
Article:
Gutacker, Nils orcid.org/0000-0002-2833-0621, Patton, Thomas Edward, Shah, Koonal et al. (1 more author) (Accepted: 2020) Using EQ-5D Data to Measure Hospital Performance:Are General Population Values Distorting Patients' Choices? Medical Decision Making. ISSN 1552-681X (In Press)

Medical Decision Making

Using EQ-5D Data to Measure Hospital Performance: Are General Population Values Distorting Patients' Choices?

Journal:	*Medical Decision Making*
Manuscript ID	MDM-19-321.R2
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	n/a
Complete List of Authors:	Gutacker, Nils; University of York Halifax College, Centre for Health Economics; University of York Centre for Health Economics, Patton, Thomas (Tom); University of York Halifax College, Team for Economic Evaluation and Health Technology Assessment (TEEHTA), Centre for Health Economics; University of York Centre for Health Economics, Shah, Koonal; University of Sheffield School of Health and Related Research, Health Economics and Decision Science (HEDS); Office of Health Economics, Parkin, David; King's College London Faculty of Life Sciences and Medicine, Health Economics, Primary Care and Public Health Sciences; Office of Health Economics, ; INTO City University London

Title: Using EQ-5D Data to Measure Hospital Performance: Are General Population Values Distorting Patients' Choices?

Running head: EQ-5D value sets and hospital performance estimates

Authors:

1. Nils Gutacker, PhD[1], Centre for Health Economics, University of York, York, UK
2. Thomas Patton, PhD, Centre for Health Economics, University of York, York, UK
3. Koonal Shah, PhD, Office of Health Economics, London, UK
4. David Parkin, DPhil, Office of Health Economics, London, UK *and* Department of Economics, City, University of London, London, UK

Manuscript word count: 4,758

Abstract word count: 275

Acknowledgements: The authors thank Helen Dakin, John Brazier, Matthijs Versteegh, three anonymous referees and participants at the 2017 PROMs conference in Oxford and the EuroQol 34th Scientific Plenary Meeting (Barcelona, 2017) for useful comments and suggestions. Financial support for this study was provided entirely by a grant from the EuroQol Research Foundation. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. NG, KS and DP are members of the EuroQol Group. No ethical approval was required for analysis of anonymised, secondary data.

[1] Corresponding author. Nils Gutacker, Centre for Health Economics, University of York, Heslington, YO10 5DD, UK. E-mail: nils.gutacker@york.ac.uk

Abstract

Background: The English NHS publishes hospital performance indicators based on average post-operative EQ-5D index scores after hip replacement surgery to inform prospective patients' choices of hospital. Unidimensional index scores are derived from multidimensional health-related quality of life data using preference weights estimated from a sample of the UK general population. This raises normative concerns if general population preferences differ from those of the patients that are to be informed. This study explores how the source of valuation affects hospital performance estimates.

Methods: Four different value sets reflecting source of valuation (general population vs. patients), valuation technique (visual analogue scale (VAS) vs. time trade-off (TTO)) and experience with health states (currently experienced vs experimentally estimated) were used to derive and compare performance estimates for 243 hospitals. Two value sets were newly estimated from EQ-5D-3L data on 122,921 hip replacement patients and 3,381 members of the UK general public. Changes in hospital ranking (nationally) and performance outlier status (nationally; amongst patients' five closest hospitals) were compared across valuations.

Results: National rankings are stable under different valuations (rank correlations > 0.92). Twenty-three (9.5%) hospitals change outlier status when using patient VAS valuations instead of general population TTO valuations, the current approach. Outlier status also changes substantially at local level. This is explained mostly by the valuation technique, not the source of valuations or experience with the health states.

Limitations: No patient TTO valuations were available. Effect of value set characteristics could only be established through indirect comparisons.

Conclusion: Different value sets may lead to prospective patients' choosing different hospitals. Normative concerns about the use of general population valuations are not supported by empirical evidence based on VAS valuations.

1 Introduction

Patients in the English National Health Service (NHS) have the right to choose among all qualified hospital providers for treatments that are deemed clinically appropriate and are publicly funded. To inform *"patients […] exercising choice"* (p.6)[1] about the quality of care they are likely to receive, the English NHS routinely collects multidimensional health-related quality of life (HRQoL) data from patients before and after undergoing planned hip and knee replacement surgical as part of the national patient-reported outcome measures (PROMs) programme. These data are then used to benchmark hospitals and calculate performance indicators in the form of case-mix adjusted average post-operative HRQoL, expressed as unidimensional composite scores, which are made publicly available on a regular basis.[2, 3]

A normative question, and the focus of this paper, is how to aggregate the multidimensional HRQoL data into unidimensional (single number) scores for the purpose of hospital performance assessment and public reporting. The PROMs programme collects HRQoL data using a generic health measurement instrument, the EQ-5D-3L[4], which comprises both a direct and indirect measure of a patient's health state. The direct measure, the EQ VAS, asks patients to provide a summary assessment of their HRQoL by marking a position on a visual analogue scale (VAS) ranging from 0 to 100, where the endpoints reflect the best and worst health states imaginable. The indirect measure uses the EQ-5D-3L descriptive system, where patients are asked to describe their current health status according to five dimensions of health (mobility, self-care, usual activities, pain & discomfort and anxiety & depression), each of which can be assigned one of three severity levels (essentially no, some or extreme problems). The resulting health profile data are aggregated into unidimensional composite ('index') scores using preference estimates of the UK general population[5], rather than of those prospective patients the PROMs programme seeks to inform. Previous research has shown many cases where preference estimates derived from specific patient populations differ systematically from those derived from the general population[6-10] although some studies find no differences [11, 12]. The current practice therefore raises normative concerns and could be inconsistent with the notion of patient sovereignty if it leads to a mismatch between the decisions patients make based on official published data, and those they would have made had the information reflected their own preferences more closely.

Ideally, the reported hospital performance should reflect prospective patients' individual preferences over relevant health states. However, the elicitation of personal preference functions is a complex and time-consuming task[13] and has therefore not (yet) found widespread adoption in the public reporting of hospital performance. Furthermore, it would imply the need to re-calculate public reports for each prospective patient based on their individual preferences, ruling out static performance reports (e.g. rankings published in newspapers) that are common currently. A pragmatic solution, that avoids both issues, is to develop a value set based on preferences elicited from a sample of patients. Such value sets are likely to reflect the preferences of prospective patients more closely than a general population value set since they are obtained from a sample of individuals with a similar age-sex structure, clinical condition, adaptation to their condition, and expectations of future health. At the same time, it would enable the calculation of EQ-5D index scores and, hence, unidimensional hospital performance indicators that could be presented alongside detailed dimension-by-dimension estimates[14] if desired.

In this paper we test whether the use of patient or general population valuations generates different hospital performance estimates for hip replacement surgery in the English NHS. We are not aware of

a UK-based patient value set that mirrors the currently used general population value set in terms of two other important aspects, namely respondents' experience of the health state to be valued as well as the valuation technique employed. This precludes a direct test of the effect of the source of valuation on hospital performance estimates. Instead, we compare hospital performance estimates generated under four published and newly-estimated value sets, out of eight possible combinations of these value set attributes. This allows us to vary one aspect at a time, holding the other two constant. The results of this indirect comparison help to demonstrate the practical implications of the normative argument about the source of health state valuations in the context of informing prospective patients about where to have surgery.

2 Valuation of health states

Amongst the desirable properties of a measure of the value of health is that it should unambiguously indicate whether a given health state, as defined by a multidimensional HRQoL profile, is better than, worse than, or equivalent to another health state. This property is most usually achieved by aggregating HRQoL data into a single number that represents the value of a health state by means of a set of preference weights. By convention, the value of a health state lies on a scale where 1 represents health which as good as possible, and 0 represents health that is either as poor as possible or is equivalent to being 'dead'. The latter allows for health states 'worse than dead' with values below 0.

Any attempt to value health in this way requires consideration of the following questions: (1) what is being valued; (2) whose values are being sought; and (3) what technique is being used to obtain the values? These are each briefly summarised below with interested readers being referred to detailed discussions elsewhere.[15-17]

2.1 What is being valued

Health state valuations are obtained as part of elicitation tasks. In these, participants may be asked to value their own health, as experienced either currently or in the past, or a set of health states that they may not be currently experiencing. For the latter, they are usually asked to value a stylised description of health, which may take the form of a health state profile comprising a series of dimensions and severity levels defined by the descriptive system of a PROM instrument, such as the EQ-5D. Such profiles are often described as 'hypothetical', but this is misleading because they are intended to reflect real health states and therefore plausible ways in which someone might self-report their health using the instrument. Since in most cases respondents will neither be experiencing or ever have experienced a health state described in the profile, they would need to imagine living in that health state in order to evaluate it. We can therefore regard these as their estimate of how they would value the health state if they were experiencing it.

2.2 Whose values are being sought

Health state valuations can be obtained from selected subgroups, such as patients with a given medical condition, or a sample of the general population.[16, 18, 19] Both approaches have merit, although advocates tend to argue their case on different grounds. Those in favour of using patient valuations typically point out that patients have first-hand experience of health states and therefore do not need to imagine the impact of an unfamiliar health state on their HRQoL.[20, 21] A common finding in the published literature - that valuations derived from specific patient populations tend to be higher than those elicited from the general population - has been attributed to patients adapting to their impaired health state over time and/or providing a more accurate assessment of the health

state based on their lived experience.[6, 7, 21] Conversely, proponents of general population valuations typically argue their case not on the grounds of validity but based on the intended use of such valuations to inform resource allocation decision in collectively funded health services, where decisions should reflect the preferences of the general population paying into the system.[18]

It is important to note that what is being valued and by whom are two separate issues. Patients may be asked to value health states that can occur as a result of their medical condition and which they may be able to imagine living in, but which they have not (yet) experienced themselves. Equally, the general population can be asked to value their currently experienced health state.[22]

2.3 What elicitation technique is being used

There are a number of techniques for valuing health states such as VAS and time trade-off (TTO).[23] The VAS involves rating the health state on a scale with imposed interval properties and well-defined endpoints, conventionally 0 and 100 (which in the EQ VAS represent worst and best imaginable health, respectively). TTO involves making a series of choices between living for a fixed amount of time in the profile under evaluation and a shorter, variable amount of time in full health, where the point at which respondents are indifferent is used to infer valuations. TTO has become the method most often recommended for the generation of values. The two methods have different assumptions underpinning them and are subject to different types of framing effects, for example VAS valuations are known to be subject to end-of-scale aversion [24] whereas respondents' time preference can have an effect on TTO valuations [25, 26]. VAS exercises are widely considered to be relatively simple and feasible to complete.[27] Previous research has shown that VAS and TTO yield different results.[28]

3 Methods

3.1 Data

We analyse EQ-5D-3L data from two independent samples. The first consists of 272,445 NHS-funded total hip replacement (THR) patients aged 15 years or over who had primary surgery in public or private hospitals in England between April 2012 and March 2016, collected as part of the English national PROMs programme[1]. Patients completed a paper questionnaire shortly before and six months after having surgery, containing the EQ-5D-3L, a condition-specific measure (the Oxford Hip Score) and other questions about their condition and treatment. The pre-operative questionnaire was administered by hospital staff at admission or the last outpatient appointment preceding admission and forwarded to a central data processor. The post-operative questionnaire was mailed directly to the patient's home address. Returned questionnaires were linked to administrative hospital records from the Hospital Episode Statistics (HES) database through a probabilistic matching algorithm. HES provides information on patient's age, place of residence, provider of care, and whether the surgery was a revision of a previous THR. Further details about the PROM data collection procedure are provided elsewhere.[29, 30] We excluded patients for whom pre- or post-operative responses were missing, either in part or completely, or where questionnaires could not be linked to HES. The sample used to estimate the patient value set in this study included 122,921 patients, which corresponds to 45.1% of all THR patients that were eligible to participate in the PROMs survey. Excluded patients were on average slightly younger and more likely to be female (Appendix Table 1). The linked HES-PROMs dataset was provided by NHS Digital.

The second sample consists of 3,381 randomly selected members of the UK general public that took part in the Measurement and Valuation of Health (MVH) study.[31] Each of the participants were

asked as part of face-to-face interviews to rate their own health status using the EQ-5D-3L questionnaire and to value 8 of 42 stylised health states using TTO[32] and VAS. The valuation data were used to derive a TTO based value set known as the MVH-A1 [5], but which we label the GP-TTO-VAL, and a VAS based value set known as the MVH-A3, but which we label the GP-VAS-VAL (Table 1).[31] The former is used in the official calculation of the hospital performance estimates reported to the public. Both value sets are anchored at 1 (full health) and 0 (dead), with scores below 0 indicating states considered worse than being dead. The MVH dataset was provided by the UK Data Services.

3.2 Estimation of experience-based value sets

A patient, current health VAS value set, which we label the PAT-VAS-OWN, was derived from the national PROMs dataset by regressing patient-reported EQ VAS scores on variables representing the levels within each dimension of the EQ-5D descriptive system, using Ordinary Least Squares. The regression model underpinning the MHV value sets include dummy variables for the main effects, a constant term reflecting any deviation from full health, and an N3 term indicating extreme problems (level 3) on any dimension.[5] To ensure comparability with these, we used the same specification. We also estimated more saturated models allowing for pairwise interactions between dimensions at level 2 and 3, but found these added little to overall fit (results available on request).

The PAT-VAS-OWN value set was estimated on data for the period April 2012 to March 2015, leaving one year of data to assess the impact of the value set on hospital rankings (see Section 2.4). It has been observed that patients' valuations of the same description of their health state may change from pre- to post-surgery, which may lead to inconsistencies when estimating patient-based value sets.[33] We focus our analysis on pre-operative survey responses since these are more likely to reflect patients' preferences at the point in time when a choice is to be made.

We also estimated a general population, current health VAS value set, which we label the GP-VAS-OWN, using the MVH study participants' EQ VAS and self-classifier responses and the same modelling structure as for the PAT-VAS-OWN value set.

Table 1 summarises the characteristics of the four value sets that we compared.

All standard errors are robust to heteroscedasticity and, in the case of the PAT-VAS-OWN value set, are clustered at hospital level. All computations were performed in Stata 14 (StataCorp LP, College Station, TX).

3.3 Deriving hospital performance estimates

Hospital performance assessment aims to identify the systematic contribution that providers make to their patients' health outcomes.[34] To allow for fair comparisons these assessments need to adjust for differences in hospital case-mix and sampling uncertainty.

Our analysis followed the published adjustment methodology of NHS England[35], in which the case-mix adjusted performance θ_j of hospital $j = 1,\ldots,J$ is estimated as

$$\hat{\theta}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}(O_{ij} - \hat{E}_{ij})$$

where O_{ij} is the observed post-operative index score for patient $i = 1,\ldots,n_j$ and E_{ij} is the expected post-operative index score for the same patient given their observable characteristics.

The expected post-operative index score is based on the official case-mix adjustment methodology developed by NHS England[35]. The adjustment takes account of age, gender, ethnicity, living arrangements, the income deprivation profile of the patients' local small areas of residence (Lower-Layer Super Output Area (LSOA)) as approximated by the 2010 Index of Deprivation[36], main diagnosis and comorbid conditions, whether patients lived alone, whether they required assistance when filling in the PROMs questionnaire or considered themselves to be disabled, the duration of symptoms, as well as their pre-operative EQ-5D index score. We estimate the case-mix adjustment model separately for each of the four value sets using data from April 2012 to March 2014.

To account for sampling uncertainty in performance scores we follow standard practice[37-39] in the NHS and calculated z-score statistics for each hospital as

$$Z_j = abs\left(\frac{\hat{\theta}_j}{SE(\hat{\theta}_j)}\right)$$

where $Z_j > 1.96$ indicates statistically significant divergent performance from the national average at the 5% level. Hospitals with $Z_j > 1.96$ were deemed to perform *well* if $\hat{\theta}_j > 0$ and *poorly* otherwise. Performance estimates that were not statistically significantly different from the national average were deemed *average*. This approach is consistent with the simplified pictorial display used to communicate performance information (green, blue and red buttons to denote good, average and poor performance) by NHS choices[2] and other hospital comparison websites[3].

3.4 Assessing the impact of different EQ-5D value sets on hospital performance estimates

We assessed the impact of different value sets on hospital performance estimates for the period between April 2015 and March 2016 through a series of head-to-head comparisons. For each hospital, we compared their performance status (i.e. whether they were judged to perform well, poorly, or average) under different value sets and quantified discrepancies. The strength of association between hospital performance rankings generated with different value sets was measured using Spearman's rank correlation coefficient ρ.

One motivation for considering patient valuations in assessing hospital performance is the desire to provide prospective patients with information that will inform their choice of hospital. Yet, most patients are unwilling to travel far for healthcare treatment[40-42], with a recent study[43] suggesting that over 92% of THR patients in the English NHS chose to attend one of their five closest hospitals in the period 2010 to 2012. We therefore also explored the impact of value sets at the local level; for each patient, we assessed how many of their five closest hospitals would be flagged as performing well or poorly under the different value sets. This 'choice set' was determined by the straight-line distance between the centroid of the patient's LSOA of residence and the hospitals' postcodes.[43]

4 Results

4.1 Descriptive statistics

Table 2 reports descriptive statistics of the data samples. Patients in the national PROMs programme sample were, on average, 68 years old and 58.7% were female. Most patients had suffered from joint-related symptoms for 1 to 5 years prior to surgery. The average improvement in HRQoL six

months after surgery was equivalent to an increase of 0.43 value points (from 0.37 to 0.80) (MVH-A1 value set), and patients' overall assessment of their health as measured by the EQ VAS increased by 12 points (from 65 to 77). Patients described their pre-operative HRQoL using 148 of the 243 possible EQ-5D-3L health states. The relative frequency of these health states was consistent with the severity of the conditions that require major joint replacement. Over 46% of patients reported extreme limitation (i.e. level 3 problems) on at least one HRQoL dimension before surgery, and >2% reported extreme limitations on three or more dimensions.

Unsurprisingly, MVH study participants reported better health on average than the patient sample, both pre- and post-surgery. They were, on average, significantly younger (mean age = 47.9 years) than the patient population but showed a similar sex split (56.7% female). Participants described their health using 77 of the 243 EQ-5D-3L health states, with 4.8% of participants having at least one extreme limitation on any of the five health dimensions. The average VAS score was 82.5 and the average EQ-5D value based on the GP-TTO-VAL value set was 0.86.

4.2 Value sets

Table 3 reports the estimated PAT-VAS-OWN and GP-VAS-OWN value sets alongside the published GP-TTO-VAL and GP-VAS-VAL value sets. Coefficient estimates represent decrements associated with some or extreme limitations on a given health dimension. The constant and the N3 term reflect global decrements that are applied in the presence of *any* limitations on *any* health dimension and at least one extreme limitation on any health dimension, respectively.

Figure 1 shows the values generated by the different value sets for the 42 stylised health states valued in the MVH study.

Both PAT-VAS-OWN and GP-VAS-OWN value sets were found to be internally consistent, i.e. more severe limitations are associated with larger decrements for each dimension. Patients assign approximately equal or smaller decrements to health problems on a given *dimension* than the general public, but they attach a larger global decrement to the presence of *any* health problems as reflected in the coefficient on the constant term. Differences are more pronounced on level 3 decrements than level 2 decrements, thus generating a wider spread of index scores across the four value sets for health states for which respondents reported at least one extreme problem. These results are consistent with previous evidence from other patient populations.[7, 44] It should be noted that because of the smaller sample size, the GP-VAS-OWN data has sparse observations in some of the levels within dimensions, particularly Mobility Level 3, which means that the coefficient estimates have very large standard errors.

Table 4 reports descriptive statistics of the pre- and post-operative index scores reported at patient level (mean, SD) as well as the range of hospital average scores calculated using the four value sets. Differences in average index scores are more pronounced prior to surgery than afterwards, which reflects the low number of patients reporting any extreme problems after surgery. The two value sets based on direct valuations of own, currently experienced health (GP-VAS-OWN, PAT-VAS-OWN) generate, on average, higher index scores as well as a smaller spread of hospital average index scores that are relevant for performance assessment. Histograms of case-mix adjusted hospital scores are presented in the online appendix.

4.3 Impact on judgements about hospital performance

Figure 2 presents scatter plots of hospital z-scores derived under different EQ-5D value sets. Each scatter point represents one hospital, with dashed lines indicating the lower and upper boundaries

at which performance estimates are deemed to be statistically significantly different from the national average. Performance estimates that would lead to differential judgement under the two value sets being compared are highlighted as diamonds (significant under the first but not the second value set) or squares (vice versa).

The GP-TTO-VAL and PAT-VAS-OWN value sets generate performance estimates that are highly correlated (ρ = 0.92) (Figure 2, Panel A). Despite this, the change in value set has a non-negligible impact on how individual hospitals are deemed to perform, with patient valuations leading to changes in outlier status for 23 hospitals in total (9.5% of 243), of which 6 (2.5%) are no longer identified as performing poorly, 10 (4.1%) are no longer identified as performing well, and seven different hospitals now appear to perform well (2.9%). At the local level, 1% fewer patients (44% vs 45% of N=65,278) receiving care between April 2015 and March 2016 would have found at least one well performing hospital within their five closest hospitals if performance estimates had been derived using the PAT-VAS-OWN value set rather than the GP-TTO-VAL (Figure 3). In contrast, patients would have been 10% more likely (34% vs 24%) to find at least one local hospital deemed to perform poorly if performance estimates had been derived using the PAT-VAS-OWN value set. Overall, at least one performance assessment for their five closest hospitals would have been different for 8.6% of patients receiving care between April 2015 and March 2016.

To further explore the reasons for this divergence, we compared hospital performance estimates derived varying one value set design characteristic (i.e. source of valuation, valuation technique, or experience with health state) while holding the others constant (Figure 2, Panels B-D). The results of this marginal analysis suggest that neither the source of valuation nor the level of experience with a health state drive the observed differences in hospital performance classifications. Instead, these differences can be explained nearly entirely by the choice of valuation technique employed, with Panel B showing many more changes in outlier status than Panel C and D.

5 Discussion

There is a strong normative rationale for using patient values to aggregate multidimensional HRQoL instruments when developing hospital performance indicators to inform prospective patients' choices of hospital. However, the standard practice in the English NHS has been to publish hospital performance indicators based on EQ-5D scores aggregated using general public values. The present study explores whether this practice may be distorting patients' choice of hospital for hip replacement surgery given that there is some evidence of discrepancies between patient and general public values. We find a larger number of hospitals are deemed to perform poorly when a patient VAS tariff (PAT-VAS-OWN) is used compared to when the UK general population TTO tariff (GP-TTO-VAL) is used. Conversely, we find only slightly fewer hospitals are deemed to perform well when using the PAT-VAS-OWN instead of the GP-TTO-VAL value set. The choice of value set therefore appears to be more important for patients seeking to avoid poorly performing hospitals. Moreover, we find that the GP-TTO-VAL tariff overvalues the relative performance of hospitals that deliver improvements in pain/discomfort and mobility compared to the PAT-VAS-OWN tariff whilst undervaluing those that perform relatively well at addressing anxiety/depression problems. Importantly, these differences appear to be driven almost entirely by the difference in the health state valuation technique employed (TTO vs VAS) rather than the source of valuations. Therefore, our results provide little empirical support for a change in reporting practice in the English PROMs programme because of normative concerns about the source of valuations.

In recent years, there has been considerable interest in the use of values that reflect individuals' own health, rather than their estimated valuations of stylised health states, to derive value sets.[22, 45] The purported rationale for using 'experience-based' values is that they avoid some of the focusing effects that can occur in the valuation of stylised health states.[20] Furthermore, any need to reflect the preferences of the tax-paying general population, which mainly arises in the context of economic evaluation of new health technologies for use in publicly-funded health systems, can be addressed by using a population survey.[22] One concern with this approach is that the data collected for the purposes of developing an experience-based value set may only contain a limited range of responses to the health state descriptive system. Our study provides further evidence to demonstrate the feasibility of developing an experience-based value set from large-scale, routinely collected PROM surveys. Patients in the hip replacement sample report their HRQoL according to 148 of the 243 possible EQ-5D-3L health states; covering a broad range of the instrument's spectrum. By design, these are also the most commonly encountered health states in this population, limiting the need to extrapolate beyond the set of valued health states in most applications.

While not the focus of our study, our findings also provide additional context to the debate about the comparability of EQ-5D-3L value sets developed in different countries. A study by Nemes and colleagues developed an experience-based VAS value set for the EQ-5D-3L using data from patients undergoing elective total hip replacement in Sweden.[46] The valuations of health dimensions in the Swedish study and those in our study are similar in that the most important dimension – both in terms of the decrements associated with the level 2 and 3 responses – is anxiety/depression (see Appendix Table 2 for estimates). Aside from this similarity, the relative importance of the various health dimensions differ systematically for the two value sets. This casts doubt on the ability to pool experienced-based value sets across countries as recently suggested for TTO value sets based on valuations of health states derived from valuation studies.[47]

There are a number of limitations to our analysis and proposed approach. First, a single patient group value set still requires aggregating valuations over a large number of patients with potentially heterogeneous preferences. While it is reasonable to assume that the mismatch between the average patient value set and individual patients' preferences is smaller than the mismatch with average general population preferences, there may be scope for further refinement. Some existing work has explored how health state valuations vary with observable characteristics of the respondent and this line of inquiry ought to be expanded.[48] Secondly, the relationship between direct valuations of health states as reflected in EQ VAS scores and patients' EQ-5D-3L health profiles has been found to change from before to after surgery.[33] The reason for this discrepancy remains unclear. We have chosen to estimate patient valuations from their pre-operative data since this reflects their ex-ante valuations at the time of their decisions. However, one may also argue that post-operative valuations are appropriate as they reflect patients' preferences over different outcomes once they have started to experience the benefits of treatment. This distinction is not the focus of this paper, although we note that it appears to have little effect on hospital performance estimates, which are highly correlated under both value sets (rho>0.99) (see Appendix Table 3 for post-operative PAT-VAS value set and the online appendix for hospital performance scatter plots). Thirdly, while we find that the source of valuation is not a major driver of hospital performance estimates when valuing health states using VAS, we cannot generalise this statement to other valuation techniques such as the TTO valuations currently used in the NHS. To test this we would require TTO data from a sample of hip replacement patients, which we do not currently have access to. Fourthly, the generalizability of the findings in our study is limited to the medical condition and the decision problem under consideration. Finally, the limited amount of provider variation in both intake and health gain following THR surgery may limit the role that valuations play in determining

hospital performance estimates.[49] As routine PROM collection becomes more prevalent, this hypothesis will become testable.

In conclusion, the choice of value set to aggregate EQ-5D-3L health profiles in the context of the English PROMs programme may have real implications for patients choosing hospitals for their THR surgery. This is particularly relevant when choices are based on simple heuristics, e.g. selection based on dichotomized performance status rather than index scores. However, this divergence does not appear to be driven by the source of health state valuations, a normative concern, but rather by the valuation technique employed, a technical matter.

Appendix

Appendix Table 1: Comparison of included and excluded patients in PROMs sample

	Hip replacement patient sample			
	Excluded		Included	
Patient age (mean, sd)	67.96	12.10	68.26	10.32
Patient gender (n, %)				
Female	90,887	61%	72,095	59%
Male	58,335	39%	50,826	41%
*Financial year of treatment (n, %) **				
2012/13	35,259	24%	28,270	23%
2013/14	35,275	24%	33,148	27%
2014/15	39,064	26%	31,591	26%
2015/16	39,624	27%	29,912	24%

** Financial years run from April to March.*

Appendix Table 2: Experienced-based VAS value sets for total hip replacement patients in England and Sweden

	England		Sweden	
	Est	SE	Est	SE
Full health	1.000		0.745	
Mobility, level 2	-0.056	0.003	-0.060	
Mobility, level 3	-0.119	0.012	-0.098	
Self-care, level 2	-0.055	0.002		
Self-care, level 3	-0.116	0.008		
Self-care, level 2 or 3			-0.038	
Usual activities, level 2	-0.029	0.003	-0.053	
Usual activities, level 3	-0.086	0.004	-0.110	
Pain/Discomfort, level 2	-0.050	0.008	-0.025	
Pain/Discomfort, level 3	-0.130	0.009	-0.124	
Anxiety/Depression, level 2	-0.088	0.002	-0.078	
Anxiety/Depression, level 3	-0.181	0.004	-0.161	
N3	0.011	0.003		
Any deviation from full health	-0.182	0.008		

Notes: Swedish values are taken from Nemes et al. (2015). Signs on coefficient estimates for England have been reversed to be compatible with Swedish values.

Appendix Table 3: PAT-VAS-OWN value sets calculated from PROMs data collected before or six months after surgery

	Before surgery		After surgery	
EQ-5D dimension	Est	SE	Est	SE
Mobility, level 2	0.056	0.003	0.075	0.001
Mobility, level 3	0.119	0.012	0.177	0.024
Self-care, level 2	0.055	0.002	0.059	0.002
Self-care, level 3	0.116	0.008	0.086	0.009
Usual activities, level 2	0.029	0.003	0.050	0.001
Usual activities, level 3	0.086	0.004	0.119	0.006
Pain/Discomfort, level 2	0.050	0.008	0.032	0.001
Pain/Discomfort, level 3	0.130	0.009	0.121	0.006
Anxiety/Depression, level 2	0.088	0.002	0.083	0.001
Anxiety/Depression, level 3	0.181	0.004	0.167	0.007
N3	-0.011	0.003	0.034	0.006
Constant	0.182	0.008	0.131	0.001

References

[1] Department of Health. Guidance on the routine collection of Patient Reported Outcome Measures (PROMs). London: The Stationery Office 2008.
[2] NHS. Find services. 2019 10/12/2019]Available from: www.nhs.uk/service-search
[3] National Joint Registry. NJR Surgeon and Hospital Profile for hip, knee, ankle, elbow and shoulder joint replacement surgery. 2019 Available from: https://surgeonprofile.njrcentre.org.uk/
[4] Brooks R. EuroQol: the current state of play. *Health Policy*. 1996; 37(1):53-72.
[5] Dolan P. Modeling Valuations for EuroQol Health States. *Medical Care*. 1997; 35(11):1095-108.
[6] De Wit GA, Busschbach JJV, De Charro FT. Sensitivity and perspective in the valuation of health status: whose values count? *Health Economics*. 2000; 9(2):109-26.
[7] Mann R, Brazier J, Tsuchiya A. A comparison of patient and general population weightings of EQ-5D dimensions. *Health Economics*. 2009; 18(3):363-72.
[8] Peeters Y, Stiggelbout AM. Health State Valuations of Patients and the General Public Analytically Compared: A Meta-Analytical Comparison of Patient and Population Health State Utilities. *Value in Health*. 2010; 13(2):306-9.
[9] Goodwin E, Green C, Hawton A. What Difference Does It Make? A Comparison of Health State Preferences Elicited From the General Population and From People With Multiple Sclerosis. *Value in Health*. 2019.
[10] Rand-Hendriksen K, Augestad L, Kristiansen I, Stavem K. Comparison of hypothetical and experienced EQ-5D valuations: relative weights of the five dimensions. *Quality of Life Research*. 2012; 21:1005-12.
[11] Pickard AS, Tawk R, Shaw JW. The effect of chronic conditions on stated preferences for health. *The European Journal of Health Economics*. 2013; 14(4):697-702.
[12] Gandhi M, Tan RS, Ng R, Choo SP, Chia WK, Toh CK, et al. Comparison of health state values derived from patients and individuals from the general population. *Quality of Life Research*. 2017; 26(12):3353-63.
[13] Devlin NJ, Shah KK, Mulhern BJ, Pantiri K, van Hout B. A new method for valuing health: directly eliciting personal utility functions. *European Journal of Health Economics*. 2019; 20(2):257-70.
[14] Gutacker N, Bojke C, Daidone S, Devlin N, Street A. Hospital Variation in Patient-Reported Outcomes at the Level of EQ-5D Dimensions: Evidence from England. *Medical Decision Making*. 2013; 33(6):804-18.
[15] Versteegh MM, Brouwer WBF. Patient and general public preferences for health states: A call to reconsider current guidelines. *Social Science & Medicine*. 2016; 165:66-74.
[16] Brazier J, Akehurst R, Brennan A, Dolan P, Claxton K, McCabe C, et al. Should Patients Have a Greater Role in Valuing Health States? *Applied Health Economics and Health Policy*. 2005; 4(4):201-8.
[17] Brazier J, Rowen D, Karimi M, Peasgood T, Tsuchiya A, Ratcliffe J. Experience-based utility and own health state valuation for a health state classification system: why and how to do it. *European Journal of Health Economics*. 2018; 19(6):881-91.
[18] Gold M, Siegel J, Russell L, Weinstein M, eds. Cost-effectiveness in health and medicine. New York: Oxford University Press 1996.
[19] Dolan P. Whose Preferences Count? *Medical Decision Making*. 1999; 19(4):482-6.
[20] Dolan P, Kahneman D. Interpretations Of Utility And Their Implications For The Valuation Of Health. *Economic Journal*. 2008; 118(525):215-34.
[21] Menzel P. Utilities for Health States: Whom to Ask. In: Culyer AJ, ed. *Encyclopedia of Health Economics*. Amsterdam: Elsevier 2014:417-24.
[22] Burström K, Sun S, Gerdtham U-G, Henriksson M, Johannesson M, Levin L-Å, et al. Swedish experience-based value sets for EQ-5D health states. *Quality of Life Research*. 2014; 23(2):431-42.

[23] Brazier J, Ratcliffe J, Saloman J, Tsuchiya A. *Measuring and Valuing Health Benefits for Economic Evaluation*. 2 ed. Oxford: Oxford University Press 2016.
[24] Shmueli A. The visual analog rating scale of health-related quality of life: an examination of end-digit preferences. *Health and Quality of Life Outcomes*. 2005; 3(1):71.
[25] Attema AE, Brouwer WBF. On the (not so) constant proportional trade-off in TTO. *Qual Life Res*. 2010; 19(4):489-97.
[26] Dolan P, Stalmeier P. The validity of time trade-off values in calculating QALYs: constant proportional time trade-off versus the proportional heuristic. *Journal of Health Economics*. 2003; 22(3):445-58.
[27] Green C, Brazier J, Deverill M. Valuing health-related quality of life. A review of health state valuation techniques. *Pharmacoeconomics*. 2000; 17(2):151-65.
[28] Robinson A, Dolan P, Williams A. Valuing health status using VAS and TTO: what lies behind the numbers? *Social Science & Medicine*. 1997; 45(8):1289-97.
[29] Gutacker N, Street A, Gomes M, Bojke C. Should English healthcare providers be penalised for failing to collect patient-reported outcome measures (PROMs)? *Journal of the Royal Society of Medicine*. 2015; 108(8):304-16.
[30] NHS Digital. *Patient Reported Outcome Measures (PROMs) in England - A guide to PROMs methodology*2017.
[31] MVH Group. *Final Report on the Modelling of Valuation Tariffs*: Centre for Health Economics, University of York; 1995.
[32] Walker S, Sculpher M, Drummond M. The Methods of Cost-Effectiveness Analysis to Inform Decisions about the Use of Health Care Interventions and Programs. In: Glied S, Smith PC, eds. *The Oxford Handbook of Health Economics* Oxford: Oxford University Press 2012.
[33] Pickard AS, Hung Y-T, Lin F-J, Lee TA. Patient Experience-based Value Sets: Are They Stable? *Medical Care*. 2017; 55(11):979-84.
[34] Jacobs R, Smith PC, Street A. *Measuring Efficiency in Health Care*. Cambridge: Cambridge University Press 2006.
[35] NHS England. *Patient Reported Outcome Measures (PROMs) in England - Update to reporting and case-mix adjusting hip and knee procedure data. Annex 1: Coefficients for primary hip replacement models*2017.
[36] McLennan D, Barnes H, Noble M, Davis J, Garrett E, Dibben C. The English Indices of Deprivation 2010. In: Government DfCaL, ed. London2011.
[37] Department of Health. Patient Reported Outcome Measures (PROMS) in England: a methodology for identifying potential outliers. London: The Stationery Office 2011.
[38] Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Statistics in Medicine*. 2005; 24(8):1185-202.
[39] National Clinical Audit Advisory Group. Detection and management of outliers. In: Department of Health, ed. London: The Stationary Office 2011.
[40] Luft HS, Hunt S, Maerki S. The Volume-Outcome Relationship: Practice-Makes-Perfect or Selective-Referral Patterns? *Health Services Research*. 1987; 22:157-82.
[41] Propper C, Damiani M, Leckie G, Dixon J. Impact of patients' socioeconomic status on the distance travelled for hospital admission in the English National Health Service. *Journal of Health Services Research & Policy*. 2007; 12(3):153-9.
[42] Varkevisser M, van der Geest SA, Schut FT. Do patients choose hospitals with high quality ratings? Empirical evidence from the market for angioplasty in the Netherlands. *Journal of Health Economics*. 2012; 31(2):371-8.
[43] Gutacker N, Siciliani L, Moscelli G, Gravelle H. Choice of hospital: Which type of quality matters? *Journal of Health Economics*. 2016; 50:230-46.
[44] Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. *Quality of Life Research*. 2003; 12(6):599-607.

[45] Dolan P. NICE should value real experiences over hypothetical opinions. *Nature*. 2009; 462:35.
[46] Nemes S, Burström K, Zethraeus N, Eneqvist T, Garellick G, Rolfson O. Assessment of the Swedish EQ-5D experience-based value sets in a total hip replacement population. *Quality of Life Research*. 2015:1-8.
[47] Olsen JA, Lamu AN, Cairns J. In search of a common currency: A comparison of seven EQ-5D-5L value sets. *Health Economics*. 2018; 27(1):39-49.
[48] Craig BM, Reeve BB, Cella D, Hays RD, Pickard AS, Revicki DA. Demographic Differences in Health Preferences in the United States. *Medical Care*. 2014; 52(4):307-13.
[49] Street A, Gutacker N, Bojke C, Devlin N, Daidone S. Variations in outcome and costs among NHS providers for common surgical procedures: econometric analyses of routinely collected data. *Health Services and Delivery Research*. 2014; 2(1).

Tables and figure legends

Table 1: Overview of value set characteristics

Value Set	Source of valuation	Valuation technique	Experience of health states
GP-TTO-VAL (Dolan 1997)	General population	TTO	Stylised description
GP-VAS-VAL	General population	Valuation VAS	Stylised description
GP-VAS-OWN	General population	EQ VAS	Current health
PAT-VAS-OWN	Patients	EQ VAS	Current health

Notes: TTO = Time trade-off, VAS = Visual analogue scale.

Table 2: Descriptive statistics of PROMs and MVH samples

Variable	Hip replacement (PROMs) sample		General population (MVH) sample	
Patient age (mean, sd)	68.26	10.32	47.86	18.37
Patient gender (n, %)				
Female	72,095	58.7%	1,917	56.7%
Male	50,826	41.3%	1,464	43.3%
Symptom duration (n, %)				
<1 year	16,414	13.4%		
1-5 years	84,015	68.3%		
6-10 years	13,967	11.4%		
>10 years	7,700	6.3%		
Not reported	825	0.7%		
Pre-operative EQ-5D responses (mean, sd)				
EQ-5D index score (GP-TTO-VAL)	0.37	0.32	0.86	0.23
EQ VAS score	65.43	21.55	82.53	16.90
Post-operative EQ-5D responses (mean, sd)				
EQ-5D index score (GP-TTO-VAL)	0.80	0.24		
EQ VAS score	77.34	17.61		
Number of level 3 problems (pre- or post-operatively) (n, %)				
none	66,170	53.8%	3,172	93.8%
1	39,068	31.8%	161	4.8%
2	14,905	12.1%	40	1.2%
3	2,405	2.0%	8	0.2%
4	314	0.3%	0	0.0%
5	59	0.0%	0	0.0%
Sample size	122,921		3,381	

Table 3: Estimated EQ-5D health dimension decrements and standard errors

	GP-TTO-VAL		GP-VAS-VAL		GP-VAS-OWN		PAT-VAS-OWN	
EQ-5D dimension	Est	SE	Est	SE	Est	SE	Est	SE
Mobility, level 2	0.069	0.005	0.071	0.004	0.059	0.010	0.047	0.002
Mobility, level 3	0.314	0.007	0.182	0.005	0.152	0.084	0.117	0.011
Self-care, level 2	0.104	0.005	0.093	0.004	0.067	0.018	0.057	0.001
Self-care, level 3	0.214	0.007	0.145	0.005	0.080	0.097	0.104	0.007
Usual activities, level 2	0.036	0.006	0.031	0.004	0.082	0.011	0.042	0.002
Usual activities, level 3	0.094	0.007	0.081	0.005	0.139	0.034	0.097	0.003
Pain/Discomfort, level 2	0.012	0.005	0.084	0.004	0.065	0.006	0.047	0.006
Pain/Discomfort, level 3	0.386	0.006	0.171	0.004	0.100	0.034	0.119	0.007
Anxiety/Depression, level 2	0.071	0.071	0.063	0.004	0.072	0.007	0.085	0.001
Anxiety/Depression, level 3	0.236	0.006	0.124	0.004	0.151	0.034	0.173	0.003
N3	0.269	0.007	0.215	0.005	0.064	0.036	-0.020	0.003
Constant	0.081	0.008	0.159	0.004	0.104	0.002	0.121	0.005
Source of valuation	General population		General population		General population		Patients	
Valuation technique	TTO		Valuation VAS		EQ VAS		EQ VAS	
Experience of health states	Stylised description		Stylised description		Current health		Current health	

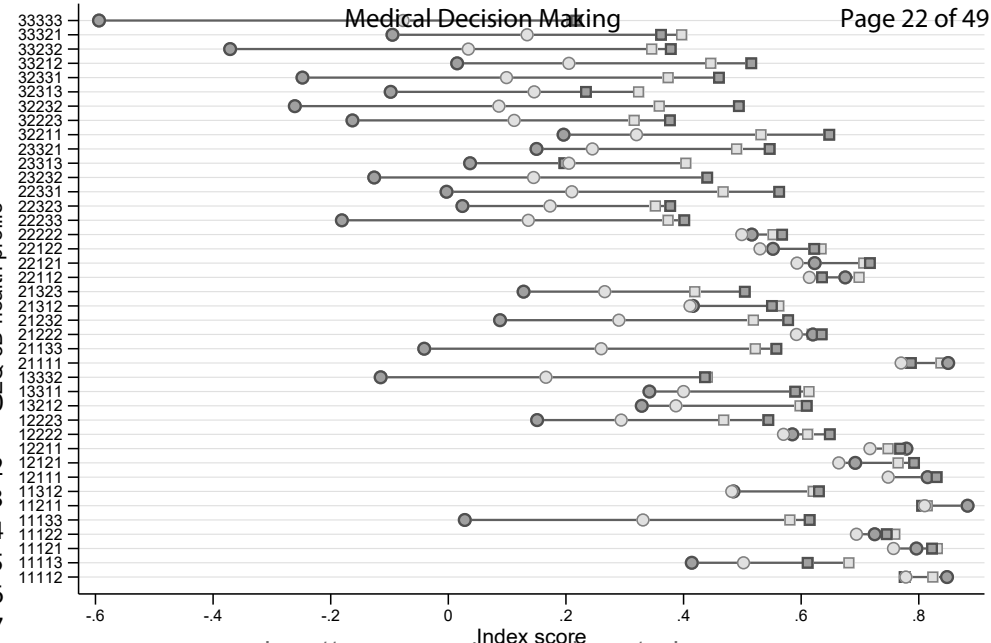
Table 4: Index scores at patient level (mean, SD) and range of scores at provider level under four value sets

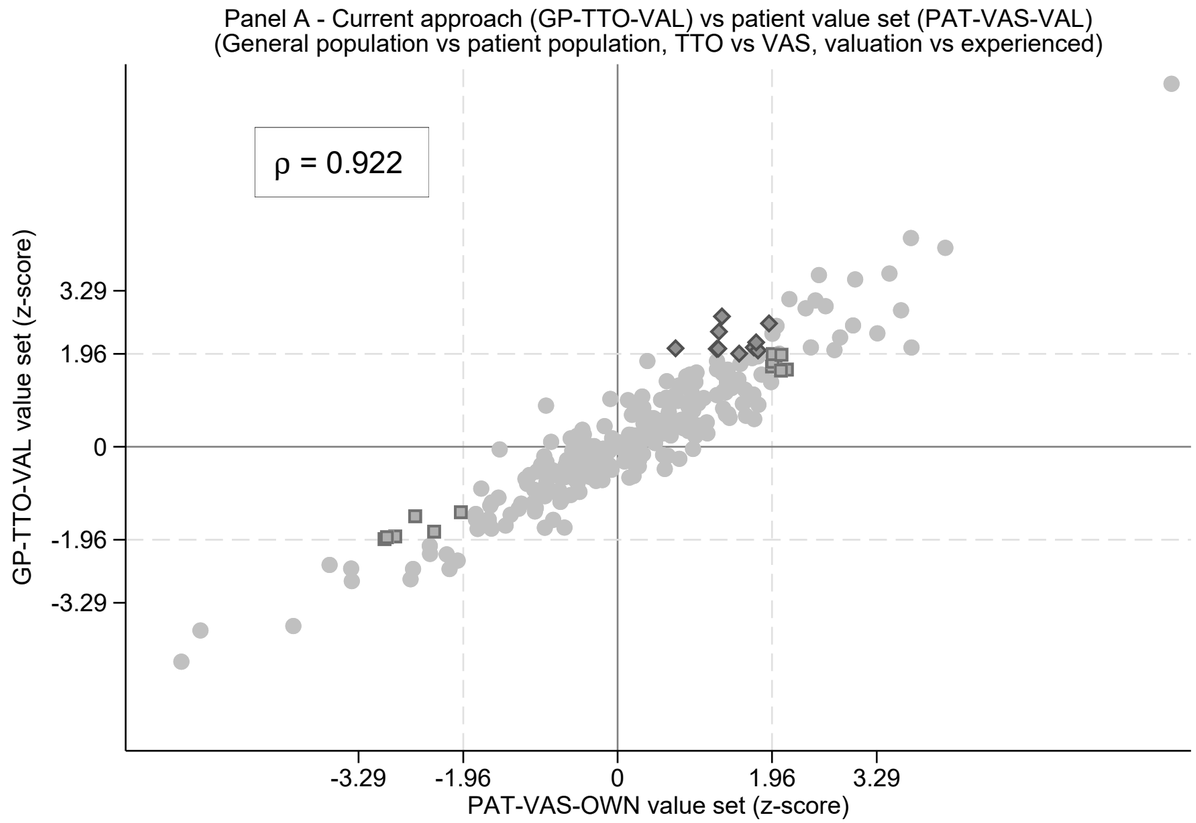
Value set	Pre-operative			Post-operative (unadjusted)			Post-operative (case-mix adjusted)		
	Mean	SD	Range of hospital mean scores	Mean	SD	Range of hospital mean scores	Mean	SD	Range of hospital mean scores
GP-TTO-VAL	0.364	0.320	0.243 to 0.576	0.802	0.239	0.568 to 1	0.804	0.216	0.632 to 1
GP-VAS-VAL	0.441	0.202	0.227 to 0.571	0.789	0.216	0.599 to 1	0.791	0.195	0.593 to 0.998
GP-VAS-OWN	0.579	0.116	0.449 to 0.673	0.826	0.173	0.687 to 1	0.828	0.155	0.629 to 0.987
PAT-VAS-OWN	0.625	0.101	0.496 to 0.711	0.832	0.162	0.708 to 1	0.834	0.144	0.646 to 0.975

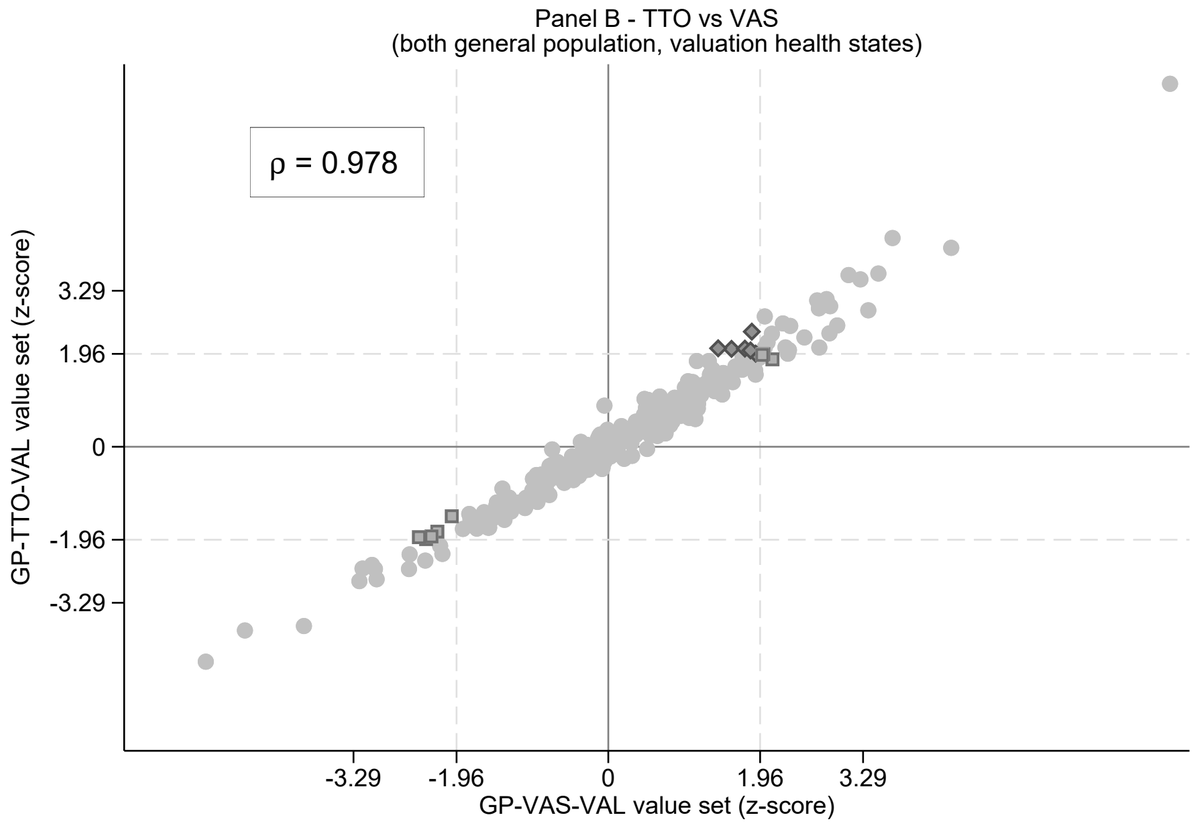
Figure 1: Selected health state valuations under different value sets

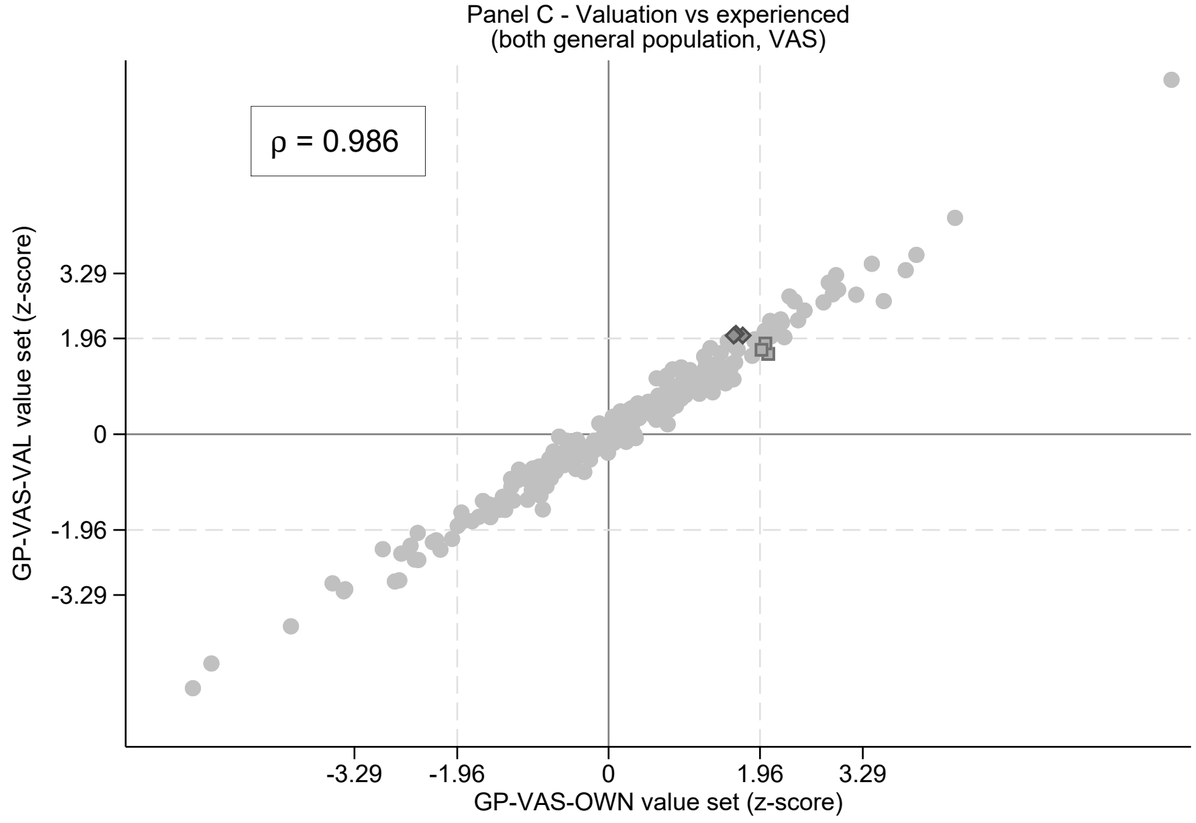
Figure 2: Relationship between hospital performance estimates under different value sets

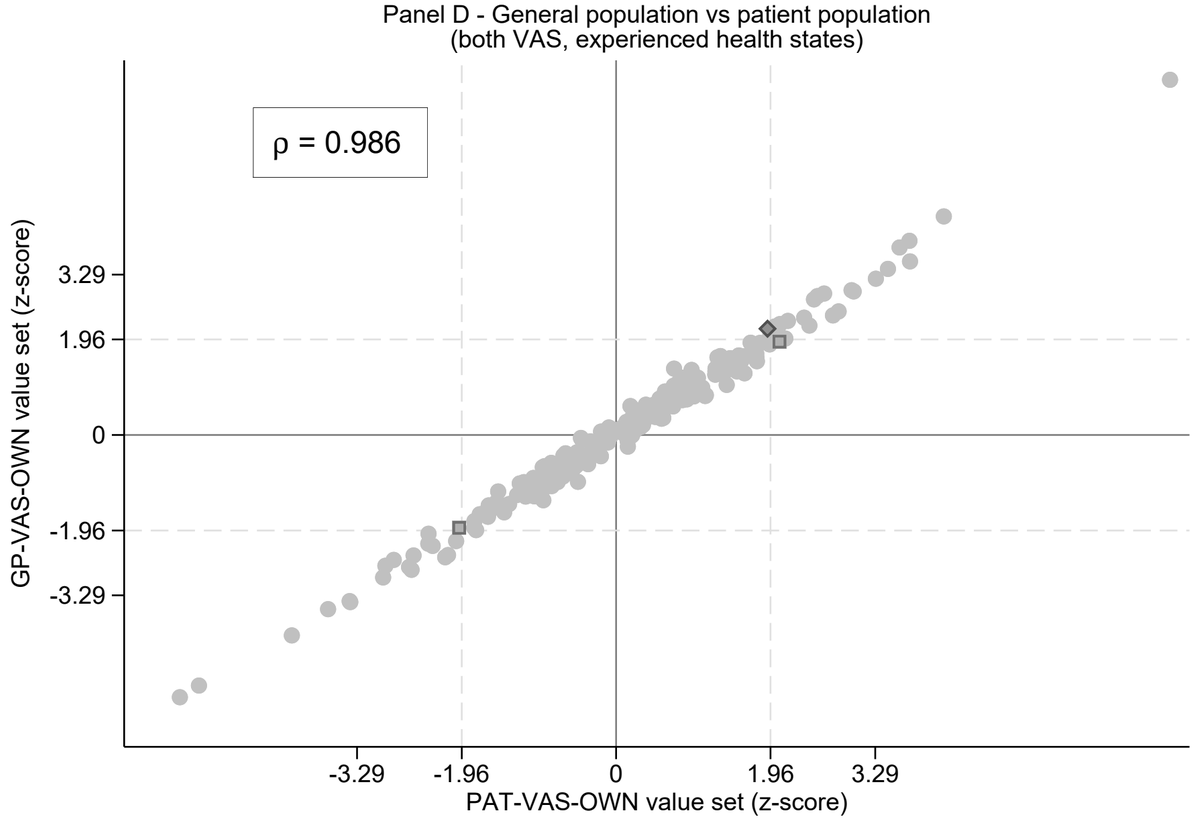
Figure 3: Number of statistically significant good/bad performers within patients' five closest hospitals under different value sets


EQ-5D health profile
33333
33321
33232
33212
32331
32313
32232
32223
32211
23321
23313
23232
22331
22323
22233
22222
22122
22121
22112
21323
21312
21232
21222
21133
21111
13332
13311
13212
12223
12222
12211
12121
12111
11312
11211
11133
11122
11121
11113
11112
-.6
-.4
-.2
0
.2
.4
.6
.8
Index score
GP-TTO-VAL
GP-VAS-VAL
GP-VAS-OWN
PAT-VAS-OWN

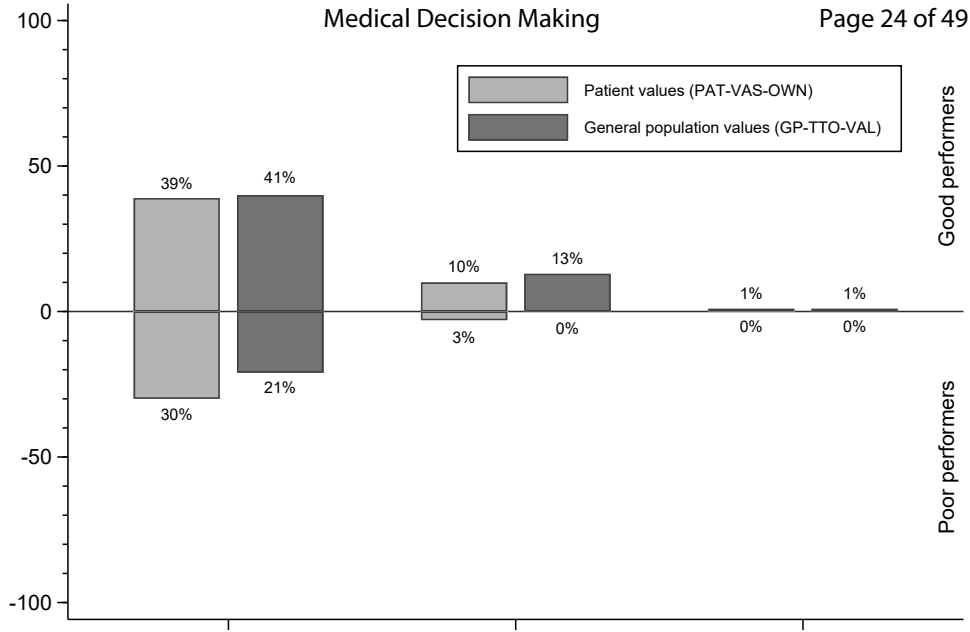








Legend: Dark grey symbols denote disagreement in performance outlier status under different value sets. Diamonds indicate outliers under value set 1 (on y-axis) but not value set 2 (on x-axis). Squares indicate outliers under value set 2 but not value set 1.

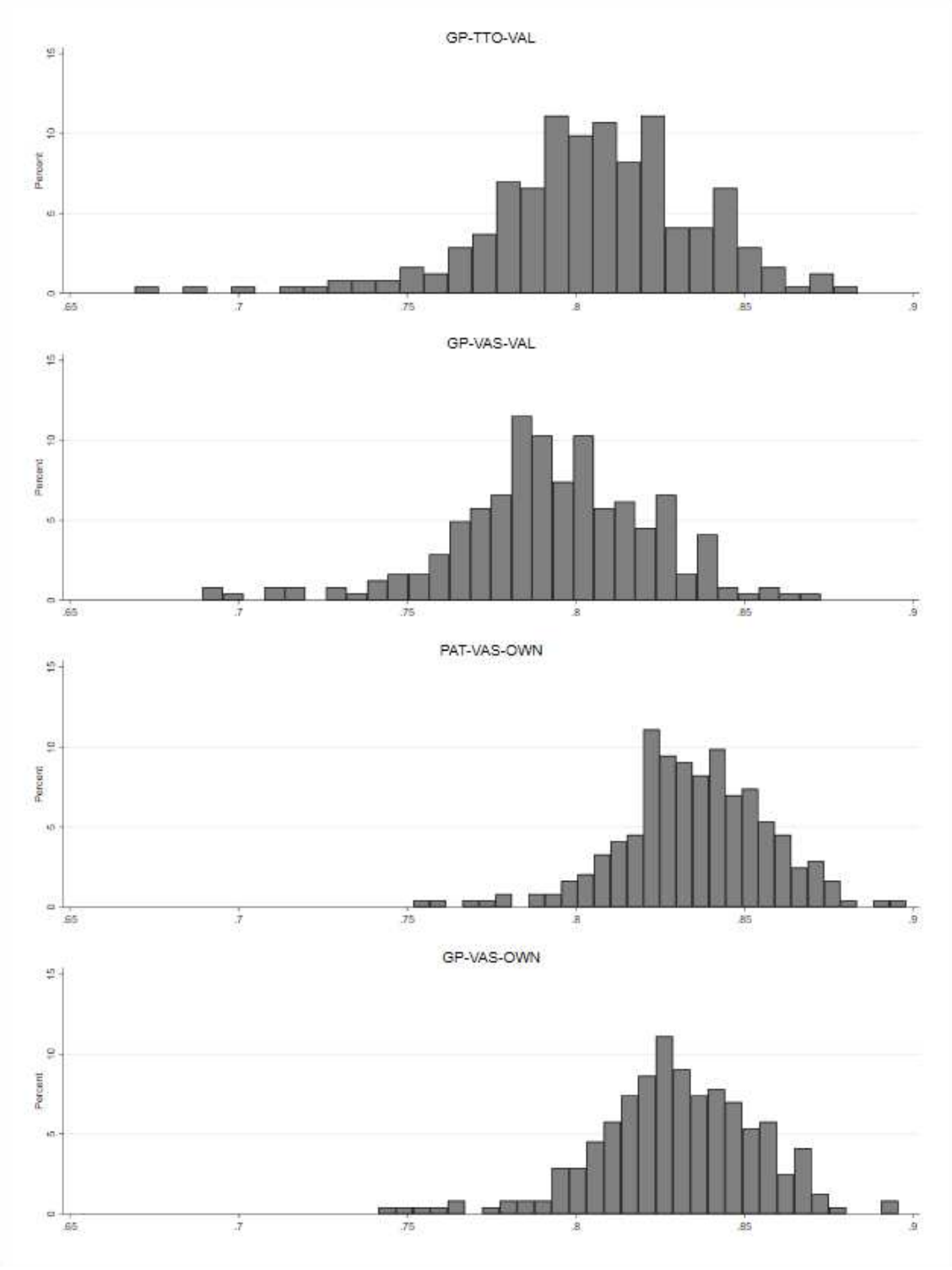


Percent of patients for which statement is true

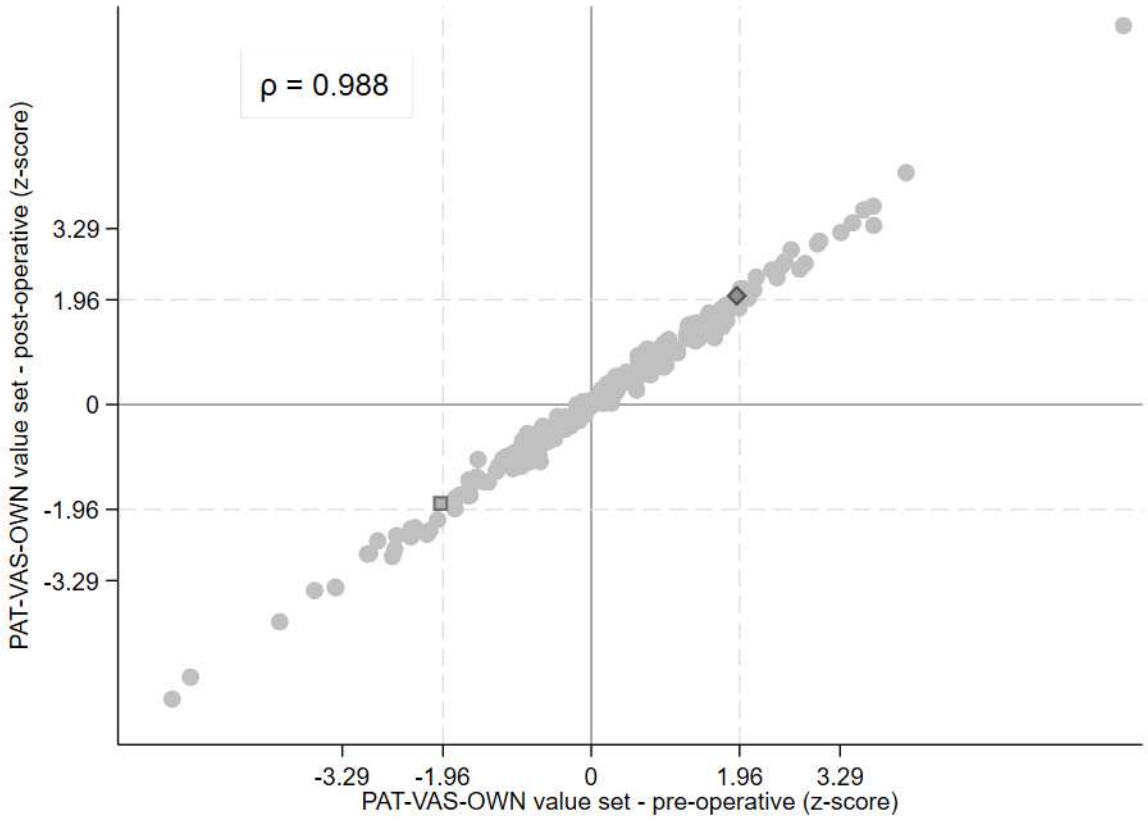
Number of hospitals identified as good/bad performers within the five closest hospital to the patient's home.

Online appendix to 'Using EQ-5D Data to Measure Hospital Performance: Are General Population Values Distorting Patients' Choices?'

Histograms of case-mix adjusted hospital scores under four different value sets



Relationship between hospital performance estimates under two PAT-VAS-OWN value sets based on pre- or post-operative PROM data



Title: Using EQ-5D Data to Measure Hospital Performance: Are General Population Values Distorting Patients' Choices?

Running head: EQ-5D value sets and hospital performance estimates

Authors:

1. Nils Gutacker, PhD[1], Centre for Health Economics, University of York, York, UK
2. Thomas Patton, PhD, Centre for Health Economics, University of York, York, UK
3. Koonal Shah, PhD, Office of Health Economics, London, UK
4. David Parkin, DPhil, Office of Health Economics, London, UK *and* Department of Economics, City, University of London, London, UK

Manuscript word count: 4,~~756~~758

Abstract word count: ~~273~~275

Acknowledgements: The authors thank Helen Dakin, John Brazier, Matthijs Versteegh, three anonymous referees and participants at the 2017 PROMs conference in Oxford and the EuroQol 34th Scientific Plenary Meeting (Barcelona, 2017) for useful comments and suggestions. Financial support for this study was provided entirely by a grant from the EuroQol Research Foundation. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. NG, KS and DP are members of the EuroQol Group. No ethical approval was required for analysis of anonymised, secondary data.

[1] Corresponding author. Nils Gutacker, Centre for Health Economics, University of York, Heslington, YO10 5DD, UK. E-mail: nils.gutacker@york.ac.uk

Abstract

Background: The English NHS publishes hospital performance indicators based on average post-operative EQ-5D index scores after hip replacement surgery to inform prospective patients' choices of hospital. Unidimensional index scores are derived from multidimensional health-related quality of life data using preference weights estimated from a sample of the UK general population. This raises normative concerns if general population preferences differ from those of the patients that are to be informed. This study explores how the source of valuation affects hospital performance estimates.

Methods: Four different value sets reflecting source of valuation (general population vs. patients), valuation technique (visual analogue scale (VAS) vs. time trade-off (TTO)) and experience with health states (currently experienced vs experimentally estimated) were used to derive and compare performance estimates for 243 hospitals. Two value sets were newly estimated from EQ-5D-3L data on 122,921 hip replacement patients and 3,381 members of the UK general public. Changes in hospital ranking (nationally~~; amongst patients' five closest hospitals~~) and performance outlier status (nationally; amongst patients' five closest hospitals) were compared across valuations.

Results: National rankings are stable under different valuations (~~Spearman~~ rank correlations > 0.~~93~~92)~~ but differ substantially at the local level~~. Twenty-three (9.5%) hospitals change outlier status when using patient VAS valuations instead of general population TTO valuations, the current approach. Outlier status also changes substantially at local level. This is explained ~~nearly entirely~~mostly by the valuation technique, not the source of valuations or experience with the health states.

Limitations: No patient TTO valuations were available. Effect of value set characteristics could only be established through indirect comparisons.

Conclusion: Different value sets may lead to prospective patients' choosing different hospitals. Normative concerns about the use of general population valuations are not supported by empirical evidence based on VAS valuations.

1 Introduction

Patients in the English National Health Service (NHS) have the right to choose among all qualified hospital providers for treatments that are deemed clinically appropriate and are publicly funded. To inform *"patients […] exercising choice"* (p.6)[1] about the quality of care they are likely to receive, the English NHS routinely collects multidimensional health-related quality of life (HRQoL) data from patients before and after undergoing planned hip and knee replacement surgical as part of the national patient-reported outcome measures (PROMs) programme. These data are then used to benchmark hospitals and calculate performance indicators in the form of case-mix adjusted average post-operative HRQoL, expressed as unidimensional composite scores, which are made publicly available on a regular basis.[2, 3]

A normative question, and the focus of this paper, is how to aggregate the multidimensional HRQoL data into unidimensional (single number) scores for the purpose of hospital performance assessment and public reporting. The PROMs programme collects HRQoL data using a generic health measurement instrument, the EQ-5D-3L[4], which comprises both a direct and indirect measure of a patient's health state. The direct measure, the EQ VAS, asks patients to provide a summary assessment of their HRQoL by marking a position on a visual analogue scale (VAS) ranging from 0 to 100, where the endpoints reflect the best and worst health states imaginable. The indirect measure uses the EQ-5D-3L descriptive system, where patients are asked to describe their current health status according to five dimensions of health (mobility, self-care, usual activities, pain & discomfort and anxiety & depression), each of which can be assigned one of three severity levels (essentially no, some or extreme problems). The resulting health profile data are aggregated into unidimensional composite ('index') scores using preference estimates of the UK general population[5], rather than of those prospective patients the PROMs programme seeks to inform. Previous research has shown many cases where preference estimates derived from specific patient populations differ systematically from those derived from the general population[6-10] although some studies find no differences [11, 12]. The current practice therefore raises normative concerns and could be inconsistent with the notion of patient sovereignty if it leads to a mismatch between the decisions patients make based on official published data, and those they would have made had the information reflected their own preferences more closely.

Ideally, the reported hospital performance should reflect prospective patients' individual preferences over relevant health states. However, the elicitation of personal preference functions is a complex and time-consuming task[13] and has therefore not (yet) found widespread adoption in the public reporting of hospital performance. Furthermore, it would imply the need to re-calculate public reports for each prospective patient based on their individual preferences, ruling out static performance reports (e.g. rankings published in newspapers) that are common currently. A pragmatic solution, that avoids both issues, is to develop a value set based on preferences elicited from a sample of patients. Such value sets are likely to reflect the preferences of prospective patients more closely than a general population value set since they are obtained from a sample of individuals with a similar age-sex structure, clinical condition, adaptation to their condition, and expectations of future health. At the same time, it would enable the calculation of EQ-5D index scores and, hence, unidimensional hospital performance indicators that could be presented alongside detailed dimension-by-dimension estimates[14] if desired.

In this paper we test whether the use of patient or general population valuations generates different hospital performance estimates for hip replacement surgery in the English NHS. We are not aware of

a UK-based patient value set that mirrors the currently used general population value set in terms of two other important aspects, namely respondents' experience of the health state to be valued as well as the valuation technique employed. This precludes a direct test of the effect of the source of valuation on hospital performance estimates. Instead, we compare hospital performance estimates generated under ~~a number~~four ~~of~~ published and newly-estimated value sets, out of eight possible combinations of these value set attributes. This allows us to vary one aspect at a time, holding the other two constant. The results of this indirect comparison help to demonstrate the practical implications of the normative argument about the source of health state valuations in the context of informing prospective patients about where to have surgery.

2 Valuation of health states

Amongst the desirable properties of a measure of the value of health is that it should unambiguously indicate whether a given health state, as defined by a multidimensional HRQoL profile, is better than, worse than, or equivalent to another health state. This property is most usually achieved by aggregating HRQoL data into a single number that represents the value of a health state by means of a set of preference weights. By convention, the value of a health state lies on a scale where 1 represents health which as good as possible, and 0 represents health that is either as poor as possible or is equivalent to being 'dead'. The latter allows for health states 'worse than dead' with values below 0.

Any attempt to value health in this way requires consideration of the following questions: (1) what is being valued; (2) whose values are being sought; and (3) what technique is being used to obtain the values? These are each briefly summarised below with interested readers being referred to detailed discussions elsewhere.[15-17]

2.1 What is being valued

Health state valuations are obtained as part of elicitation tasks. In these, participants may be asked to value their own health, as experienced either currently or in the past, or a set of health states that they may not be currently experiencing. For the latter, they are usually asked to value a stylised description of health, which may take the form of a health state profile comprising a series of dimensions and severity levels defined by the descriptive system of a PROM instrument, such as the EQ-5D. Such profiles are often described as 'hypothetical', but this is misleading because they are intended to reflect real health states and therefore plausible ways in which someone might self-report their health using the instrument. Since in most cases respondents will neither be experiencing or ever have experienced a health state described in the profile, they would need to imagine living in that health state in order to evaluate it. We can therefore regard these as their estimate of how they would value the health state if they were experiencing it.

2.2 Whose values are being sought

Health state valuations can be obtained from selected subgroups, such as patients with a given medical condition, or a sample of the general population.[16, 18, 19] Both approaches have merit, although advocates tend to argue their case on different grounds. Those in favour of using patient valuations typically point out that patients have first-hand experience of health states and therefore do not need to imagine the impact of an unfamiliar health state on their HRQoL.[20, 21] A common finding in the published literature - that valuations derived from specific patient populations tend to be higher than those elicited from the general population - has been attributed to patients adapting to their impaired health state over time and/or providing a more accurate assessment of the health

state based on their lived experience.[6, 7, 21] Conversely, proponents of general population valuations typically argue their case not on the grounds of validity but based on the intended use of such valuations to inform resource allocation decision in collectively funded health services, where decisions should reflect the preferences of the general population paying into the system.[18]

It is important to note that what is being valued and by whom are two separate issues. Patients may be asked to value health states that can occur as a result of their medical condition and which they may be able to imagine living in, but which they have not (yet) experienced themselves. Equally, the general population can be asked to value their currently experienced health state.[22]

2.3 What elicitation technique is being used

There are a number of techniques for valuing health states such as VAS and time trade-off (TTO).[23] The VAS involves rating the health state on a scale with imposed interval properties and well-defined endpoints, conventionally 0 and 100 (which in the EQ VAS represent worst and best imaginable health, respectively). TTO involves making a series of choices between living for a fixed amount of time in the profile under evaluation and a shorter, variable amount of time in full health, where the point at which respondents are indifferent is used to infer valuations. TTO has become the method most often recommended for the generation of values. The two methods have different assumptions underpinning them and are subject to different types of framing effects, for example VAS valuations are known to be subject to end-of-scale aversion [24] whereas respondents' time preference can have an effect on TTO valuations [25, 26]. VAS exercises are widely considered to be relatively simple and feasible to complete.[27] Previous research has shown that VAS and TTO yield different results.[28]

3 Methods

3.1 Data

We analyse EQ-5D-3L data from two independent samples. The first consists of 272,445 NHS-funded total hip replacement (THR) patients aged 15 years or over who had primary surgery in public or private hospitals in England between April 2012 and March 2016, collected as part of the English national PROMs programme[1]. Patients completed a paper questionnaire shortly before and six months after having surgery, containing the EQ-5D-3L, a condition-specific measure (the Oxford Hip Score) and other questions about their condition and treatment. The pre-operative questionnaire was administered by hospital staff at admission or the last outpatient appointment preceding admission and forwarded to a central data processor. The post-operative questionnaire was mailed directly to the patient's home address. Returned questionnaires were linked to administrative hospital records from the Hospital Episode Statistics (HES) database through a probabilistic matching algorithm. HES provides information on patient's age, place of residence, provider of care, and whether the surgery was a revision of a previous THR. Further details about the PROM data collection procedure are provided elsewhere.[29, 30] We excluded patients for whom pre- or post-operative responses were missing, either in part or completely, or where questionnaires could not be linked to HES. The sample used to estimate the patient value set in this study included 122,921 patients, which corresponds to 45.1% of all THR patients that were eligible to participate in the PROMs survey. Excluded patients were on average slightly younger and more likely to be female (Appendix Table 1). The linked HES-PROMs dataset was provided by NHS Digital.

The second sample consists of 3,381 randomly selected members of the UK general public that took part in the Measurement and Valuation of Health (MVH) study.[31] Each of the participants were

asked as part of face-to-face interviews to rate their own health status using the EQ-5D-3L questionnaire and to value 8 of 42 stylised health states using TTO[32] and VAS. The valuation data were used to derive a TTO based value set known as the MVH-A1 [5], but which we label the GP-TTO-VAL, and a VAS based value set known as the MVH-A3, but which we label the GP-VAS-VAL (Table 1).[31] The former is used in the official calculation of the hospital performance estimates reported to the public. Both value sets are anchored at 1 (full health) and 0 (dead), with scores below 0 indicating states considered worse than being dead. The MVH dataset was provided by the UK Data Services.

3.2 Estimation of experience-based value sets

A patient, current health VAS value set, which we label the PAT-VAS-OWN, was derived from the national PROMs dataset by regressing patient-reported EQ VAS scores on variables representing the levels within each dimension of the EQ-5D descriptive system, using Ordinary Least Squares. The regression model underpinning the MHV value sets include dummy variables for the main effects, a constant term reflecting any deviation from full health, and an N3 term indicating extreme problems (level 3) on any dimension.[5] To ensure comparability with these, we used the same specification. We also estimated more saturated models allowing for pairwise interactions between dimensions at level 2 and 3, but found these added little to overall fit (results available on request).

The PAT-VAS-OWN value set was estimated on data for the period April 2012 to March 2015, leaving one year of data to assess the impact of the value set on hospital rankings (see Section 2.4). It has been observed that patients' valuations of the same description of their health state may change from pre- to post-surgery, which may lead to inconsistencies when estimating patient-based value sets.[33] We focus our analysis on pre-operative survey responses since these are more likely to reflect patients' preferences at the point in time when a choice is to be made.

We also estimated a general population, current health VAS value set, which we label the GP-VAS-OWN, using the MVH study participants' EQ VAS and self-classifier responses and the same modelling structure as for the PAT-VAS-OWN value set.

Table 1 summarises the characteristics of the four value sets that we compared.

All standard errors are robust to heteroscedasticity and, in the case of the PAT-VAS-OWN value set, are clustered at hospital level. All computations were performed in Stata 14 (StataCorp LP, College Station, TX).

3.3 Deriving hospital performance estimates

Hospital performance assessment aims to identify the systematic contribution that providers make to their patients' health outcomes.[34] To allow for fair comparisons these assessments need to adjust for differences in hospital case-mix and sampling uncertainty.

Our analysis followed the published adjustment methodology of NHS England[35], in which the case-mix adjusted performance θ_j of hospital $j = 1,\ldots,J$ is estimated as

$$\hat{\theta}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}(O_{ij} - \hat{E}_{ij})$$

where O_{ij} is the observed post-operative index score for patient $i = 1,\ldots,n_j$ and E_{ij} is the expected post-operative index score for the same patient given their observable characteristics. ~~The case-mix adjusted performance estimates θ_j are themselves expressed as index scores.~~

The expected post-operative index score is based on the official case-mix adjustment methodology developed by NHS England[35]. The adjustment takes account of age, gender, ethnicity, living arrangements, the income deprivation profile of the patients' local small areas of residence (Lower-Layer Super Output Area (LSOA)) as approximated by the 2010 Index of Deprivation[36], main diagnosis and comorbid conditions, whether patients lived alone, whether they required assistance when filling in the PROMs questionnaire or considered themselves to be disabled, the duration of symptoms, as well as their pre-operative EQ-5D index score. We estimate the case-mix adjustment model separately for each of the four value sets using data from April 2012 to March 2014.

To account for sampling uncertainty in performance scores we follow standard practice[37-39] in the NHS and calculated z-score statistics for each hospital as

$$Z_j = abs\left(\frac{\hat{\theta}_j}{SE(\hat{\theta}_j)}\right)$$

where $Z_j > 1.96$ indicates statistically significant divergent performance from the national average at the 5% level. Hospitals with $Z_j > 1.96$ were deemed to perform *well* if $\hat{\theta}_j > 0$ and *poorly* otherwise. Performance estimates that were not statistically significantly different from the national average were deemed *average*. This approach is consistent with the simplified pictorial display used to communicate performance information (green, blue and red buttons to denote good, average and poor performance) by NHS choices[2] and other hospital comparison websites[3].

3.4 Assessing the impact of different EQ-5D value sets on hospital performance estimates

We assessed the impact of different value sets on hospital performance estimates for the period between April 2015 and March 2016 through a series of head-to-head comparisons. For each hospital, we compared their performance status (i.e. whether they were judged to perform well, poorly, or average) under different value sets and quantified discrepancies. The strength of association between hospital performance ~~estimates~~ rankings generated with different value sets ~~is~~ was measured using ~~Pearson's~~ Spearman's rank correlation coefficient ρ.

One motivation for considering patient valuations in assessing hospital performance is the desire to provide prospective patients with information that will inform their choice of hospital. Yet, most patients are unwilling to travel far for healthcare treatment[40-42], with a recent study[43] suggesting that over 92% of THR patients in the English NHS chose to attend one of their five closest hospitals in the period 2010 to 2012. We therefore also explored the impact of value sets at the local level; for each patient, we assessed how many of their five closest hospitals would be flagged as performing well or poorly under the different value sets~~examined whether the performance ranking for the five closest hospitals changed when different value sets were used to derive health state values~~. This 'choice set' was determined by the straight-line distance between the centroid of the patient's LSOA of residence and the hospitals' postcodes.[43] ~~Additionally, we assessed how many of these hospitals would be flagged as performing well or poorly under the different value sets. This approach is consistent with the simplified pictorial display used to communicate performance information (green, blue and red buttons to denote good, average and poor performance) by NHS choices[2] and other hospital comparison websites[3].~~

4 Results

4.1 Descriptive statistics

Table 2 reports descriptive statistics of the data samples. Patients in the national PROMs programme sample were, on average, 68 years old and 58.7% were female. Most patients had suffered from joint-related symptoms for 1 to 5 years prior to surgery. The average improvement in HRQoL six months after surgery was equivalent to an increase of 0.43 value points (from 0.37 to 0.80) (MVH-A1 value set), and patients' overall assessment of their health as measured by the EQ VAS increased by 12 points (from 65 to 77). Patients described their pre-operative HRQoL using 148 of the 243 possible EQ-5D-3L health states. The relative frequency of these health states was consistent with the severity of the conditions that require major joint replacement. Over 46% of patients reported extreme limitation (i.e. level 3 problems) on at least one HRQoL dimension before surgery, and >2% reported extreme limitations on three or more dimensions.

Unsurprisingly, MVH study participants reported better health on average than the patient sample, both pre- and post-surgery. They were, on average, significantly younger (mean age = 47.9 years) than the patient population but showed a similar sex split (56.7% female). Participants described their health using 77 of the 243 EQ-5D-3L health states, with 4.8% of participants having at least one extreme limitation on any of the five health dimensions. The average VAS score was 82.5 and the average EQ-5D value based on the GP-TTO-VAL value set was 0.86.

4.2 Value sets

Table 3 reports the estimated PAT-VAS-OWN and GP-VAS-OWN value sets alongside the published GP-TTO-VAL and GP-VAS-VAL value sets. Coefficient estimates represent decrements associated with some or extreme limitations on a given health dimension. The constant and the N3 term reflect global decrements that are applied in the presence of *any* limitations on *any* health dimension and at least one extreme limitation on any health dimension, respectively.

Figure 1 shows the values generated by the different value sets for the 42 stylised health states valued in the MVH study.

Both PAT-VAS-OWN and GP-VAS-OWN value sets were found to be internally consistent, i.e. more severe limitations are associated with larger decrements for each dimension. Patients assign approximately equal or smaller decrements to health problems on a given *dimension* than the general public, but they attach a larger global decrement to the presence of *any* health problems as reflected in the coefficient on the constant term. Differences are more pronounced on level 3 decrements than level 2 decrements, thus generating a wider spread of index scores across the four value sets for health states for which respondents reported at least one extreme problem. These results are consistent with previous evidence from other patient populations.[7, 44] It should be noted that because of the smaller sample size, the GP-VAS-OWN data has sparse observations in some of the levels within dimensions, particularly Mobility Level 3, which means that the coefficient estimates have very large standard errors.

Table 4 reports descriptive statistics of the pre- and post-operative index scores reported at patient level (mean, SD) as well as the range of hospital average scores calculated using the four value sets. Differences in average index scores are more pronounced prior to surgery than afterwards, which reflects the low number of patients reporting any extreme problems after surgery. The two value sets based on direct valuations of own, currently experienced health (GP-VAS-OWN, PAT-VAS-OWN)

generate, on average, higher index scores as well as a smaller spread of hospital average index scores that are relevant for performance assessment. Histograms of case-mix adjusted hospital scores are presented in the online appendix.

4.3 Impact on judgements about hospital performance

Figure 2 presents scatter plots of hospital z-scores derived under different EQ-5D value sets. Each scatter point represents one hospital, with dashed lines indicating the lower and upper boundaries at which performance estimates are deemed to be statistically significantly different from the national average. Performance estimates that would lead to differential judgement under the two value sets being compared are highlighted as diamonds (significant under the first but not the second value set) or squares (vice versa).

The GP-TTO-VAL and PAT-VAS-OWN value sets generate performance estimates that are highly correlated (ρ = 0.~~93~~92) (Figure 2, Panel A). Despite this, the change in value set has a non-negligible impact on how individual hospitals are deemed to perform, with patient valuations leading to changes in outlier status for 23 hospitals in total (9.5% of 243), of which 6 (2.5%) are no longer identified as performing poorly, 10 (4.1%) are no longer identified as performing well, and seven different hospitals now appear to perform well (2.9%). At the local level, 1% fewer patients (44% vs 45% of N=65,278) receiving care between April 2015 and March 2016 would have found at least one well performing hospital within their five closest hospitals if performance estimates had been derived using the PAT-VAS-OWN value set rather than the GP-TTO-VAL (Figure 3). In contrast, patients would have been 10% more likely (34% vs 24%) to find at least one local hospital deemed to perform poorly if performance estimates had been derived using the PAT-VAS-OWN value set. Overall, at least one performance assessment for their five closest hospitals would have been different for 8.6% of patients receiving care between April 2015 and March 2016.

To further explore the reasons for this divergence, we compared hospital performance estimates derived varying one value set design characteristic (i.e. source of valuation, valuation technique, or experience with health state) while holding the others constant (Figure 2, Panels B-D). The results of this marginal analysis suggest that neither the source of valuation nor the level of experience with a health state drive the observed differences in hospital performance classifications. Instead, these differences can be explained nearly entirely by the choice of valuation technique employed, with Panel B showing many more changes in outlier status than Panel C and D.

5 Discussion

There is a strong normative rationale for using patient values to aggregate multidimensional HRQoL instruments when developing hospital performance indicators to inform prospective patients' choices of hospital. However, the standard practice in the English NHS has been to publish hospital performance indicators based on EQ-5D scores aggregated using general public values. The present study explores whether this practice may be distorting patients' choice of hospital for hip replacement surgery given that there is some evidence of discrepancies between patient and general public values. We find a larger number of hospitals are deemed to perform poorly when a patient VAS tariff (PAT-VAS-OWN) is used compared to when the UK general population TTO tariff (GP-TTO-VAL) is used. Conversely, we find only slightly fewer hospitals are deemed to perform well when using the PAT-VAS-OWN instead of the GP-TTO-VAL value set. The choice of value set therefore appears to be more important for patients seeking to avoid poorly performing hospitals. Moreover, we find that the GP-TTO-VAL tariff overvalues the relative performance of hospitals that

deliver improvements in pain/discomfort and mobility compared to the PAT-VAS-OWN tariff whilst undervaluing those that perform relatively well at addressing anxiety/depression problems. Importantly, these differences appear to be driven almost entirely by the difference in the health state valuation technique employed (TTO vs VAS) rather than the source of valuations. Therefore, our results provide little empirical support for a change in reporting practice in the English PROMs programme because of normative concerns about the source of valuations.

In recent years, there has been considerable interest in the use of values that reflect individuals' own health, rather than their estimated valuations of stylised health states, to derive value sets.[22, 45] The purported rationale for using 'experience-based' values is that they avoid some of the focusing effects that can occur in the valuation of stylised health states.[20] Furthermore, any need to reflect the preferences of the tax-paying general population, which mainly arises in the context of economic evaluation of new health technologies for use in publicly-funded health systems, can be addressed by using a population survey.[22] One concern with this approach is that the data collected for the purposes of developing an experience-based value set may only contain a limited range of responses to the health state descriptive system. Our study provides further evidence to demonstrate the feasibility of developing an experience-based value set from large-scale, routinely collected PROM surveys. Patients in the hip replacement sample report their HRQoL according to 148 of the 243 possible EQ-5D-3L health states; covering a broad range of the instrument's spectrum. By design, these are also the most commonly encountered health states in this population, limiting the need to extrapolate beyond the set of valued health states in most applications.

While not the focus of our study, our findings also provide additional context to the debate about the comparability of EQ-5D-3L value sets developed in different countries. A study by Nemes and colleagues developed an experience-based VAS value set for the EQ-5D-3L using data from patients undergoing elective total hip replacement in Sweden.[46] The valuations of health dimensions in the Swedish study and those in our study are similar in that the most important dimension – both in terms of the decrements associated with the level 2 and 3 responses – is anxiety/depression (see Appendix Table 2 for estimates). Aside from this similarity, the relative importance of the various health dimensions differ systematically for the two value sets. This casts doubt on the ability to pool experienced-based value sets across countries as recently suggested for TTO value sets based on valuations of health states derived from valuation studies.[47]

There are a number of limitations to our analysis and proposed approach. First, a single patient group value set still requires aggregating valuations over a large number of patients with potentially heterogeneous preferences. While it is reasonable to assume that the mismatch between the average patient value set and individual patients' preferences is smaller than the mismatch with average general population preferences, there may be scope for further refinement. Some existing work has explored how health state valuations vary with observable characteristics of the respondent and this line of inquiry ought to be expanded.[48] Secondly, the relationship between direct valuations of health states as reflected in EQ VAS scores and patients' EQ-5D-3L health profiles has been found to change from before to after surgery.[33] The reason for this discrepancy remains unclear. We have chosen to estimate patient valuations from their pre-operative data since this reflects their ex-ante valuations at the time of their decisions. However, one may also argue that post-operative valuations are appropriate as they reflect patients' preferences over different outcomes once they have started to experience the benefits of treatment. This distinction is not the focus of this paper, although we note that it appears to have little effect on hospital performance estimates, which are highly correlated under both value sets (rho>0.99) (see Appendix Table 3 for post-operative PAT-VAS value set and the online appendix for hospital performance scatter plots).

Thirdly, while we find that the source of valuation is not a major driver of hospital performance estimates when valuing health states using VAS, we cannot generalise this statement to other valuation techniques such as the TTO valuations currently used in the NHS. To test this we would require TTO data from a sample of hip replacement patients, which we do not currently have access to. Fourthly, the generalizability of the findings in our study is limited to the medical condition and the decision problem under consideration. Finally, the limited amount of provider variation in both intake and health gain following THR surgery may limit the role that valuations play in determining hospital performance estimates.[49] As routine PROM collection becomes more prevalent, this hypothesis will become testable.

In conclusion, the choice of value set to aggregate EQ-5D-3L health profiles in the context of the English PROMs programme may have real implications for patients choosing hospitals for their THR surgery. This is particularly relevant when choices are based on simple heuristics, e.g. selection based on dichotomized performance status rather than index scores. However, this divergence does not appear to be driven by the source of health state valuations, a normative concern, but rather by the valuation technique employed, a technical matter.

Appendix

Appendix Table 1: Comparison of included and excluded patients in PROMs sample

	Hip replacement patient sample			
	Excluded		Included	
Patient age (mean, sd)	67.96	12.10	68.26	10.32
Patient gender (n, %)				
Female	90,887	61%	72,095	59%
Male	58,335	39%	50,826	41%
*Financial year of treatment (n, %) **				
2012/13	35,259	24%	28,270	23%
2013/14	35,275	24%	33,148	27%
2014/15	39,064	26%	31,591	26%
2015/16	39,624	27%	29,912	24%

** Financial years run from April to March.*

Appendix Table 2: Experienced-based VAS value sets for total hip replacement patients in England and Sweden

	England		Sweden	
	Est	SE	Est	SE
Full health	1.000		0.745	
Mobility, level 2	-0.056	0.003	-0.060	
Mobility, level 3	-0.119	0.012	-0.098	
Self-care, level 2	-0.055	0.002		
Self-care, level 3	-0.116	0.008		
Self-care, level 2 or 3			-0.038	
Usual activities, level 2	-0.029	0.003	-0.053	
Usual activities, level 3	-0.086	0.004	-0.110	
Pain/Discomfort, level 2	-0.050	0.008	-0.025	
Pain/Discomfort, level 3	-0.130	0.009	-0.124	
Anxiety/Depression, level 2	-0.088	0.002	-0.078	
Anxiety/Depression, level 3	-0.181	0.004	-0.161	
N3	0.011	0.003		
Any deviation from full health	-0.182	0.008		

Notes: Swedish values are taken from Nemes et al. (2015). Signs on coefficient estimates for England have been reversed to be compatible with Swedish values.

Appendix Table 3: PAT-VAS-OWN value sets calculated from PROMs data collected before or six months after surgery

	Before surgery		After surgery	
EQ-5D dimension	Est	SE	Est	SE
Mobility, level 2	0.056	0.003	0.075	0.001
Mobility, level 3	0.119	0.012	0.177	0.024
Self-care, level 2	0.055	0.002	0.059	0.002
Self-care, level 3	0.116	0.008	0.086	0.009
Usual activities, level 2	0.029	0.003	0.050	0.001
Usual activities, level 3	0.086	0.004	0.119	0.006
Pain/Discomfort, level 2	0.050	0.008	0.032	0.001
Pain/Discomfort, level 3	0.130	0.009	0.121	0.006
Anxiety/Depression, level 2	0.088	0.002	0.083	0.001
Anxiety/Depression, level 3	0.181	0.004	0.167	0.007
N3	-0.011	0.003	0.034	0.006
Constant	0.182	0.008	0.131	0.001

References

[1] Department of Health. Guidance on the routine collection of Patient Reported Outcome Measures (PROMs). London: The Stationery Office 2008.
[2] NHS. Find services. 2019 10/12/2019]Available from: www.nhs.uk/service-search
[3] National Joint Registry. NJR Surgeon and Hospital Profile for hip, knee, ankle, elbow and shoulder joint replacement surgery. 2019 Available from: https://surgeonprofile.njrcentre.org.uk/
[4] Brooks R. EuroQol: the current state of play. *Health Policy*. 1996; 37(1):53-72.
[5] Dolan P. Modeling Valuations for EuroQol Health States. *Medical Care*. 1997; 35(11):1095-108.
[6] De Wit GA, Busschbach JJV, De Charro FT. Sensitivity and perspective in the valuation of health status: whose values count? *Health Economics*. 2000; 9(2):109-26.
[7] Mann R, Brazier J, Tsuchiya A. A comparison of patient and general population weightings of EQ-5D dimensions. *Health Economics*. 2009; 18(3):363-72.
[8] Peeters Y, Stiggelbout AM. Health State Valuations of Patients and the General Public Analytically Compared: A Meta-Analytical Comparison of Patient and Population Health State Utilities. *Value in Health*. 2010; 13(2):306-9.
[9] Goodwin E, Green C, Hawton A. What Difference Does It Make? A Comparison of Health State Preferences Elicited From the General Population and From People With Multiple Sclerosis. *Value in Health*. 2019.
[10] Rand-Hendriksen K, Augestad L, Kristiansen I, Stavem K. Comparison of hypothetical and experienced EQ-5D valuations: relative weights of the five dimensions. *Quality of Life Research*. 2012; 21:1005-12.
[11] Pickard AS, Tawk R, Shaw JW. The effect of chronic conditions on stated preferences for health. *The European Journal of Health Economics*. 2013; 14(4):697-702.
[12] Gandhi M, Tan RS, Ng R, Choo SP, Chia WK, Toh CK, et al. Comparison of health state values derived from patients and individuals from the general population. *Quality of Life Research*. 2017; 26(12):3353-63.
[13] Devlin NJ, Shah KK, Mulhern BJ, Pantiri K, van Hout B. A new method for valuing health: directly eliciting personal utility functions. *European Journal of Health Economics*. 2019; 20(2):257-70.
[14] Gutacker N, Bojke C, Daidone S, Devlin N, Street A. Hospital Variation in Patient-Reported Outcomes at the Level of EQ-5D Dimensions: Evidence from England. *Medical Decision Making*. 2013; 33(6):804-18.
[15] Versteegh MM, Brouwer WBF. Patient and general public preferences for health states: A call to reconsider current guidelines. *Social Science & Medicine*. 2016; 165:66-74.
[16] Brazier J, Akehurst R, Brennan A, Dolan P, Claxton K, McCabe C, et al. Should Patients Have a Greater Role in Valuing Health States? *Applied Health Economics and Health Policy*. 2005; 4(4):201-8.
[17] Brazier J, Rowen D, Karimi M, Peasgood T, Tsuchiya A, Ratcliffe J. Experience-based utility and own health state valuation for a health state classification system: why and how to do it. *European Journal of Health Economics*. 2018; 19(6):881-91.
[18] Gold M, Siegel J, Russell L, Weinstein M, eds. Cost-effectiveness in health and medicine. New York: Oxford University Press 1996.
[19] Dolan P. Whose Preferences Count? *Medical Decision Making*. 1999; 19(4):482-6.
[20] Dolan P, Kahneman D. Interpretations Of Utility And Their Implications For The Valuation Of Health. *Economic Journal*. 2008; 118(525):215-34.
[21] Menzel P. Utilities for Health States: Whom to Ask. In: Culyer AJ, ed. *Encyclopedia of Health Economics*. Amsterdam: Elsevier 2014:417-24.
[22] Burström K, Sun S, Gerdtham U-G, Henriksson M, Johannesson M, Levin L-Å, et al. Swedish experience-based value sets for EQ-5D health states. *Quality of Life Research*. 2014; 23(2):431-42.

[23] Brazier J, Ratcliffe J, Saloman J, Tsuchiya A. *Measuring and Valuing Health Benefits for Economic Evaluation*. 2 ed. Oxford: Oxford University Press 2016.
[24] Shmueli A. The visual analog rating scale of health-related quality of life: an examination of end-digit preferences. *Health and Quality of Life Outcomes*. 2005; 3(1):71.
[25] Attema AE, Brouwer WBF. On the (not so) constant proportional trade-off in TTO. *Qual Life Res*. 2010; 19(4):489-97.
[26] Dolan P, Stalmeier P. The validity of time trade-off values in calculating QALYs: constant proportional time trade-off versus the proportional heuristic. *Journal of Health Economics*. 2003; 22(3):445-58.
[27] Green C, Brazier J, Deverill M. Valuing health-related quality of life. A review of health state valuation techniques. *Pharmacoeconomics*. 2000; 17(2):151-65.
[28] Robinson A, Dolan P, Williams A. Valuing health status using VAS and TTO: what lies behind the numbers? *Social Science & Medicine*. 1997; 45(8):1289-97.
[29] Gutacker N, Street A, Gomes M, Bojke C. Should English healthcare providers be penalised for failing to collect patient-reported outcome measures (PROMs)? *Journal of the Royal Society of Medicine*. 2015; 108(8):304-16.
[30] NHS Digital. *Patient Reported Outcome Measures (PROMs) in England - A guide to PROMs methodology*2017.
[31] MVH Group. *Final Report on the Modelling of Valuation Tariffs*: Centre for Health Economics, University of York; 1995.
[32] Walker S, Sculpher M, Drummond M. The Methods of Cost-Effectiveness Analysis to Inform Decisions about the Use of Health Care Interventions and Programs. In: Glied S, Smith PC, eds. *The Oxford Handbook of Health Economics* Oxford: Oxford University Press 2012.
[33] Pickard AS, Hung Y-T, Lin F-J, Lee TA. Patient Experience-based Value Sets: Are They Stable? *Medical Care*. 2017; 55(11):979-84.
[34] Jacobs R, Smith PC, Street A. *Measuring Efficiency in Health Care*. Cambridge: Cambridge University Press 2006.
[35] NHS England. *Patient Reported Outcome Measures (PROMs) in England - Update to reporting and case-mix adjusting hip and knee procedure data. Annex 1: Coefficients for primary hip replacement models*2017.
[36] McLennan D, Barnes H, Noble M, Davis J, Garrett E, Dibben C. The English Indices of Deprivation 2010. In: Government DfCaL, ed. London2011.
[37] Department of Health. Patient Reported Outcome Measures (PROMS) in England: a methodology for identifying potential outliers. London: The Stationery Office 2011.
[38] Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Statistics in Medicine*. 2005; 24(8):1185-202.
[39] National Clinical Audit Advisory Group. Detection and management of outliers. In: Department of Health, ed. London: The Stationary Office 2011.
[40] Luft HS, Hunt S, Maerki S. The Volume-Outcome Relationship: Practice-Makes-Perfect or Selective-Referral Patterns? *Health Services Research*. 1987; 22:157-82.
[41] Propper C, Damiani M, Leckie G, Dixon J. Impact of patients' socioeconomic status on the distance travelled for hospital admission in the English National Health Service. *Journal of Health Services Research & Policy*. 2007; 12(3):153-9.
[42] Varkevisser M, van der Geest SA, Schut FT. Do patients choose hospitals with high quality ratings? Empirical evidence from the market for angioplasty in the Netherlands. *Journal of Health Economics*. 2012; 31(2):371-8.
[43] Gutacker N, Siciliani L, Moscelli G, Gravelle H. Choice of hospital: Which type of quality matters? *Journal of Health Economics*. 2016; 50:230-46.
[44] Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. *Quality of Life Research*. 2003; 12(6):599-607.

[45] Dolan P. NICE should value real experiences over hypothetical opinions. *Nature*. 2009; 462:35.
[46] Nemes S, Burström K, Zethraeus N, Eneqvist T, Garellick G, Rolfson O. Assessment of the Swedish EQ-5D experience-based value sets in a total hip replacement population. *Quality of Life Research*. 2015:1-8.
[47] Olsen JA, Lamu AN, Cairns J. In search of a common currency: A comparison of seven EQ-5D-5L value sets. *Health Economics*. 2018; 27(1):39-49.
[48] Craig BM, Reeve BB, Cella D, Hays RD, Pickard AS, Revicki DA. Demographic Differences in Health Preferences in the United States. *Medical Care*. 2014; 52(4):307-13.
[49] Street A, Gutacker N, Bojke C, Devlin N, Daidone S. Variations in outcome and costs among NHS providers for common surgical procedures: econometric analyses of routinely collected data. *Health Services and Delivery Research*. 2014; 2(1).

Tables and figure legends

Table 1: Overview of value set characteristics

Value Set	Source of valuation	Valuation technique	Experience of health states
GP-TTO-VAL (Dolan 1997)	General population	TTO	Stylised description
GP-VAS-VAL	General population	Valuation VAS	Stylised description
GP-VAS-OWN	General population	EQ VAS	Current health
PAT-VAS-OWN	Patients	EQ VAS	Current health

Notes: TTO = Time trade-off, VAS = Visual analogue scale.

Table 2: Descriptive statistics of PROMs and MVH samples

Variable	Hip replacement (PROMs) sample		General population (MVH) sample	
Patient age (mean, sd)	68.26	10.32	47.86	18.37
Patient gender (n, %)				
Female	72,095	58.7%	1,917	56.7%
Male	50,826	41.3%	1,464	43.3%
Symptom duration (n, %)				
<1 year	16,414	13.4%		
1-5 years	84,015	68.3%		
6-10 years	13,967	11.4%		
>10 years	7,700	6.3%		
Not reported	825	0.7%		
Pre-operative EQ-5D responses (mean, sd)				
EQ-5D index score (GP-TTO-VAL)	0.37	0.32	0.86	0.23
EQ VAS score	65.43	21.55	82.53	16.90
Post-operative EQ-5D responses (mean, sd)				
EQ-5D index score (GP-TTO-VAL)	0.80	0.24		
EQ VAS score	77.34	17.61		
Number of level 3 problems (pre- or post-operatively) (n, %)				
none	66,170	53.8%	3,172	93.8%
1	39,068	31.8%	161	4.8%
2	14,905	12.1%	40	1.2%
3	2,405	2.0%	8	0.2%
4	314	0.3%	0	0.0%
5	59	0.0%	0	0.0%
Sample size	122,921		3,381	

Table 3: Estimated EQ-5D health dimension decrements and standard errors

	GP-TTO-VAL		GP-VAS-VAL		GP-VAS-OWN		PAT-VAS-OWN	
EQ-5D dimension	Est	SE	Est	SE	Est	SE	Est	SE
Mobility, level 2	0.069	0.005	0.071	0.004	0.059	0.010	0.047	0.002
Mobility, level 3	0.314	0.007	0.182	0.005	0.152	0.084	0.117	0.011
Self-care, level 2	0.104	0.005	0.093	0.004	0.067	0.018	0.057	0.001
Self-care, level 3	0.214	0.007	0.145	0.005	0.080	0.097	0.104	0.007
Usual activities, level 2	0.036	0.006	0.031	0.004	0.082	0.011	0.042	0.002
Usual activities, level 3	0.094	0.007	0.081	0.005	0.139	0.034	0.097	0.003
Pain/Discomfort, level 2	0.012	0.005	0.084	0.004	0.065	0.006	0.047	0.006
Pain/Discomfort, level 3	0.386	0.006	0.171	0.004	0.100	0.034	0.119	0.007
Anxiety/Depression, level 2	0.071	0.071	0.063	0.004	0.072	0.007	0.085	0.001
Anxiety/Depression, level 3	0.236	0.006	0.124	0.004	0.151	0.034	0.173	0.003
N3	0.269	0.007	0.215	0.005	0.064	0.036	-0.020	0.003
Constant	0.081	0.008	0.159	0.004	0.104	0.002	0.121	0.005
Source of valuation	General population		General population		General population		Patients	
Valuation technique	TTO		Valuation VAS		EQ VAS		EQ VAS	
Experience of health states	Stylised description		Stylised description		Current health		Current health	

Table 4: Index scores at patient level (mean, SD) and range of scores at provider level under four value sets

Value set	Pre-operative			Post-operative (unadjusted)			Post-operative (case-mix adjusted)		
	Mean	SD	Range of hospital mean scores	Mean	SD	Range of hospital mean scores	Mean	SD	Range of hospital mean scores
GP-TTO-VAL	0.364	0.320	0.243 to 0.576	0.802	0.239	0.568 to 1	0.804	0.216	0.632 to 1
GP-VAS-VAL	0.441	0.202	0.227 to 0.571	0.789	0.216	0.599 to 1	0.791	0.195	0.593 to 0.998
GP-VAS-OWN	0.579	0.116	0.449 to 0.673	0.826	0.173	0.687 to 1	0.828	0.155	0.629 to 0.987
PAT-VAS-OWN	0.625	0.101	0.496 to 0.711	0.832	0.162	0.708 to 1	0.834	0.144	0.646 to 0.975

Figure 1: Selected health state valuations under different value sets

Figure 2: Relationship between hospital performance estimates under different value sets

Figure 3: Number of statistically significant good/bad performers within patients' five closest hospitals under different value sets

MDM-19-321-R1. Using EQ-5D Data to Measure Hospital Performance: Are General Population Values Distorting Patients' Choices?

We thank the reviewer for the helpful set comments and suggestions. A response to each point is provided. The comments from the reviewer are in **bold** (numbered for ease of reference) with our responses below.

Reviewer 1

R1) In abstract it first implies that rankings did not change overall, but changed on the local level. This is confusing- but in the paper itself, it seems there were no rankings on the local level, but rather a determination of outlier status. Abstract may need correction.

The reviewer is correct. We have amended the abstract from:

> Changes in hospital ranking (nationally; amongst patients' five closest hospitals) and performance outlier status were compared across valuations.

to

> Changes in hospital ranking (nationally) and performance outlier status (nationally; amongst patients' five closest hospitals) were compared across valuations.

R2) I still find the last paragraph of the Introduction confusing, especially as it emphasizes what is not done or possible to do. Could it be written more directly- e.g. include sentences such as: "In this paper we compute EQ-5D-3L scores for hip replacement patients in the English NHS based on valuations that differ in whether they were obtained from patients or the general population, based on stylized descriptions or own health, and based on VAS versus TTO. Out of 8 possible combinations of these factors, we have data on four (see Table 1), allowing comparison of general population and patients based on valuation of own health on VAS, comparison of own health versus stylized description for VAS based valuations in the general population and VAS based versus TTO valuations for stylized descriptions in the general population. We are not aware of a UK-based patient value set that mirrors the currently used general population value set to allow direct comparisons of VAS versus TTO valuations or own versus stylized health valuations within the patient group. However, available data allows us to test one aspect of valuations at a time while keeping the others constant."

We thank the reviewer for this suggestion. However, the suggested text does not match with the motivation for our paper, which is to compare the use of general population valuations to that of patient valuations and analyse the impact that this has on hospital performance measurement. The key point is that we do not have data with which to test this directly, so we use indirect comparisons, which are limited by the available value sets and data. The suggested text implies we are testing, with equal importance, the effect of valuation technique, experience with health state and source of valuations. Our paper does not have that broader motivation.

We have therefore not changed this paragraph substantially, but have edited the text to clarify that we are comparing four of out of eight possible value sets.

R3) It is mentioned that Standard errors are robust, taking into account hospital clustering for patients. However, SD (Table 4) do not take hospital clustering into account. To do so, one would need to report separate within and between hospital SD.

Table 4 does not report standard *errors*. The table contains descriptive statistics, including standard *deviations*, that show the distribution of index scores over hospitals. Our statement about standard errors is correct. We have therefore not altered the text or the data in the table.

R4) Not sure what this means (page 6): the case-mix adjusted performance estimates θj are themselves expressed as index scores.

We have removed this sentence. It does not add anything to the paper, so there would be no point in explaining it.

R5) Page 7: Pearson correlation is not "rank correlation". Which is intended "Pearson correlation" or "Spearman rank correlation"?

Thank you for pointing out this potentially misleading label. The correlations are, as the reviewer suggests, Spearman rank correlations. These are Pearson correlations applied to ranked data, but we accept that our description is ambiguous, so we have updated the text and Figure 2 accordingly.

R6) Not sure the information on change in HRQoL at the beginning of Results was anticipated in Methods.

The start of the results section provides descriptive statistics to familiarise the reader with the data, rather than to address a specific research question. We believe this is consistent with the MDM house style for this type of auxiliary information.