen Varicose Vein Questionnaire (AVVQ). In addition, this section also presents an outline of comparisons between other generic and specific instruments used in the literature.

Section four presents the empirical approach followed in the study. The empirical analysis is carried out only for one of the procedures for which PROMs are collected: hip replacement. First of all, this section presents the characteristics of the sample of patients undergoing hip replacement. Given that we only consider this intervention, the measures used in the analysis are EQ-5D and EQ-VAS as generic and OHS as disease-specific. Afterwards, we present a description of the methodology applied and a summary of some previous papers considering this methodology for similar purposes.

In section five we present descriptive statistics and the estimation results obtained from a fixed effects model. In this section we also present the comparison of the hospital effects between each of the generic measures and the disease-specific measure, i.e. EQ-5D vs. OHS and EQ-VAS vs. OHS and between the two generic measures, i.e. EQ-5D and EQ-VAS.

Section six concludes the paper.

2 PROMS AS MEASURES OF PERFORMANCE

In this section we present the main characteristics of PROMs in general and the advantages and disadvantages of generic and disease-specific instruments. We then go on to present a description of the instruments currently collected in the NHS PROM survey.

2.1 Generic and disease-specific measures of patient-reported health

The English Department of Health (DH) defines PROMs as “self-completed questionnaires administered to patients to assess their self-reported health status before

and after certain elective healthcare interventions funded by the NHS" (Department of Health, 2008, pp. 5). The consideration of these measures is based on the idea that the best source of information of how a patient feels is the patient himself (Devlin et al., 2010).

Different authors have also defined PROMS as measurements of any aspect of a patients' health status, obtained directly from the patients, i.e. without the help of physicians or other observers (Ousey and Cook, 2001; Valderas et al., 2008; Valderas and Alonso, 2008; Wylde et al., 2009). As such, PROMs consider the patient's view and satisfaction, increasing with that, their participation in health care (Marshall et al., 2006)

Across PROMs we can distinguish between generic and disease-specific measures. The generic measures of health-related quality of life are being collected for all procedures in the PROMs survey. They allow for comparisons of hospitals for individual procedures as well as across interventions. Despite this positive aspect, the generic measures present some disadvantages. For example, the items they include are broader and not directly related to the condition. Therefore, the patient's utility score may include health aspects not related to the surgery for which the questionnaire has been completed.

The disease-specific measures are particular for each procedure. They are hypothesised to be more sensitive to changes in health status within a given procedure, as they only consider information for the particular disease they analyse. Therefore, they can help to examine that the generic measures do not miss a relevant aspect of patient health related to the medical condition for which patient has received treatment (Devlin and Appleby, 2010). However, this implies that we can only use them to make comparisons within a particular procedure and not across patients presenting with different conditions (Devlin and Appleby, 2010).

2.2 Instruments included in the current NHS PROM initiative

Currently in the NHS both, generic as well as disease-specific measures are being collected. The generic measures are EQ-5D and EQ-VAS, while the specific measures for each of the interventions are: OHS, OKS and AVVQ respectively for hip and knee replacement and varicose vein surgery, not having any specific measure for the case of hernia repair. Here we present the main characteristics for each of them.

2.2.1 EQ-5D descriptive system

The EQ-5D descriptive system measures patients' self-reported health-related quality of life in terms of five health domains (mobility, self-care, usual activity, pain/discomfort and anxiety/depression). For each of these domains, patients indicate the degree of problems they experience using a three-point scale (no problems (=1), some problems (=2), extreme problems (=3)). As a result, the EQ-5D can describe 243 different health states, where the health state defined as 11111 (which would correspond to a person with no problems in any of the dimensions) reflects full health, and 33333 is the worst possible health state. This health profile can be translated into a weighted utility score using the UK population weights (Dolan, 1997). The resulting EQ-5D index is a cardinal measure and ranges from one, representing perfect health, to -0.594, where zero represents a state equivalent to being dead and utility scores lower than zero represent health states worse than being dead.

Using EQ-5D to measure patient health and health-related quality of life has many advantages. For example, it is simple in use and has been found to be responsive to change and reliable (Hurst et al., 1997). However, the disadvantages are that it can lead to losses of information when obtaining the EQ-5D index from the EQ-5D profile (Devlin and Appleby, 2010; Gutacker et al., 2012). For example, the differences between scores can be related to a particular dimension and it may be interesting knowing in which dimension the differences in health arise.

2.2.2 EQ-VAS

The EQ-5D also contains a visual analogue scale (EQ-VAS). The EQ-VAS can be defined as a measure of the patient's valuation of their own global health status. This scale ranges from zero to one hundred where zero is the worst health state and 100 is the best state that patients can report. Patients are asked to report their health-related quality of life by indicating the point on the scale that reflects their current health state. The EQ-VAS is not based on a utility theory and, therefore, it is not much used by economists.

Following the terminology used by the English Department of Health (DH), we will refer to the EQ-5D descriptive system simply as the EQ-5D and will treat the EQ-5D descriptive system and the EQ-VAS as two independent measures (NHS, 2011). Note that

this is at odds with the terminology used by the EuroQol group who developed and maintain the EQ-5D.

2.2.3 OHS and OKS

The OHS and OKS instruments are designed to evaluate disability in patients undergoing total hip and knee replacement respectively (Dawson et al., 1998; Wylde et al., 2009). Each questionnaire contains a total of 12 questions about pain and physical limitations which have existed during the past four weeks due to the hip or knee. Ten of the questions are identical in the OHS and OKS, while the remaining two are specific to the condition. Each of the questions has five categories of response, from least to most difficulty or severity, resulting in more than 244 million possible health states (Oppe et al., 2011). Each of the answers results in an item score between zero and four and the total score is obtained by adding the individual item scores. The total score ranges from zero to 48 were lower scores indicate higher disability. Both measures have been found to be “practical, reliable, valid and sensitive to clinically important changes over the time” (Dawson et al., 1998, pp.63)

2.2.4 AVVQ

This specific questionnaire for the analysis of varicose vein surgery consists of 13 questions. The responses can be aggregated into an index taking values from zero to 100 (Garratt et al., 1993) using weights provided by the developers. A score of zero is defined to be the best health state and higher scores reflect worse health states.

3 LITERATURE REVIEW

As a first step in assessing the performance of generic versus disease-specific PROMs we conducted a literature review on the comparative performance of PRO measures for the purpose of performance assessment.

MEDLINE and Pre-MEDLINE, Embase, Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts of Review of Effects (DARE), Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Methodology Register (CMR), Health Technology Assessments (HTA) and NHS Economic Evaluation Database (NHS EED) databases were searched. Specifically we searched for papers that use generic terms for PROMs, quality of

life questionnaires, and quality-adjusted life year measurement tools, combined with search terms for specific instruments. These searches were not limited by date range or restricted to English publication, although, we searched on English terms only. Duplicate records were eliminated after careful consideration.

The literature search resulted in 804 unique papers. In order to reduce this to a manageable quantity, we applied the following criteria (a summary of the process can be seen in Figure 1): One researcher (AGC) reviewed all titles and selected 195 for further investigation. These studies considered either one of the four interventions (hip or knee replacement, varicose vein surgery or hernia repair) or focused on one of the measures of interest (EQ-5D, EQ-VAS, OHS, OKS, or AVVQ). After this reduction, three researchers (CB, AGC and NG) proceed to read the abstracts in order to obtain a set of relevant papers. This revision obtained a total of 102 papers. From these, 25 considered the generic measures (EQ-5D descriptive system and EQ-VAS), 31 consider the OHS, 27 the OKS and 19 use the AVVQ.

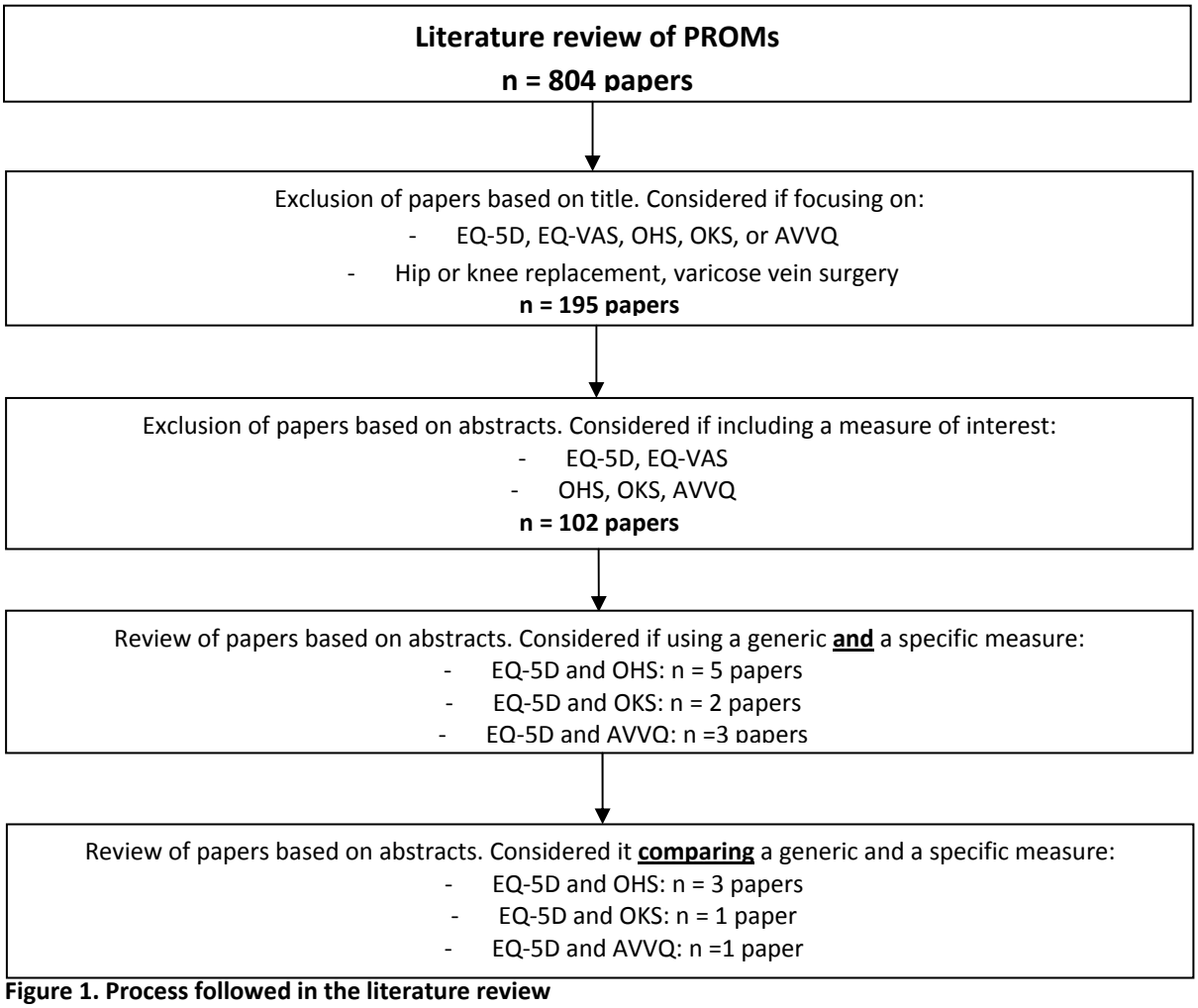

Figure 1. Process followed in the literature review

The vast majority of these papers include only one of the measures of interest and thus does not allow for assessment of comparative performance of these instruments. Only studies that contain information on both a disease-specific and a generic measure were considered further.

We found five papers analysing EQ-5D and OHS (Bilberg et al., 2011; Dawson et al., 2001; Ostendorf et al., 2004; Oppe et al., 2009, and Oppe et al., 2011); two considering EQ-5D and OKS (Xie et al., 2007, and Baker et al., 2012), and three focusing on EQ-5D and AVVQ (Nesbitt et al., 2011; Nesbitt et al., 2012 and Smith et al., 2002). Two papers focus on more than two measures at the same time (EQ-5D, OHS and OKS) (Chard et al., 2001 and Browne et al., 2008) and one paper contain information on all measures, generic and disease-specific, that are included in the PROM survey (Soljak et al., 2009).

We did not consider some of those papers for several reasons: although some of them carried out comparisons those comparisons were made across patient population rather than between instruments (Bilberg et al., 2011) or referred to cross-cultural adaptation comparisons (Xie et al., 2007); others referred to conference abstracts without a relevant publication (Oppe et al., 2009; Nesbitt et al., 2011). Finally, one paper was excluded because it analyses the value of colour duplex in pre-operative marking for varicose vein surgery and was hence judged unrelated to our research question (Smith et al., 2002). Regarding the papers using more than two measures, some of them compare the NHS with the Independent Sector Treatment Centres (ISTCs) (Browne et al., 2008 and Chard et al., 2001) and other ones focus on the analysis of socioeconomic differences in health (Soljak et al., 2009).

After applying these various exclusion criteria, we ended up with five relevant papers which specifically compared instruments: Oppe et al. (2011), Dawson et al. (2001), Ostendorf et al. (2004), Baker et al. (2012) and Nesbitt et al. (2012)., from which we proceeded to read the full texts. These papers are relevant for us given that, in general, they analyse whether the answers to one of the questionnaires are in line with the answers given to the other one

used in the comparison. Particularly for each paper, we present a summary of the work developed in each of them[1].

Oppe et al. (2011) evaluate the comparability of the information reported by the OHS and EQ-5D, and investigate whether the OHS can be mapped onto the EQ-5D responses. They use data on English patients undergoing unilateral hip replacement and apply Principal Component Analysis (PCA) to assess whether certain items carry information about a more general construct. Items with high intercorrelation were supposed to reflect the same construct. In addition, they also performed a mapping exercise in order to find a single model with which to link the OHS responses to EQ-5D utility scores. In this analysis they use the EQ-5D as dependent variable and the items from the OHS are used as the explanatory variables. They obtain a moderate correlation between OHS and EQ-5D with respect to the pre-operative data (0.33), while it is higher when analysing the data after operation (0.51). Furthermore, they also find that the lowest correlation is associated with the anxiety/depression dimension of EQ-5D. Via mapping they find that this anxiety/depression construct is not related to any of the OHS items and that those mapping models do not estimate the utilities of the health states correctly, underestimating the utilities of the mild states and overestimating those for the severe state. Therefore, they assert that there are conceptual differences between the instruments which prohibit a satisfactory mapping.

Dawson et al. (2001) compare the EQ-5D and OHS data to ascertain the validity of OHS with respect to changes in health-related quality of life as measured by the EQ-5D and to analyse the sensitivity of the OHS instrument. Patients are asked to complete both questionnaires before surgery and one year after surgery. These responses are then translated into change scores for each instrument, defined as the difference between the postoperative and the preoperative scores, and effect sizes, which measures the change in a standardised way and permits making direct comparisons between instruments. They found that there is a correlation between OHS and EQ-5D in the preoperative scores (0.67) as well as in the postoperative ones (0.77) and interpret this as a high level of agreement between measures. The correlation is also high (0.59) when they consider the change scores. Regarding the sensitivity of the instruments, they find that when assessing change, the EQ-

[1] For Nesbitt et al. (2012) we do not present any summary, given that the full text for this paper could not be obtained from the library in York directly of the British Library in London.

5D is less sensitive than the OHS if there have been any revisions of hip replacement in the past, and conclude that the OHS is more sensitive to improvements in hip-related health.

Ostendorf et al. (2004) perform a comparison of the characteristics of OHS and EQ-5D after Total Hip Replacement (THR). Apart from those instruments they also consider the disease-specific WOMAC and the generics SF-36 and SF-12. They compare patients' baseline scores as measured by these instruments and analyse the sensitivity to change of the different questionnaires. In order to make this comparison, they collect data before surgery and three and twelve months afterwards. The comparison between the instruments is made using the Spearman correlation coefficients, which allows for the analysis of whether a change in one of the measures is related with a change in another measure. They find that all the scores, except the SF-36, improve at one year after operation. Furthermore, they find high and statistically significant correlations and correlations in the change between the pre-operative and post-operative scores between the measures, especially between OHS and EQ-5D among others (0.51). Apart from this, they also find that many outcome scores showed a very important ceiling effect at one year after surgery. When comparisons are made between instruments, for example between the disease-specific measures WOMAC and OHS, they recommend the use of OHS given that it is "shorter, more site-specific and responsive" and it does not show high ceiling effect after surgery. Regarding the generic measures they recommend the use of SF-12 and they point out that EQ-5D would be especially useful in situations where the utility values are needed.

The literature review identified only one relevant paper on knee replacement. This paper refers to the work developed by Baker et al. (2012). They compare both the disease-specific OKS and the generic EQ-5D in unicondylar and total knee replacement (UKR and TKR). They determine which variables explain the largest proportion of the variance in the different outcomes by means of linear regression and logistic regression. They show that without adjusting for patient characteristics all instruments detect improvements. After obtaining risk-adjustment, they find that the most important variable in the models is the preoperative score. This implies that patients that were worse at the baseline show the highest improvement and those who were better present a ceiling effect showing the inability to improve to the same extent. However, there are not statistically significant differences in the improvements measured by the different instruments.

Apart from these studies, the literature review resulted in several papers that compare other unrelated generic and disease-specific instruments. An overview of those papers can be found in Table 1. The most commonly featured generic measure is the SF-36, and the most commonly featured disease-specific instrument is the WOMAC.

Paper	Generic measure	Disease-specific measure
Bak et al., 2001	SF-36	WOMAC
Bombardier et al., 1995	SF-36	WOMAC
Ghanem et al., 2010	SF-36	WOMAC
Hawker et al., 1995	SF-36	WOMAC
Impellizzeri et al., 2011	SF-12	OKS, WOMAC
Kirschner et al., 2003	SF-36	SMFA-D, WOMAC
Larsen et al., 2010	EQ-5D, SF-36	HHS
Lingard et al., 2001	SF-36	WOMAC
Robertsson & Dunbar, 2001	SF-36, SF-12, NHP	OKS, WOMAC
Shepherd et al., 2011	SF-12	AVVQ, SQOR-V

Table 1. Papers comparing generic and disease-specific measures

From the analysis of the literature we have reviewed, we can draw two main conclusions. The first of them is that most of the papers report a high correlation between the instruments (around 0.50), although some papers report better correlations than other papers. These high correlations indicate that, in general, the answers given to the different measures (generic and disease-specific) will lead to similar results. However, it cannot be generalised, given that sometimes some measures have not reflected changes that other measures have detected. The second one is that we have not found any papers comparing instruments for the analysis of provider assessment, which is the focus of this study.

4 METHODS

In this section we present the main characteristics of the sample used in the study as well as the methodology used in the estimation.

4.1 Data

We extract data from the patient-reported outcome survey database for all NHS-funded patients undergoing elective hip replacement in the period April 2009 to March 2010. Patients undergoing this intervention are asked to complete the pre-operative PROMs questionnaire during the initial outpatient appointment preceding the admission. Post-operative questionnaires are sent to patients 6 months after discharge.

We link these data to routinely collected inpatient records as recorded in the Hospital Episode Statistics (HES) database. This allows us to obtain information on a large range of patient characteristics as well as to identify whether the patient underwent primary or revision surgery.

The initial data consisted of 63,761 patients. After cleaning the data we ended up with 20,509 patients, which are clustered in 153 different hospitals. This cleaning process eliminated observations because of missing participation in the PROM survey, because the patient had died before completing the second questionnaire or because the patient underwent emergency surgery and was thus not eligible to fill in a PROM questionnaire. Although this reduction seems large, it is the mechanism behind this reduction that is important. For example, if patients that do not experiencing a positive change in the HRQoL after surgery are less likely to fill in the post-operative questionnaire, the estimates of hospital effects may be biased. For the purposes of the study we will assume that the observations we observe are representative of the observations we do not observe. Therefore we assume that the magnitude and nature of the missing data does not cause bias in our estimates.

4.2 Empirical approach

Here, we present the characteristics of the multilevel models used in the analysis. Afterwards, we explain the procedure we have followed to identify unusual hospitals ("outliers").

4.2.1 Multilevel model

Patients are clustered in hospitals, leading to a hierarchical data structure. Multilevel models are commonly used when the data that are to be analysed fall into a "hierarchical structure consisting of multiple macro units and multiple micro units within each macro units" (Rice and Jones, 1997, pp. 562). Our data present this multilevel structure, as we have many hospitals and many individuals within each hospital.

We estimate multilevel models with individual patient characteristics and hospital fixed effects, which will allow us to investigate whether some hospitals have a differential performance than others based on the outcome data provided by patients within different

hospitals. Furthermore, this methodology allows us to account for the observed heterogeneity in patient characteristics rather than rely on aggregate data about patient severity at hospital level (Laudicella et al., 2010).

The dependent variable used in these models is the change in patient health or health-related quality of life as measured by the different PROM instruments. We are interested in the unexplained variation at hospital captured by the hospital fixed effect. Because we control for patient characteristics, any remaining variation at provider level (the "hospital effect") can be interpreted as systematic variation in performance.

Street et al. (2012) use fixed effect multilevel models to explain why the resource use (costs or length of stay) differs among patients and hospitals. In order to do it, they make use of a two-stage model. In the first stage they analyse the influence of a set of patient level explanatory variables on individual resource use and extract hospital fixed effects. In the second stage, they analyse these fixed effects to identify hospital level factors that are associated with performance variation.

Laudicella et al. (2010) use the same approach to examine to what extent costs of obstetrics departments are explained by the characteristics of patients admitted to their respective diagnosis related groups (DRGs). After controlling for those characteristics, they analyse why some departments have higher costs than other. Again, estimates of departmental fixed effects are expected to reflect the relative performance of each department, with values above the national average indicating worse performance (higher average costs).

We follow these two examples in specifying our empirical approach. We estimate three different models: one for OHS, one for EQ-5D and one for EQ-VAS. The model is specified as follows:

$$change_{ij} = \beta x_{ij} + u_j + e_{ij} \quad [1]$$

where the subscript i refers to the individual patient while the subscript j is used to identify the hospitals. The vector x_{ij} comprises the set of explanatory variables containing patient characteristics. We consider the age and sex of the patients, whether the operation

was a revision of a previous hip replacement, as well as the Charlson index of comorbidity[2] (Charlson et al., 1987). In order to allow for a non-linear relationship between age and change in health status, we also include a squared term of age. Finally, we include the initial health status reported in the pre-operative questionnaire. The rationale for this is that not all patients are likely to improve to the same extent. Potentially, those patients who are in worse health before surgery are more likely to experience a greater change as surgery can have higher impacts on those individuals. Similarly, those patients who report to be healthier at the baseline cannot improve to the same extent given what has been termed a "ceiling effect" and the "inability of the scores to detect top-end differences" (Baker et al., 2012, pp. 924).

The term u_j is the fixed effect for the jth hospital. This is the term we are interested in, as it reflects the performance of each individual hospital j. In our case, higher values of this term will indicate that that particular hospital is performing better than another one with a smaller value of u_j, given that patients express a higher scores in the different questionnaires.

Finally, the term e_{ij} denotes the random error. This error term is assumed to fulfil the standard properties of the disturbances (Woolridge, 2009): random disturbances, homoskedasticity, no serial correlation and normal distribution.

4.2.2 Identification of outliers

After the estimation of the different hospital effects we identify unusual hospitals, which are often termed "outliers". These outliers are hospitals that perform differently from the average with respect to at least one of the PROMs considered. Note that some hospitals may only be considered outlier on one PROM, whereas others may be outliers on both PROMs which are compared.

The procedure we have followed to identify these outliers is as follows: we obtain an estimate of the hospital effect together with its 95% confidence intervals. We defined a variable including those observations for which the confidence intervals contains the value zero and that, therefore, are not statistically different from the average. For those hospitals

[2] The comorbidities we are considering are shown in the Appendix 1.

for which the confidence intervals do not include zero, we classify them as positive outlier if the confidence intervals only contain positive values and as negative outliers if the confidence intervals only contain negative values. Hence, positive outliers are those hospitals that perform above the average while negative outliers are those below the average.

5 RESULTS

In this section we present the results obtained from the application of the previously described methodology. We begin by presenting the descriptive statistics of our sample and describing correlation pattern of the unadjusted scores. We then present the findings obtained from the application of regression models. As mentioned before the data analysed refers to patients undergoing hip replacement. Therefore, the questionnaires we use in the analysis are the generics EQ-5D and EQ-VAS and the disease-specific OHS.

5.1 Descriptive statistics

Our final data consisted of 20,509 patients. Out of these 20,509 patients the 41.45% (n = 8,501) were males while the 58.55% (n = 12,008) were females. The age of patients ranged from 15 to 94, with an average age of 68 years. 7.40% (n = 1,516) of patients underwent a revision of a previous hip replacement.

We obtain summary statistics for the PROM responses both, at the individual level and the hospital level. The hospital level data is obtained by computing the mean PROM score for each of the hospitals. Once computed, we drop all the observations except from one, which is going to be representative of that particular hospital. Accordingly, hospital level data are not weighted by hospital volume.

Patient level	**OHS**		**EQ-5D**		**EQ-VAS**	
	Pre-op	Post-op	Pre-op	Post-op	Pre-op	Post-op
Mean	18.38	38.16	0.357	0.765	66.36	75.52
Min	0	0	-0.594	-0.594	0	0
Max	48	48	1	1	100	100

Table 2. Descriptive statistics. Patient level

Hospital level	**OHS**		**EQ-5D**		**EQ-VAS**	
	Pre-op	Post-op	Pre-op	Post-op	Pre-op	Post-op
Mean	18.02	37.67	0.346	0.753	65.54	74.57
Min	13.67	29.3	0.185	0.604	55.65	60
Max	29	44.5	0.639	0.908	72.78	81.5

Table 3. Descriptive statistics. Hospital level

Tables 2 and 3 present the mean, minimum and maximum of the unadjusted scores at patient and hospital level. The mean pre-operative score is 18.38 for OHS, 0.357 for EQ-5D and 66.36 for EQ-VAS. The scores at hospital level are very similar to those at patient level. For the post-operative score all the mean values experience an improvement. Specifically these values are 38.16, 0.765 and 75.52 respectively for OHS, EQ-5D and EQ-VAS. This improvement is particularly big in the case of OHS and EQ-5D both at the individual level (207.62% increase and 214.29% increase respectively) and at the provider level (209.06% and 217.63% respectively). Improvements in patient health as measured by EQ-VAS are substantially smaller (13.80% for individual-data and 13.78% for hospital-data). Furthermore, an interesting issue in the analysis of patient level data is related to the number of patients reporting perfect health before and after surgery in each of the measures. In the case of OHS the percentage of people reporting perfect health increases from 0.05% before surgery to 11.67% after surgery. For EQ-5D this difference goes from 3.12% in the pre-operative score to 35.62% with respect to post-operative score. In the case of EQ-VAS there is also an increase in this percentage, although this is smaller. Specifically it is 1.77% of patient reporting perfect health before surgery and 4.55% reporting perfect health after surgery.

The differences between PROMs at patient level are also reflected in the provider level statistics. For example, for the case of OHS, the minimum value before surgery is 13.67 while after surgery the minimum value is 29.3. The same is observed at the top. Before surgery the highest score, in for example OHS, is 29, while after surgery this score increased to 44.5.

Average scores at hospital level are more dispersed before surgery than after surgery. For example, the difference between the minimum and maximum values on the EQ-5D is 0.454 points, while after surgery this reduces to 0.304. This reduction in variance is also observed for the OHS, albeit less pronounced. However, in the case of EQ-VAS, the difference between the minimum and maximum value after surgery is bigger than before surgery. This can be also observed in the histograms for the pre-operative and post-operative scores presented in Figures 2 to 7. In these figures we can also observe the general improvement experienced by the mean values in the different questionnaires.

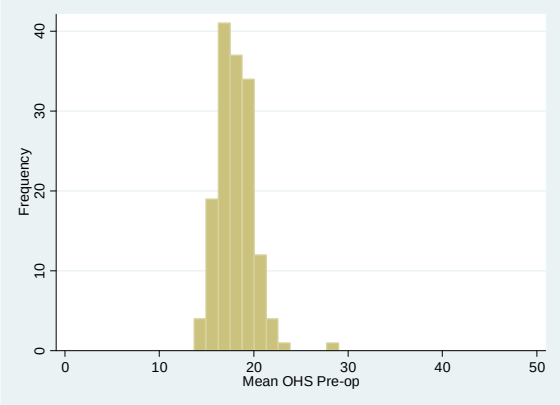

Figure 2. Histogram for pre-op score. OHS

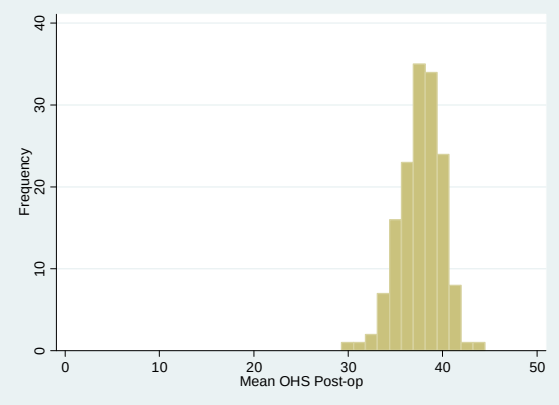

Figure 3. Histogram for post-op score. OHS

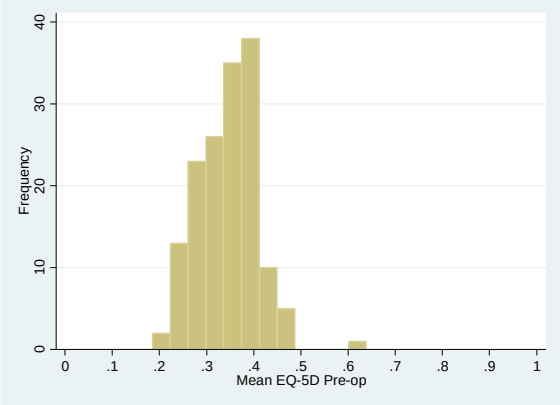

Figure 4. Histogram for pre-op score. EQ-5D

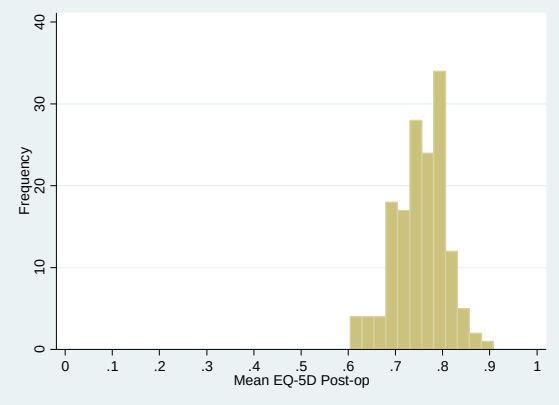

Figure 5. Histogram for post-op score. EQ-5D

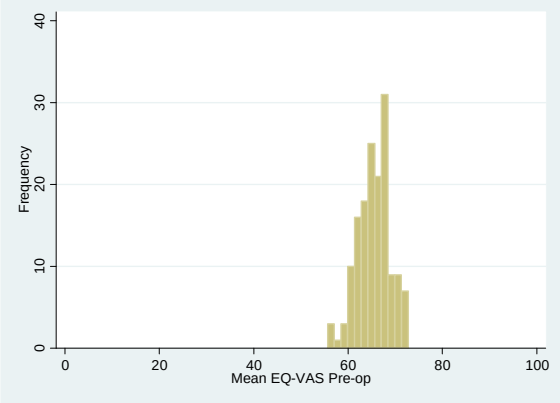

Figure 6. Histogram for pre-op score. EQ-VAS

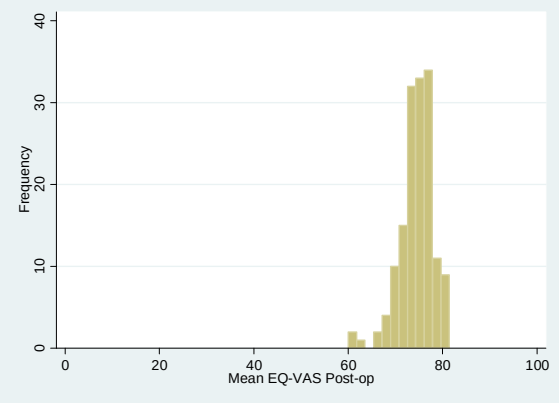

Figure 7. Histogram for post-op score. EQ-VAS

We also computed the correlations between the different PRO measures, again for both, individual and hospital level data. This information is presented in Tables 4 and 5. The correlations showed in the tables refer to the answers given to the pre-operative questionnaires, to the post-operative questionnaires as well as to the change experienced by the two scores.

Patient level	Pre-op	Post-op	CHANGE
OHS and EQ-5D	0.7325	0.7617	0.6311
OHS and EQ-VAS	0.3783	0.5982	0.3535
EQ-5D and EQ-VAS	0.3586	0.6386	0.3234

Table 4. Correlation between PROMs and derived change at patient level

Hospital level	Pre-op	Post-op	CHANGE
OHS and EQ-5D	0.8904	0.9337	0.7571
OHS and EQ-VAS	0.5496	0.7893	0.5106
EQ-5D and EQ-VAS	0.5500	0.8282	0.5145

Table 5. Correlation between PROMs and derived change at hospital level

The highest correlation is observed between the OHS and the EQ-5D, at individual and hospital level. This is consistent with the findings from the previous literature. The EQ-VAS correlates less well with the two other instruments. All correlations are higher at the hospital level than the individual level.

We also derived a series of indicators that record the percentage of patients who report improvements or deteriorations in their health status after surgery or who did not experienced any changes (neutral). These percentages are reported in Tables 6 and 7.

Patient level	Better	Neutral	Worse
OHS	96.07%	0.53%	3.40%
EQ-5D	87.45%	6.16%	6.39%
EQ-VAS	61.49%	11.49%	27.02%

Table 6. Indicators better/neutral/worse. Patient level

Hospital level	Better	Neutral	Worse
OHS	98.04%	0%	1.96%
EQ-5D	87.58%	7.84%	4.58%
EQ-VAS	60.79%	15.03%	24.18%

Table 7. Indicators better/neutral/worse. Hospital level

More patients report improvements when outcomes are measured by the OHS instead of the generic measure. This may be because the OHS is measuring more specific aspects of health of the particular intervention than the EQ-5D or EQ-VAS do. Therefore, the former measure may take into account smaller changes in patient's health status, which are related to the surgery performed. Comparing the two generic measures, we observe that there are more people showing an improvement when considering the EQ-5D rather than EQ-VAS. Additionally, an interesting value is the percentage of people being worse-off when we consider EQ-VAS. This value is over 20% in both cases, at patient and at hospital level data.

5.2 Econometric analysis

We now move to the results of the econometric analysis. We first report the output obtained after the estimation, analysing the individual significance of the variables as well as the proportion of the change in the measure they explain. Afterwards, we present the results obtained from the comparisons of the hospital effects between the generic and disease-specific measures.

5.2.1 Case-mix adjustment

Here we present a relevant subset of the coefficients used in the estimation (Table 8). Full results are presented in Appendix 2.

	OHS	EQ-5D	EQ-VAS
Age of patients	0.354*** (0.043)	0.010*** (0.001)	0.441*** (0.0861)
Age of patients^2	-0.003*** (0.000)	-0.0008*** (0.000)	-0.004*** (0.001)
Sex	0.761*** (0.124)	0.017*** (0.003)	0.375 (0.240)
Pre-op score	-0.656*** (0.007)	-0.774*** (0.005)	-0.716*** (0.006)
Revision procedure	-6.106*** (0.231)	-0.111*** (0.006)	-4.990*** (0.451)
Adj. R^2 without hospital effect	0.299	0.519	0.445
Adj. R^2 with hospital effect	0.313	0.527	0.451

Table 8. Estimation output

All coefficients are significant, except for the case of some comorbidities included in the Charlson index. The significance of the coefficients of the comorbidities varies when we consider the different questionnaires. Some of them are significantly associated with changes in health in all the estimations (chronic obstructive pulmonary disease, rheumatoid disease, mild liver disease and diabetes), others are significant in some estimations but not in others (congestive heart failure, peptic ulcer disease, diabetes and complications, hemiplegia or paraplegia, cancer and metastic cancer) and a range of comorbidities are insignificant in all the estimations (acute myocardial infarction, peripheral vascular disease, cerebrovascular disease, dementia, renal disease or moderate/severe liver disease).

Both age effects are significant and the coefficients for the square of age have a negative sign, resulting in a u-shape effect. Regarding the sex of patients, the variable is significant in every model, except for the case when we analyse EQ-VAS ($p = 0.118$), where

we do not find any statistical significant differences between men and women. For models that consider OHS and EQ-5D as dependent variables, we find that the coefficient for male gender is positive, indicating that men experience a greater change than women and profit more from surgery.

Considering the pre-operative score obtained in each of the questionnaires, we can see that its sign is negative, indicating that those patients being worse at the beginning are experiencing a higher improvement that those ones being better. This is consistent with the results presented in Baker et al. (2012). However, as pointed out by Baker et al. (2012) the reason can be that those patients with a better pre-operative score are not able to improve to the same extent, given the existence of a ceiling effect.

Regarding the revision variable, the negative sign is indicating that those patients undergoing surgery as a revision procedure experience a smaller change than those ones who are admitted as a primary procedure.

Finally, with respect to the estimation we computed the R^2 for two situations: with and without the hospital effects. This value represents the proportion of variation in the PRO measures explained by the variables we are using in the analysis, with and without the hospital effects. We can see from the table presented above that this value changes, so there is a proportion of the change in the measures explained by the hospital effects. However, these changes in the R^2 are very small.

5.2.2 Comparison of hospital effects for OHS and EQ-5D

Our main interest lies on the analysis of the hospital effects, which are represented by the term u_j. We could have represented the effect for each of the questionnaires separately; however, the results are more informative when we plot the hospital effects obtained from one of the estimations against another one. This procedure will allow us to compare the performance of the different hospitals according to the different measures as well as to identify the different outliers.

The hospital effects for the case of OHS and EQ-5D are represented in Figure 8. We observe a strong positive association between both measures. This positive association indicates that better performance on EQ-5D is associated with better performance in OHS.

This Pearson correlation coefficient is 0.8979. This strong correlation between the hospitals effects analysed with the two measures is consistent with the correlations between the unadjusted measures. Previous literature also found high correlations, however, the value we obtain is different given that it is obtained considering the different hospitals and not all aggregated at the patient level.

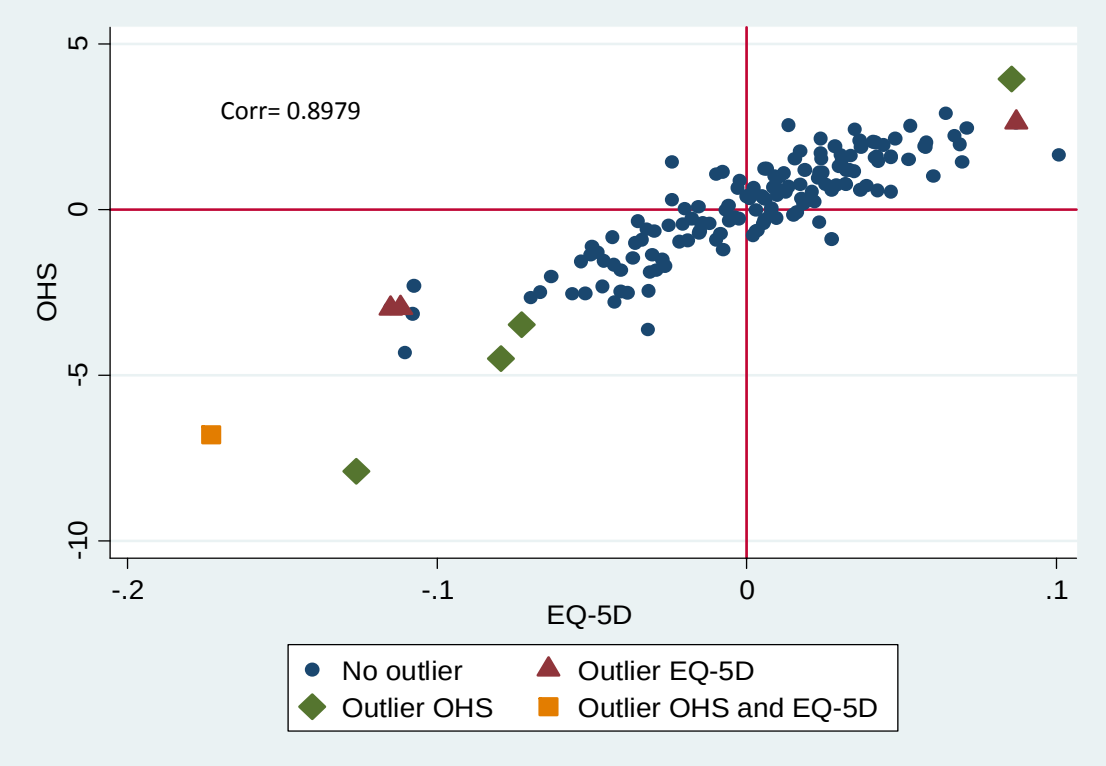

Figure 8. Hospital Effects for OHS and EQ-5D

We find eight hospitals that are outliers (i.e. above/below average performance) on one measure but on the other. These particular hospitals are highlighted in Figure 8 by using different markers. Three of the observations are outliers only with respect to EQ-5D, one of them performing above the average and the other two performing below the average. With respect to OHS we find four outliers, three of them performing below the average and one of them performing above the average.

In addition, there was one hospital which was performing below the average with respect to both EQ-5D and OHS at the same time. When we represented the confidence intervals for this particular hospital we observed that they were very wide compared to the confidence intervals obtained in the other hospitals. This characteristic is observed in both of the measures, EQ-5D and OHS.

This analysis of the outliers according to EQ-5D and OHS can be summarised in the table presented bellow. In this table we can see the number of outliers identified separately for each of the measures as well as the ones identified by both measures.

		OHS			TOTAL
		+ outlier	Average	- outlier	
EQ-5D	+ outlier	0	1	0	**1**
	Average	1	145	3	**149**
	- outlier	0	2	1	**3**
TOTAL		**1**	**148**	**4**	**153**

Table 9. Outliers OHS and EQ-5D

5.2.3 Comparison of hospital effects for OHS and EQ-VAS

The hospital effects for OHS plotted against those for the EQ-VAS are presented in Figure 9. Again, the association between the measures is positive. Specifically the correlation between the two measures is 0.8087, indicating that high values in OHS are associated with high values in EQ-VAS and vice versa. Even if the correlation is weaker than before there is still a strong correlation, and higher than the correlations found at the individual level in the previous literature.

Again, and in spite of this general concordance we have also identified several outliers following the same procedure described before. This procedure obtained a total of eight outliers. As in the previous case we have found one hospital being a negative outlier according to the two measures, and no hospital being classified as positive outlier in the two measures. Furthermore, there are three hospitals which are positive outliers according to EQ-VAS but not according to OHS. One of those hospitals are performing above the average with respect to EQ-VAS (positive outlier) while the other two are performing below the average. For the case of OHS, there is one hospital with a better score than the average and three in which it is below the average.

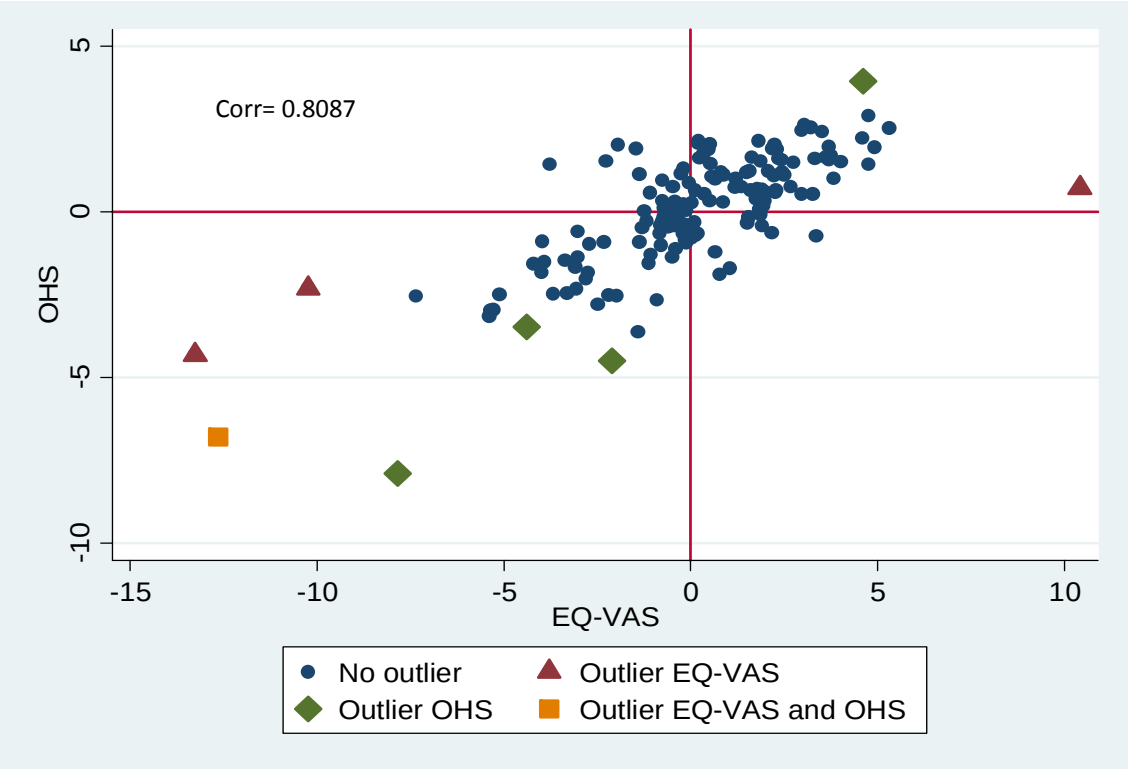

Figure 9. Hospital Effects for OHS and EQ-VAS

In Table 10 we present a summary of the number of hospitals being outliers according to the generic EQ-VAS and the disease-specific OHS separetely and the outliers we have when we consider both of them.

		OHS			TOTAL
		+ outlier	Average	- outlier	
EQ-VAS	+ outlier	0	1	0	**1**
	Average	1	145	3	**149**
	- outlier	0	2	1	**3**
TOTAL		**1**	**148**	**4**	**153**

Table 10. Outliers OHS and EQ-VAS

5.2.4 Comparison of hospital effects for EQ-5D and EQ-VAS

The hospital effects for the two generic measures are presented in Figure 10. The association between the two generic measures is smaller than the association of each of the generics with the disease-specific. Specifically the correlation between the two measures is 0.7578.

Again, we have identified several outliers following the same procedure described before. In this case we have identified a total of seven outliers. As before, we have found one hospital being a negative outlier according to the two measures, and no hospital being

classified as positive outlier in the two measures. Furthermore, we have also identified several hospitals being outlier in only one of the measures. Specifically, there are three outliers with EQ-5D and three outliers with EQ-VAS. In each case, one of the three hospitals are performing above the average with respect to one of the measures (positive outlier) while with respect to the other measure they are performing on the average. The other two hospitals are negative outliers in one of the measures and are performing on the average with respect to the other one. Despite the fact that there are three outliers in each case, one positive and two negative, these outliers do not refer to the same hospitals.

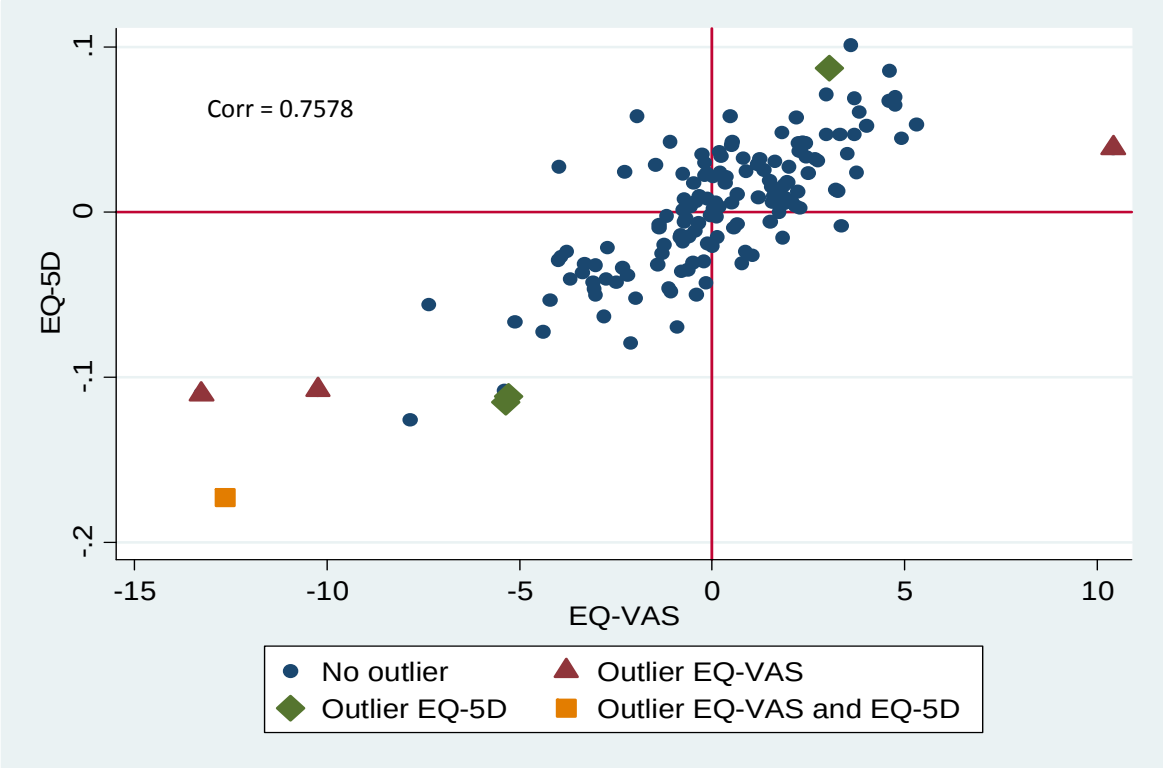

Figure 10. Hospital Effects for EQ-5D and EQ-VAS

This analysis of the outliers on the generic measure is summarised in Table 11. This table shows the number of outliers that each of the measures has identified separately as well as the outliers identified by both measures.

		EQ-VAS			TOTAL
		+ outlier	Average	- outlier	
EQ-5D	+ outlier	0	1	0	**1**
	Average	1	146	2	**149**
	- outlier	0	2	1	**3**
TOTAL		**1**	**149**	**3**	**153**

Table 11. Outliers EQ-5D and EQ-VAS

Furthermore, a particular aspect in the three analyses presented before is the similarities existing between the outliers in the different comparisons. In the first two analyses (OHS vs. EQ-5D and OHS vs. EQ-VAS) we found that there are four hospitals being outliers according to OHS but performing on the average in both cases, when the comparison is made against EQ-5D and when the comparison is made with EQ-VAS. These four hospitals are the same in the two comparisons. Therefore, none of the outliers that the specific measure identifies are considered as such under the generic measures. Contrary to this, we also identified three hospitals performing on the average with respect to OHS and being outliers according to the generic measures. However, these three hospitals are different when we consider EQ-5D than when we use EQ-VAS as the comparative measure for OHS. These three outliers identified in each of the generic measures are the outliers we mentioned before when we explained the third of the analyses comparing EQ-5D and EQ-VAS.

Finally, the hospital which was observed as being a negative outlier with respect to both generic as well as disease-specific is the same when the comparison is made between EQ-VAS and OHS than when it is made between EQ-5D and OHS. In addition, this is hospital is also a negative outliers when the performance comparison is made between the two generic measures. Therefore, this particular hospital is performing below the average according to all the measures we have considered in the analysis.

6 SUMMARY AND DISCUSSION

The inclusion of PROMs in the analysis of hospital quality performance allow for the consideration of patients' perspective, which has been stated to provide value information about their health related quality of life. The general problem is that we have no gold standard measurement to which we can compare alternative measures. Rather we have a disease-specific measure which is a priori hypothesised to be more sensitive to changes in disease condition, but has a naive scoring system that converts patient responses to 48 dimensions/questions to a cardinal number without taking into account the relative value of those dimensions. The generic measures, especially the EQ-5D, has the benefit of a theory based scoring system but may be too blunt an instrument to detect variation in provider performance. Our objective then is to compare the three measures and to observe whether

the messages we would draw from each differ i.e. what is the general correlation between results and would we identify different outliers as a result of different measures?

In order to analyse this issue, we have carried out a review of the literature on PROMs comparison. We searched for papers using the measures currently considered by the NHS for four different procedures: hip replacement, knee replacement, varicose vein surgery and hernia repair. These measures refer to EQ-5D and EQ-VAS as the generic questionnaires and Oxford Hip Score (OHS), Oxford Knee Score (OKS) and Aberdeen Varicose Vein Questionnaire (AVVQ), as the disease-specific measures for hip and knee replacement and varicose vein surgery respectively. We identified only a small number of studies that made direct comparisons between generic and specific instruments. These studies found a strong correlation (of around 50% or more) between the generic and disease-specific measures. This correlation was higher when it is obtained with the post-operative scores rather than with the pre-operative scores. No study compared PROM for the purpose of hospital quality performance assessment.

To fill this gap in the literature, we have carried out an empirical analysis of PROM data that have been collected during April 2009 and March 2010 in the English National Health Service (NHS). This analysis was focused on the hip replacement intervention. The aim of this empirical approach was to analyse the responses on the generic EQ-5D and EQ-VAS and the disease-specific measure OHS and study whether these provide similar inferences with regard to hospital performance. In order to do that we estimated multilevel models with fixed effects for each of the measures using the change experience by every measure between the post-operative and pre-operative scores as dependent variable. As explanatory variables we used a set of variables describing the characteristics of patients. This approach allowed us to analyse the effect of each hospital separately and therefore, to perform comparisons between the measures at the level of hospital instead of considering the data aggregated at the patient level.

The main insights we obtained from the analysis is that, in general older females with a higher pre-operative score are the ones improving at a lower extent. We obtained the R^2 for the model with and without considering the hospital effects. We saw that the model considering the hospital effects explained marginally more the variation in the measures

that the model without the hospital effects. Furthermore, the model explaining more variation was the model related to EQ-5D, rather than the ones considering OHS or EQ-VAS. However, the change experienced by the R^2 was not very big, indicating that the hospital effects are not explaining a big proportion of the change showed by each of the questionnaires.

We made three comparisons, one using EQ-5D and OHS, another one using EQ-VAS and OHS, and the last one comparing the two generic measures, EQ-5D and EQ-VAS. We found that, in general, when we analyse the performance assessment using generic measures and disease-specific instruments the correlation is high. However, for individual hospitals there can be differences in the measure we use.

Looking at what have been defined as outliers, only one hospital was identified as a negative outlier according to all the measures considered. Several other hospitals were being classified as outliers with respect to only one of the measures but not the other. Specifically, we found three hospitals being outliers with respect to OHS, but not to any of the generic EQ-5D or EQ-VAS. Two of them were identified as negative outliers while the other one was identified as being performing above the average (positive outlier). Furthermore, we found three hospitals being outliers with respect to the generic EQ-5D but performing on the average according to OHS. One of those three hospitals was performing above the average and two below the average. Similarly, we found three different hospitals performing on the average with respect to OHS but being outliers when EQ-VAS is the measure of performance. One of these three hospitals was performing above the average and the other two were performing below the average. Additionally, the hospitals identified as outliers according to the generic measures when they were compared with OHS (three outliers in EQ-5D and three in EQ-VAS), were also outliers when the comparison was made between EQ-5D and EQ-VAS. Therefore, we must be cautious when we compare the measures, given that we will not always obtain the same results with generic and disease-specific patient reported outcome measures.

One possible explanation for these differences in performance assessment may be that disease-specific measures are exclusive for a particular disease and hence, may be more sensitive to smaller changes in the patients' hip related health status. Therefore, these

measures will be able to reflect smaller improvements or worsening that the generic measures cannot detect given that it considers wider aspects of the patients' health. In the case that the EQ-5D or EQ-VAS was able to reflect more specific changes maybe the differences between generic and disease-specific measures would have been smaller and we would not have identified any outlier or the number of outliers identified would have been smaller. Another possible answer is that the generic measures can reflect aspects of patients' health status that are not related to the particular surgery we are considering, but that are affecting to patients' general health.

As a solution to those limitations and as an issue to be considered in future research we could use the new version of EQ-5D, the EQ-5D-5L (Herdman et al., 2010). This questionnaire considers the same five dimensions that the previous one (mobility, self-care, usual activity, pain/discomfort and anxiety/depression). However, it includes five levels of answers instead of three which could increase the discriminatory power of the questionnaire. The five levels correspond to no problems, slight problems, moderate problems, severe problems, and extreme problems. Another aspect to be considered in further research would be the use of multivariate multilevel analysis, which would allow for simultaneous observation and analysis of more than one outcome variable (Zellner, 1992).

Furthermore, a more detailed analysis would consist of the consideration of other procedures, such as knee replacement or varicose vein surgery. We considered these interventions in the literature review. However, due to time constraints we were not able to carry out the empirical analysis for all of them. Therefore, a further study including all the procedures would show whether the results we found in the analysis of hip replacement are similar when considering generic and disease-specific instruments for the study of knee replacement as well as varicose vein surgery.

Finally, this study has contributed to the existing literature in the sense that, to the best of our knowledge, none of the previous studies using PROMs carried out comparisons of generic and disease-specific instruments considering the effects of the different hospitals. Our study highlight that relying on the PRO instruments solely may be problematic for performance assessment purposes because different instruments may measure different aspects of health and health-related quality of life.

7 CONCLUSIONS

In this study we have analysed the provision of health services in the NHS according to different patient-reported outcome measures (PROMs). To conclude, we observe that the different measures of health outcome are highly correlated although they identify different outliers. If we had to choose one of them we would recommend the use of EQ-5D given that it is simpler, shorter and can be used across conditions. Furthermore, the inclusion of the specific measures does not seem to provide any important additional information. However, it should be checked whether this conclusion holds when we consider the other two conditions, knee replacement and varicose vein surgery.

ACKNOWLEDGMENTS

Ana Luisa Godoy-Caballero is grateful for the FPI grant obtained by the *Gobierno de Extremadura* (Decreto 146/2010, de 2 de julio, DOE nº 130, de 08/07/10).

REFERENCES

Papers

Bak, P. et al. (2001). Generic and specific health-related quality of life at short-term follow-up after total hip arthroplasty and inpatient rehabilitation program. *Physikalische Medizin Rehabilitationsmedizin Kurortmedizin,* 11(4), 129-132.

Baker, P. N. et al. (2012). Comparison of patient-reported outcome measures following total and unicondylar knee replacement. *Journal of Bone & Joint Surgery - British Volume,* 94(7), 919-927.

Bilberg, R. et al. (2011). Preoperative mental health and quality of life in patients with shoulder and hip pain. *European Journal of Pain Supplements,* 5 (1), 283.

Bombardier, C. et al. (1995). Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. *Medical Care,* 33(4 Suppl), AS131-144.

Browne, J. et al. (2008). Case-mix & patients' reports of outcome in Independent Sector Treatment Centres: Comparison with NHS providers. *BMC Health Services Research,* 8(1), 78.

Chard, J. et al. (2011). Outcomes of elective surgery undertaken in independent sector treatment centres and NHS providers in England: audit of patient outcomes in surgery. *BMJ,* 343, d6404.

Charlson, M. E. et al. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases,* 40(5), 373-383.

Chua, C. L., Palangkaraya, A. and Yong, J. (2010). A two-stage estimation of hospital quality using mortality outcome measures: an application using hospital administrative data. *Health Economics,* 19(12), 1404-1424.

Dawson, J. et al. (2001). Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. *Journal of Bone & Joint Surgery - British Volume,* 83(8), 1125-1129.

Dawson, J. et al. (1998). Questionnaire on the perceptions of patients about total knee replacement. *Journal of Bone & Joint Surgery - British Volume,* 80(1), 63-69.

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

Devlin, N. J. and Appleby, J. (2010). Getting the most out of PROMs. *London: Kings Fund*.

Devlin, N. J., Parkin, D. and Browne, J. (2010). Patient-reported outcome measures in the NHS: new methods for analysing and reporting EQ-5D data. *Health economics,* 19(8), 886-905.

Dimick, J. B. et al. (2012). Composite Measures for Rating Hospital Quality with Major Surgery. *Health Services Research*.

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical care*, 1095-1108.

Garratt, A. M. et al. (1993). Towards measurement of outcome for patients with varicose veins. *Quality in Health Care,* 2(1), 5-10.

Ghanem, E. et al. (2010). Limitations of the Knee Society Score in evaluating outcomes following revision total knee arthroplasty. *Journal of Bone & Joint Surgery - American Volume,* 92(14), 2445-2451.

Gutacker, N., Bojke, C., Daidone, S., Devlin, N. and Street, A. (2012). Analysing hospitals variation in health outcome at the level of EQ-5D dimension, CHE Research Paper 74, Centre for Health Economics, University of York.

Hawker, G. et al. (1995). Comparison of a generic (SF-36) and a disease specific (WOMAC) (Western Ontario and McMaster Universities Osteoarthritis Index) instrument in the measurement of outcomes after knee replacement surgery. *Journal of Rheumatology,* 22(6), 1193-1196.

Herdman, M. et al. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research,* 20(10), 1727-1736.

Hurst, N. et al. (1997). Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). *Rheumatology,* 36(5), 551-559.

Impellizzeri, F. M. et al. (2011). Comparison of the reliability, responsiveness, and construct validity of 4 different questionnaires for evaluating outcomes after total knee arthroplasty. *Journal of Arthroplasty,* 26(6), 861-869.

Kirschner, S. et al. (2003). German short musculoskeletal function assessment questionnaire (SMFA-D): comparison with the SF-36 and WOMAC in a prospective evaluation in patients with primary osteoarthritis undergoing total knee arthroplasty. *Rheumatology International,* 23(1), 15-20.

Larsen, K. et al. (2010). Patient-reported outcome after fast-track hip arthroplasty: a prospective cohort study. *Health & Quality of Life Outcomes,* 8, 144.

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), 1872-1881.

Lingard, E. A. et al. (2001). Validity and responsiveness of the Knee Society Clinical Rating System in comparison with the SF-36 and WOMAC. *Journal of Bone & Joint Surgery - American Volume,* 83-A(12), 1856-1864.

Marshall, S., Haywood, K. and Fitzpatrick, R. (2006). Impact of patient-reported outcome measures on routine practice: a structured review. *Journal of evaluation in clinical practice,* 12(5), 559-568.

National Institute for Health and Clinical Excellence, (2008). Guide to the methods of technological appraisal. ISBN: 1-84629-741-9

Nesbitt, C. et al. (2012). Interpretation of patient-reported outcome measures for varicose vein surgery. *Phlebology,* 27(4), 173-178.

Nesbitt, C., Wilson, W. R. W. and Stansby, G. (2011). A critical interpretation of UK patient-reported outcome measures for varicose veins surgery. *Phlebology,* 26 (6), 258.

Oppe, M. and Devlin, N. (2009). To map or not to map? The Oxford hip score and EQ-5D compared. *Value in Health,* 12 (7), A397.

Oppe, M., Devlin, N. and Black, N. (2011). Comparison of the underlying constructs of the EQ-5D and Oxford Hip Score: implications for mapping. *Value in Health,* 14(6), 884-891.

Ostendorf, M. et al. (2004). Patient-reported outcome in total hip replacement. A comparison of five instruments of health status. *Journal of Bone & Joint Surgery - British Volume,* 86(6), 801-808.

Ousey, K. and Cook, L. (2011). Understanding patient reported outcome measures (PROMs). *British Journal of Community Nursing,* 16(2), 80-82.

Rice, N. and Jones, A. (1997). Multilevel models and health economics. *Health Economics,* 6(6), 561-575.

Robertsson, O. and Dunbar, M. J. (2001). Patient satisfaction compared with general health and disease-specific questionnaires in knee arthroplasty patients. *Journal of Arthroplasty,* 16(4), 476-482.

Selim, A. J. et al. (2002). Risk-adjusted mortality rates as a potential outcome indicator for outpatient quality assessments. *Medical care,* 40(3), 237-245.

Shepherd, A. C. et al. (2011). A study to compare disease-specific quality of life with clinical anatomical and hemodynamic assessments in patients with varicose veins. *Journal of Vascular Surgery,* 53(2), 374-382.

Smith, J. J. et al. (2002). Randomised trial of pre-operative colour duplex marking in primary varicose vein surgery: outcome is not improved. *European Journal of Vascular & Endovascular Surgery,* 23(4), 336-343.

Soljak, M. et al. (2009). Is there an association between deprivation and pre-operative disease severity? A cross-sectional study of patient-reported health status. *International Journal for Quality in Health Care,* 21(5), 311-315.

Street, A. et al. (2012). HOW WELL DO DIAGNOSIS-RELATED GROUPS EXPLAIN VARIATIONS IN COSTS OR LENGTH OF STAY AMONG PATIENTS AND ACROSS HOSPITALS? METHODS FOR ANALYSING ROUTINE PATIENT DATA. *Health Economics,* 21(S2), 6-18.

Thomas, N., Longford, N. T. and Rolph, J. E. (1994). Empirical Bayes methods for estimating hospital-specific mortality rates. *Statistics in medicine,* 13(9), 889-903.

Valderas, J. et al. (2008). The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Quality of Life Research,* 17(2), 179-193.

Valderas, J. M. and Alonso, J. (2008). Patient reported outcome measures: a model-based classification system for research and clinical practice. *Quality of Life Research,* 17(9), 1125-1135.

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach*. South-Western Pub.

Wylde, V. et al. (2009). Patient-reported outcomes after total hip and knee arthroplasty: comparison of midterm results. *Journal of Arthroplasty,* 24(2), 210-216.

Xie, F. et al. (2007). Cross-cultural adaptation and validation of Singapore English and Chinese Versions of the Oxford Knee Score (OKS) in knee osteoarthritis patients undergoing total knee replacement. *Osteoarthritis & Cartilage,* 15(9), 1019-1024.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association,* 57(298), 348-368.

Web pages

NHS (2011) The Information Centre. http://www.ic.nhs.uk/statistics-and-data-collections/hospital-care/patient-reported-outcome-measures-proms/finalised-patient-reported-outcome-measures-proms-in-england--april-2009-to-march-2010-pre-and-post-operative-data-experimental-statistics.

EuroQoL. http://www.euroqol.org/.

Appendix 1. Charlson index of comorbidity

VARIABLE	DESCRIPTION
AMI	Acute Myocardial Infarction
CHF	Congestive Heart Failure
PVD	Peripheral Vascular Disease
CD	Cerebrovascular Disease
Dem	Dementia
COPD	Chronic Obstructive Pulmonary Disease
RD	Rheumatoid Disease
PED	Peptic Ulcer Disease
MLD	Mild Liver Disease
Dia	Diabetes
Dia + Com	Diabetes + Complications
H/P	Hemiplegia or Paraplegia
RD	Renal Disease
Cancer	Cancer
M/SLD	Moderate/Severe Liver Disease
MC	Metastic Cancer
AIDS	AIDS

Appendix 2. Output estimations

	OHS	EQ-5D	EQ-VAS
Age of patients	0.354*** (0.043)	0.010*** (0.001)	0.441*** (0.0861)
Age of patients^2	-0.003*** (0.000)	-0.0008*** (0.000)	-0.004*** (0.001)
Sex	0.761*** (0.124)	0.017*** (0.003)	0.375 (0.240)
Q1 score	-0.656*** (0.007)	-0.774*** (0.005)	-0.716*** (0.006)
Revision procedure	-6.106*** (0.231)	-0.111*** (0.006)	-4.990*** (0.451)
AMI	-0.805 (0.651)	-0.122 (0.178)	-1.556 (1.271)
CHF	-1.227 (0.718)	-0.072*** (0.020)	-4.596** (1.403)
PVD	-0.188 (0.657)	-0.025 (0.018)	-0.805 (1.283)
CD	-.0667 (0.925)	-0.029 (0.025)	0.043 (1.807)
Dem	0.880 (1.558)	0.005 (0.043)	-2.877 (3.041)
COPD	-1.480*** (0.201)	-0.046*** (0.006)	5.022*** (0.392)
RD	-0.995** (0.344)	-0.080*** (0.009)	-6.094*** (0.671)
PED	-3.050* (1.531)	-0.064 (0.042)	-4.550 (2.989)
MLD	-3.460* (1.585)	-0.138** (0.044)	-10.014** (3.093)
Dia	-1.939*** (0.224)	-0.049*** (0.006)	-4.417*** (0.438)
Dia+Com	-3.241* (1.483)	-0.071 (0.041)	-6.061* (2.896)
H/P	-3.117 (1.842)	-0.036 (0.051)	-9.699** (3.597)
RD	0.397 (0.426)	-0.001 (0.012)	-1.524 (0.831)
Cancer	-0.778 (0.645)	-0.027 (0.018)	-3.829** (1.260)
M/SLD	-1.848 (3.803)	-0.135 (0.105)	-6.166 (7.423)
MC	-2.042 (1.774)	-0.064 (0.049)	-10.442** (3.463)
Adj. R^2 without hospital effect	0.299	0.519	0.445
Adj. R^2 with hospital effect	0.313	0.527	0.451
*** p <0.05; **p < 0.01; p < 0.001**			