Addressing Missing Data in Patient-Reported Outcome Measures (PROMs): Implications for Comparing Provider Performance

Manuel Gomes, Nils Gutacker, Chris Bojke, Andrew Street

CHE Research Paper 101

Addressing missing data in patient-reported outcome measures (PROMs): implications for comparing provider performance

[1]Manuel Gomes
[2]Nils Gutacker
[2]Chris Bojke
[2]Andrew Street

[1]Department of Health Services Research & Policy, London School of Hygiene and Tropical Medicine, UK
[2]Centre for Health Economics, University of York, UK

Background to series

Acknowledgments

Funding

Disclaimer

Further copies

Centre for Health Economics
Alcuin College
University of York
York, UK
www.york.ac.uk/che

Abstract

Patient-reported outcome measures (PROMs) are now routinely collected in the English National Health Service (NHS) and used to compare and reward hospital performance within a high-powered pay-for-performance scheme. However, PROMs are prone to missing data. For example, hospitals often fail to administer the pre-operative questionnaire at hospital admission, or patients may refuse to participate or fail to return their post-operative questionnaire. A key concern with missing PROMs is that the individuals with complete information tend to be an unrepresentative sample of patients within each provider, and inferences based on the complete cases will be misleading. This study proposes a strategy for addressing missing data in the English PROMs survey using multiple imputation techniques, and investigates its impact on assessing provider performance. We find that inferences about relative provider performance are sensitive to the assumptions made about the reasons for the missing data.

Keywords: Missing data, Multiple Imputation, Patient-reported outcome measures, Provider performance, Missing not at Random

1.   Introduction

Non-response is a major concern in health surveys because individual non-respondents tend to be systematically different from those providing complete data (Cohen and Duffy, 2002; Perneger *et al.*, 2005; Schenker *et al.*, 2006). The reasons for the non-response are rarely completely independent from both observed and unobserved values, meaning that data are not missing completely at random (MCAR). Consequently, inferences based solely on the respondents will be misleading. If the differences between respondents and non-respondents can be explained entirely by differences in the observed data, such as characteristics of the patients, data are said to be missing at random (MAR). If so, it is possible to condition analyses on observed factors, thereby correcting the bias caused by missing data. However, differences between respondents and non-respondents may depend on unobserved values, in which case data are missing not at random (MNAR). If missingness is associated with unmeasured factors, conditioning on the observed data will not eliminate entirely potential bias.

One area where missing data has recently raised important concerns is in the assessment of hospital performance (Kirkham, 2008; Gale *et al.*, 2011; Groene *et al.*, 2014). Missing data may bias performance assessments through several routes. Firstly, within each provider, individuals with complete information tend to be systematically different from those with missing data. Secondly, provider assessments that are based on smaller samples will lead to increased uncertainty in the estimation of provider effects. Indeed, providers with large proportions of non-response will be less likely to be identified as statistically significantly better or worse than the benchmark. Thirdly, the reasons for the missing data may be related to the provider, not just the patient. For example, hospitals may differ in their data collection according to observed characteristics such as their volume of activity and their staffing arrangements (McCall *et al.*, 2004; Hutchings *et al.*, 2014).

The English patient reported outcome measures (PROMs) programme involves collecting survey responses from patients in order to facilitate comparative performance assessment of different healthcare providers. Since April 2009 all providers of publicly-funded inpatient care in the National Health Service (NHS) have been required to collect both generic and condition-specific PROMs for four elective procedures: unilateral hip and knee replacements, varicose vein surgery, and groin hernia repairs. Patients having these procedures are invited to report their health status before and three or six months after surgery using paper-based questionnaires. By comparing these before and after measures, changes in health can be identified and used to better understand differences in the systematic effect that health providers have on their patients' health (Appleby and Devlin, 2004). However, as with other health surveys, patients are not obliged to participate so responses will be missing if some fail to do so. Data may also be missing because providers differ in the effort they exert in overseeing data collection. Indeed, their efforts in this regard may be linked to their performance in terms of improving their patients' health status.

Recent incentive schemes linking payments to achievements in terms of health improvements have explicitly contemplated the issue of missing data. For example, the English 'best practice tariff' (BPT) pay-for-performance scheme comprises two components (Monitor, 2013). First, providers qualify for bonus payments if they do not perform statistically significantly below a national benchmark with respect to risk-adjusted improvements in patients' health status. Second, in order to receive this bonus, providers must ensure they collect PROMs for over 50% of the patients eligible for the survey. This policy creates an incentive for providers to meet the minimum standard for data collection, but this does not necessary eliminate the problem of missing data (Gutacker *et al.*, 2014). An important concern is whether the assessment of hospital's relative performance crucially hinges on the assumptions made about the reasons for the missing data.

This paper presents a strategy for addressing the missing data in PROMs, and assesses its impact on the use of PROMS for comparing provider performance. Here, we consider multiple imputation methods, which offer particular advantages for addressing missing data in performance assessments compared to other commonly used approaches such as maximum likelihood and inverse probability weighting. In the next section, we describe the data and the different missing data patterns in the PROMs survey. Section 3 presents the methods for estimating provider-specific outcomes and illustrates the implications of non-response for reporting provider performance. Section 4 describes the approach for dealing with the missing data under MAR and sensitivity analyses to investigate potential departures from MAR. Section 5 reports the results on provider performance according to different assumptions made about the missing data. The last section discusses the findings and highlights some priorities for future research.

2.   PROMs data

Our sample includes all patients aged 12 or over, who underwent primary, unilateral hip replacement surgery during the period April 2011 to March 2012. All providers of publicly-funded inpatient care in the English NHS are required to offer a pre-operative PROM questionnaire (Q1) to all patients deemed fit for surgery. Patients complete this questionnaire, usually during the last outpatient appointment preceding the surgery or on the day of admission. Patients are surveyed again approximately six months after surgery via another questionnaire sent by mail (Q2). This post-operative questionnaire is administered at the national level by an organisation contracted by the Department of Health.

Patient's health status before and after surgery is measured using a condition-specific measure, the Oxford Hip Score (OHS), and a generic quality-of-life measure, the EuroQol-5 dimensions (EQ-5D). Here, we focus on the former given it is used for BPT arrangements (Monitor, 2013). OHS consists of 12 components (questions) on different aspects of hip problems such as pain and functioning (Dawson *et al.*, 1996; Murray *et al.*, 2007). Each component has five possible responses, scored from 0 (most severe symptoms) to 4 (no symptoms). The overall score is a simple unweighted sum of all individual components, ranging from 0 (most severe level and highest number of symptoms) to 48 (least symptoms).

Administrative data about all patients having hip replacement, irrespective of whether they completed a PROM survey, are available in the Hospital Episodes Statistics (HES). HES includes detailed patient-level hospital records about all NHS-funded inpatient care provided by public and private hospitals in England. HES data are linked to the PROM survey responses through a matching variable provided by the Health & Social Care Information Centre (HSCIC). This linkage allows us to (i) ascertain the full population of hip replacement patients who were eligible to complete the PROMs survey; and (ii) obtain important clinical and socio-demographic information about the patients that would not otherwise be available in PROMs.

The PROM data collection and linkage process may result in different types of missing data. For example, pre-operative questionnaires may be missing because: i) the HSCIC was unable to match the Q1 PROM record to the HES episode, because the NHS number was missing or there was an insufficient match; ii) the provider failed to administer the Q1 questionnaire; or (iii) the patient refused to complete it. By design, all patients that received a Q2 questionnaire had to have filled in a Q1 questionnaire which had successfully been linked to the corresponding HES record. In other words, response to Q2 is conditional on response to Q1, and therefore, patients with missing Q1 questionnaires will also have missing Q2 questionnaires. Finally, for patients whose Q1 questionnaire was completed and successfully linked to HES, there may be missing data in the Q2 questionnaire. For example, patients may fail to return Q2 or may leave some questions unanswered (item non-response).

3. Missing data in the assessment of provider performance

3.1. Estimation of provider-specific outcomes

For provider-specific outcomes to be comparable, adjustment for the different case-mix of patients within each provider is required. This is typically undertaken using a regression framework, which we will denote as the *analysis* model. In this paper, we built upon the NHS case-mix adjustment methodology to estimate provider-specific outcomes (Nuttall *et al.*, 2013). Let $y_{2,ij}$ be the post-operative observed health outcome for patient i treated in provider j. We adjusted $y_{2,ij}$ for key patient characteristics (X_{ij}), such as age, gender, co-morbidities and socio-economic status (measured using an index of multiple deprivation), as well as the pre-operative health outcome ($y_{1,ij}$). The analysis model is defined as:

$$y_{2,ij} = \alpha + X_{ij}\beta + y_{1,ij}\gamma + u_j + \varepsilon_{ij,} \qquad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon) \qquad u_j \sim N(0, \sigma_u) \qquad (1)$$

Both the provider-specific unobserved effects (u_j) and the error term (ε_{ij}) are assumed to be normally distributed with zero mean and constant variance. Provider-specific effects can be estimated using a fixed-effects or random-effects model. Here, we considered the latter as it is typically more efficient[1]. To estimate how provider j performs relative to the national average we considered an indirect standardisation approach recommended by the NHS case-mix adjustment (Nuttall *et al.*, 2013):

$$\tilde{y}_{2,j} = \rho_j \bar{y}_{2,} \qquad \hat{\rho}_j = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_{2,ij}}{\hat{y}_{2,ij}}\right) \qquad (2)$$

The adjusted provider-specific outcome ($\tilde{y}_{2,j}$) is obtained by multiplying the national average outcome ($\bar{y}_{2,}$) by a provider-specific factor ($\hat{\rho}_j$), which reflects the extent to which provider's observed outcome ($y_{2,ij}$) compares with its expected outcome ($\hat{y}_{2,ij}$).
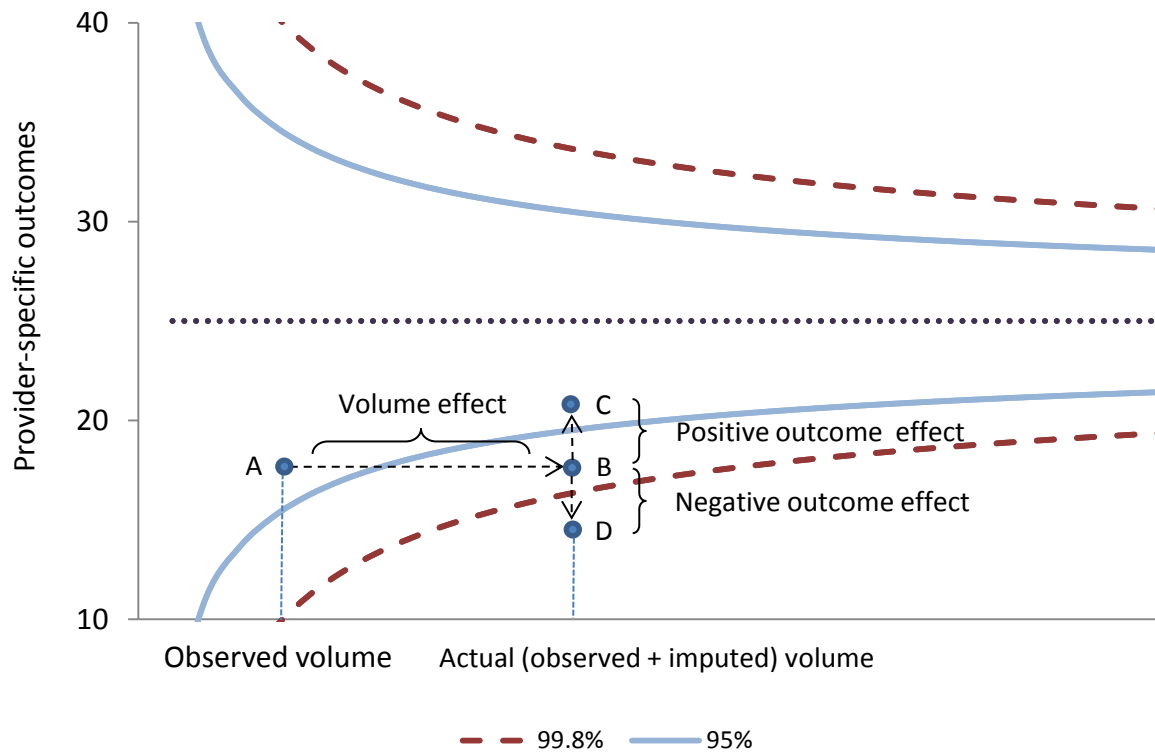
3.2. Funnel plots and missing data

It is common to compare provider performance using funnel plots (Spiegelhalter, 2005; Department of Health, 2012). Here, outcomes are plotted against volume (the number of patients treated) with 95% (2 standard deviations) and 99.8% (3 standard deviations) control limits used to indicate those providers that perform better or worse than expected (Figure 1). Accordingly, those providers located above the 95% and 99.8% control limits are judged to have a positive *alert* and *alarm* status, respectively, while negative *alerts* and *alarms* are those located below the 95% and 99.8% control limits, which are often under greater scrutiny. The pay-for-performance BPT scheme requires that providers do not perform statistically significantly below the national average, i.e. being located outside the 99.8% control limit (negative *alarm*), to be eligible for bonus payments. The BPT guidance sets out the requirement to use funnel plots to report provider performance, although other approaches such as caterpillar plots or z-scores are also available (Goldstein and Spiegelhalter, 1996).

With missing outcomes, the assessment of provider performance requires careful consideration. Ideally, we would like to remove the impact of non-response from inferences about provider performance, in order to locate the hospital near its 'true' mean outcome and volume on the funnel plot. Consider a hospital with a low response rate such that its observed mean outcome and volume

---

[1] Nuttall *et al.* (2013) reports coefficient estimates obtained using both fixed and random effects estimators. While the Hausman test rejects the null hypothesis that the random effects estimator is unbiased, the results on comparative provider performance are similar between the two approaches.

are at point A (*in control*) on a standard funnel plot (Figure 1). Instead, if all patients were observed for this hospital, and the unobserved outcomes were similar to those observed (data were MCAR), then the hospital would be located at point B (negative *alert*). In this particular stylised example, because of the missing data, the provider would be judged to be *in control* rather than a negative *alert*. From the hospital's point of view, this is a more desirable status, but such an assessment is misleading.

Figure 1. Missing data and its implications for assessing provider performance via funnel plots

The assumption that the patients with observed outcomes are similar to those who have missing outcomes (MCAR) is unlikely to hold. Missing data will be dependent on observed factors (MAR), other than those included in the analysis model, and might depend on unobserved values (MNAR). Therefore, the handling of missing data under MAR or MNAR will have three main effects for assessing provider performance via funnel plots[2]:

i) **Volume effect**. This will always shift the provider location to the right, for example, from point A to B (Figure 1). *Ceteris paribus*, moving from the observed to the actual volume increases the probability of being located outside the control limits.

ii) **Mean outcome effect**. Under MAR or MNAR, the provider-specific mean outcome may shift downwards or upwards depending on whether the imputed outcomes are better or worse than the observed outcomes given the observed data. In the example above, the hospital would move from point B to C (*in control*) if the patients for whom data are missing have a relatively better profile than do those for which data are observed, or to point D (negative *alarm*) otherwise (Figure 1).

---

[2]All these effects also apply for assessments based on z-scores or caterpillar plots.

iii)   **Variability effect**. The control limits will be narrower or wider depending on whether the variability in the full population (observed and imputed outcomes) is smaller or larger than that for the sample of patients whose outcomes are observed.

In sum, the impact of addressing the missing data on the location of the provider in the funnel plot will be chiefly determined by the provider response rate (size of volume effect) and how different missing outcomes are from those observed (size and direction of mean outcome effect).

4.   Methods for addressing the missing data

4.1.   Complete case analysis

A common approach for dealing with missing data is to discard patients for whom any outcome or covariate is missing. While complete case analysis (CCA) is simple to implement, this approach is only valid when data are MCAR. This implicitly assumes that individuals with complete data are representative of those with missing data, conditional on the variables included in the analysis model. The current official approach to performance assessment using PROMs is to apply CCA.

4.2.   Multiple imputation

With MI, each missing value is replaced by a set of plausible values, which are drawn from the posterior distribution of the missing outcomes given the observed data. Standard implementation of MI assumes that the probability of observing the outcomes is independent of any unobserved values, given the observed data (MAR). After imputation, the analysis model is applied to each multiple imputed dataset to estimate the parameters of interest. The multiple imputed estimates are typically combined using Rubin's rules (Rubin, 1987), which properly reflect the variation between and within imputations.

A key feature of MI is that the model for the missing data is estimated separately from the analysis model for estimating the parameters of interest. This allows us to include in the imputation model *auxiliary* variables in addition to those used in the analysis model that are associated both with the outcome and missingness. This is an important advantage of MI when compared with commonly used maximum likelihood approaches, which makes a potentially stronger MAR assumption that all observed factors that give rise to the missing data are included in the pre-specified analysis model. Including these *auxiliary* variables in the imputation model can reduce bias, improve precision and help make the MAR assumption more plausible. An additional advantage of MI is that its framework naturally extends to the assessment of alternative assumptions about the missing data mechanism (section 3.2).

For MI to provide valid inferences, the imputation model must accommodate the structure and the distribution of the data. In PROMs, the imputation model needs to recognise that the probability of non-response may be more similar within than across providers. Indeed, missingness may depend on observed patient-level characteristics, which tend to be more similar within the provider, and on provider-level characteristics such as whether the provider is a public (NHS) or private hospital. Compatible with the analysis model, this can be achieved by including provider-specific random effects (Gomes *et al.*, 2013).

The distribution of the outcome measure, the Oxford Hip Score, is left skewed, bimodal and with spike at 48. Typically, MI assumes Normality for continuous outcomes, and finding a suitable transformation to help normalise this outcome can be difficult. An alternative approach is to address the missing data in the individual components of the Oxford Hip Score, which are ordered (0 to 4). With ordinal components, we can consider a latent Normal variable for each component of the score. An important advantage of the latent Normal approach (Albert and Chib, 1993), which is equivalent to the probit model, is that it naturally links with the multivariate normal imputation model, easily implemented in standard software (Carpenter *et al.*, 2011). Hence, we can impute these latent variables, assuming their variance is restricted to one, along with other continuous variables, for example missing covariates.

Let h_{ij}^k be the observed kth component ($k = 1, ..., K$), with M ordinal categories ($m = 1, ..., M$), of the self-reported score for individual i in provider j (for the OHS score, $K = 12$ and $M = 5$). Let

$\pi_{ij,m}^k = \Pr(h_{ij}^k = m)$ and $\gamma_{ij,m}^k = \Pr(h_{ij}^k \leq m)$. By considering the ordered probit link model (Green, 2003), $\text{probit}(\gamma_{ij,m}^k) = \Phi^{-1}(\gamma_{ij,m}^k) = \alpha_m^k$, then h_{ij}^k can be described as a latent normal variable, Z_{ij}^k, with the following threshold model:

$$h_{ij}^k = \begin{cases} 1 \; if \; Z_{ij}^k \leq \alpha_1^k \\ m \; if \; \alpha_{m-1}^k < Z_{ij}^k \leq \alpha_m^k, \quad m = 2, \dots, M-1 \\ M \; if \; Z_{ij}^k > \alpha_{M-1}^k \end{cases} \tag{3}$$

The threshold parameters, α_m^k, define the mth category of the component k. The multivariate latent model is then given by $\mathbf{Z}_{ij} = \boldsymbol{\beta X}_{ij} + \boldsymbol{u}_j + \boldsymbol{e}_{ij}$, with

$$\boldsymbol{e}_{ij} \sim N\left[\boldsymbol{0}, \boldsymbol{\Omega}_e = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}\right] \quad \boldsymbol{u}_j \sim N\left[\boldsymbol{0}, \boldsymbol{\Omega}_u = \begin{pmatrix} \tau_1^2 & \cdots & \phi\tau_1\tau_K \\ \vdots & \ddots & \vdots \\ \phi\tau_K\tau_1 & \cdots & \tau_K^2 \end{pmatrix}\right]$$

Where $\mathbf{Z}_{ij} = Z_{ij}^1, \dots, Z_{ij}^K$, $\mathbf{X}_{ij} = X_{ij}^1, \dots, X_{ij}^K$ are the auxiliary variables, $\boldsymbol{u}_j = u_j^1, \dots, u_j^K$ are the provider-specific random effects, and $\boldsymbol{e}_{ij} = e_{ij}^1, \dots, e_{ij}^K$ are the error terms. The level-1 variance (σ_k^2) is constrained to 1. Level-2 correlation (ϕ) is often set to zero to avoid over-parameterisation at level-2. The MCMC algorithm to impute k ordinal components is provided in Appendix 1.

We implemented a distinct imputation model for missing data pattern at Q1 and Q2. Imputation models included all predictors included in the analysis model, and other auxiliary variables which were predictive of missingness and associated with the post-operative outcome, identified in previous studies (Hutchings *et al.*, 2012; Gutacker *et al.*, 2014; Hutchings *et al.*, 2014). For example, in addition to the risk-adjustment predictors, we included post-operative variables such as length of stay, and provider-level characteristics such as whether the provider was public or private.

We conducted 100 imputations and 10,000 MCMC iterations, with each set of imputed values being drawn from the posterior distribution at every 100th iteration of the MCMC chain. After imputation, we combined the individual OHS components into an overall OHS score. Then, we applied the analysis model (equation 1) to each multiple imputed dataset to estimate our parameter of interest, the adjusted provider-specific post-operative OHS (equation 2), and combined the results using Rubin's rules. All analyses were implemented in Stata, version 13, with imputations conducted in the software REALCOM-impute called from Stata (Carpenter *et al.*, 2011).

4.3. Sensitivity analysis

The approach taken to handling missing requires careful consideration of the different reasons that may give rise to the missing data (Little and Rubin, 2002). Since the true missing data mechanism is unknown, it is important to examine whether inferences about comparative provider performance are robust to alternative assumptions concerning the reasons for the missing data. A practical approach is to undertake sensitivity analyses after MI under MAR. Here, we considered a weighting approach after MI which uses the concept of importance resampling (Carpenter *et al.*, 2007).

Let $R_{ij} = 1$ if the outcome ($y_{2,ij}$) is observed, 0 otherwise. Let the probability of observing the outcome, $P(R_{ij} = 1)$, depend on observed patient (W_{ij}) and provider (Z_j) characteristics, but also on the underlying outcome, $y_{2,ij}$.

$$P(R_{ij} = 1 | W_{ij}, Z_j, y_{2,ij}) = \eta_0 + W_{ij}\eta_1 + Z_j\eta_2 + \varphi_j + y_{2,ij}\delta + \epsilon_{ij} \tag{4}$$

Equation 4 collapses to a MAR mechanism when $\delta = 0$. The basic idea is to explore the sensitivity of the results as δ departs from 0. After we have generated M multiple imputed datasets under MAR, we apply the analysis model to each dataset and obtain M estimates. Then, instead of a simple average, a weighted average is computed after assigning a relatively higher weight to those imputations judged to have a more plausible MNAR mechanism (for a chosen δ).

Suppose we order the data so that patients $i = 1, \dots, n_1$ have missing outcomes, and patients $i = n_{1+1}, \dots, N$ have complete data. For patients with incomplete data, let y_{ij}^m denote the mth imputed value under MAR, $m = 1, 2, \dots, M$. The weights (w_m) are computed as follows:

$$w_m = \frac{\widetilde{w}_m}{\sum_{i=1}^{m} \widetilde{w}_m} \qquad\qquad \widetilde{w}_m = \exp\left(\sum_{i=1}^{n1} -\phi y_i^m\right) \qquad\qquad (5)$$

Then, under the MNAR model implied by δ, the parameter of interest ($\tilde{y}_{2,j}$) and its variance can be calculated using a re-weighted version of Rubin's rules:

$$\tilde{y}_{2,j}^{MNAR} = \sum_{m=1}^{M} w_m \tilde{y}_{2,j} \qquad\qquad Var_{\tilde{y}_{2,j}}^{MNAR} = \tilde{V}_W + \left(1 + \frac{1}{M}\right)\tilde{V}_B \qquad\qquad (6)$$

Where $\tilde{V}_W = \sum_{m=1}^{M} w_m \hat{\sigma}_m^2$ is the within-imputation variance, and $\tilde{V}_B = \sum_{m=1}^{M} w_m \left(\tilde{y}_{2,j} - \tilde{y}_{2,j}^{MNAR}\right)^2$ is the between-imputation variance. For example, when δ is positive, the probability of response is higher for patients reporting better health. This means that for imputations under MAR patients reporting poorer outcomes will be under-represented. The weights correct for this by up-weighting the estimates from those imputations where the imputed outcomes are lower.

5.   Results

Table 1 reports descriptive statistics of patient characteristics (included in the risk adjustment) and outcomes for our sample of individuals undergoing elective hip replacement surgery in 2011-2012. Most patients were women, over 55 years old and white. Typically, these patients had 0 or 1 co-morbidities and symptoms during the last 5 years or longer. On average, patients had a substantial health improvement six months after surgery, as measured by their OHS (mean OHS more than doubled). The overall proportion of individuals with both incomplete pre- and post-operative outcomes and covariates was 48%.

Table 1. Descriptive statistics of patient characteristics and outcomes included in the risk adjustment model, and their proportion of missingness, for individuals undergoing hip replacement in 2011-2012 (N=71,821).

Variable	N (%) or Mean (SD)	% observed
Male	28,979 (40%)	100%
Age		100%
Under 55	8,694 (12%)	
55-65	15,736 (22%)	
65-75	25, 133 (35%)	
Over 75	22, 376 (31%)	
Ethnicity (Non-White)	8,027 (11%)	89%
Charlson Comorbidity Index		
0	51,957 (72%)	
1	15,213 (21%)	
2 or more	4,769 (7%)	
Index of Multiple Deprivation quintile		100%
Most deprived	9,163 (13%)	
2nd	12,468 (17%)	
3rd	16,524 (23%)	
4th	18,399 (26%)	
Least deprived	15,166 (21%)	
Duration of symptoms		61%
Up to 1 year	6,182 (14%)	
1-5 years	29,365 (67%)	
Over 5 years	8,479 (19%)	
Pre-operative Oxford Hip Score	17.5 (8.4)	61%
Post-operative Oxford Hip Score	38.1 (9.5)	52%

Figure 2 shows the distribution of the post-operative OHS after risk adjustment, for both the complete cases and after imputation. MI led to a slightly lower mean (standard deviation) post-operative OHS when compared with CCA: 37.3 (3.26) versus 38.1 (3.95). This suggests that patients with missing outcomes were associated with a somewhat poorer profile according to observed factors included in the imputation model.
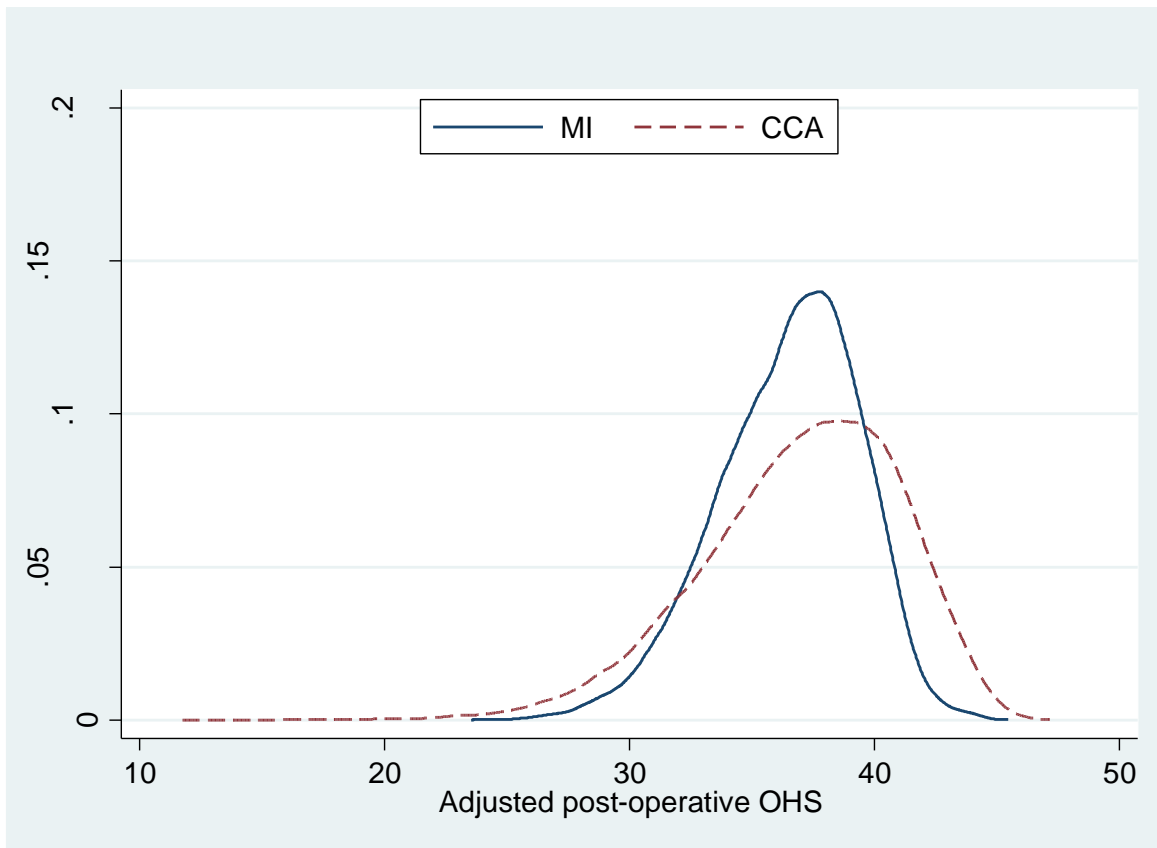
Figure 2. Kernel density of the risk-adjusted post-operative OHS for complete case analysis (CCA) and multiple imputation (MI).

Figure 3 illustrates the funnel plots of the provider-specific OHS according to CCA and MI. The width of the control limits is similar between the two approaches. That is, conditional on the observed data, the variability in the individual outcomes is relatively similar between the observed sample and full population of patients. The proportion of providers performing statistically above and below the national average (according to both 95% and 99.8% control limits) is higher after MI. The changes in the performance status appear to be dominated by the volume effect, as illustrated in Figure 3b and 3c.
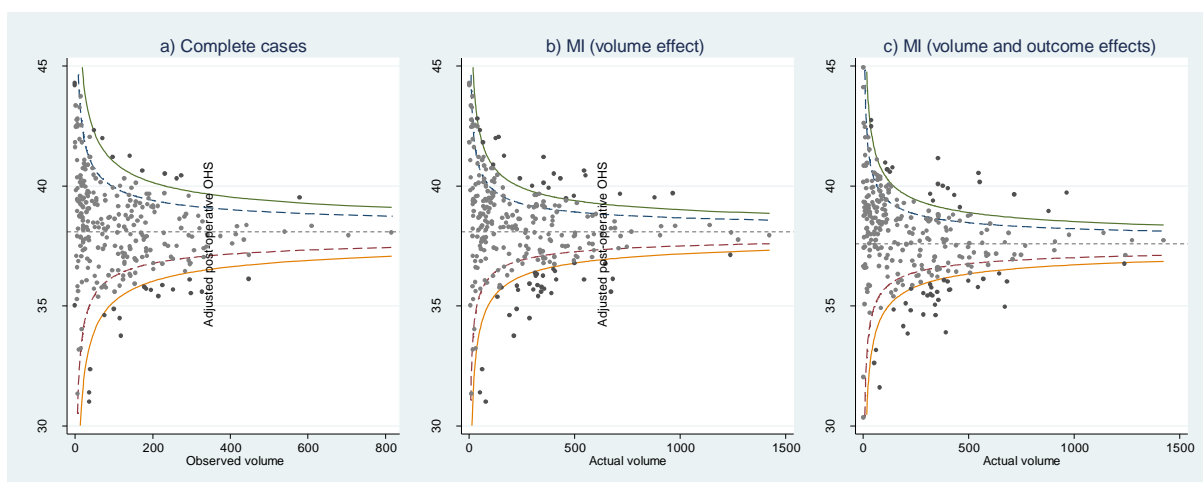
Figure 3. Funnel plots of provider-specific outcomes according to complete cases (N=279), and after multiple imputation: volume and mean outcome effects (N=298).

Table 2 reports the number of providers under each performance category according to CCA and MI. Overall, CCA leads to type-II errors by failing to detect statistically significant outperformers (positive *alarms*) and underperformers (negative *alarms*). For example, the proportion of negative *alarms* (outside the lower 99.8% control limit) varied from 7% (N=20) with complete cases to 11% (N=32) after MI. Most of the providers that moved to an *alarm* status after MI were previously judged *alerts* based on CCA (more details are available in Appendix A1 in the online supplementary material). Under the new BPT, the number of providers who would be ineligible for a bonus (based on their performance status and the minimum 50% response rate) scheme was 67 (22%) according to CCA and 78 (26%) after MI (Appendix A1).

Table 2. Provider performance status according to complete-case analysis (CCA) and multiple imputation (MI).

	CCA		MI (volume effect)		MI (Volume and outcome effects)	
	N	**%**	**N**	**%**	**N**	**%**
Negative *alarm*	20	7.2	32	10.7	32	10.7
Negative *alert*	22	7.9	27	9.2	24	8.1
In control	214	76.7	185	61.9	187	62.8
Positive *alert*	14	5.0	32	10.7	35	11.7
Positive *alarm*	9	3.2	22	7.4	20	6.7
Total	279	100.0	298	100.0	298	100.0

Note: PROMs were entirely missing for 19 providers, and hence, these were not assessed under complete case analysis.

Table 3 reports the results for $-0.3 \leq \delta \leq 0.3$. The relative provider performance remained fairly similar beyond these values, as a relatively lower number of imputations receive high weights. For example, $\delta = 0.3$ means that the odds of response is 1.35 (exp(0.3)) times greater for patients with an additional unit in the post-operative OHS. There is some evidence that patients reporting poorer outcomes were somewhat under-represented in the imputations under MAR. Indeed, for positive values of δ, there was a relatively higher number of providers with negative *alarm* status and fewer providers with a positive *alarm* status. The distribution of the relative provider performance remained fairly unchanged for negative values of δ, i.e. when there is a negative association between the probability of observing the outcome and the outcome level.

Table 3. Sensitivity analyses to departures from missing at random (MAR) represented by alternative missing not at random (MNAR) mechanisms.

	Performance status according to Oxford Hip Score				
	Positive *alarm*	Positive *alert*	*In control*	Negative *alert*	Negative *alarm*
MNAR, $\delta = -0.3$	22 (7.4%)	28 (9.5%)	193 (65.4%)	22 (7.5%)	31 (10.5%)
MNAR, $\delta = -0.2$	21 (7.1%)	29 (9.8%)	192 (65.1%)	23 (7.8%)	31 (10.5%)
MNAR, $\delta = -0.1$	18 (6.1%)	28 (9.5%)	193 (65.4%)	22 (7.5%)	34 (11.5%)
MAR, $\delta = 0$	20 (6.7%)	35 (11.7%)	187 (62.8%)	24 (8.1%)	32 (10.7%)
MNAR, $\delta = 0.1$	14 (4.8%)	23 (7.8%)	192 (65.1%)	26 (8.8%)	40 (13.6%)
MNAR, $\delta = 0.2$	13 (4.4%)	21 (7.1%)	191 (64.8%)	21 (7.1%)	43 (14.6%)
MNAR, $\delta = 0.3$	13 (4.4%)	18 (6.1%)	195 (66.1%)	25 (8.5%)	44 (14.9%)

6.   Discussion

The English PROMs survey provides valuable evidence about the effectiveness of the health care delivered by the NHS in terms of health improvements to patients. An important aim of this programme is to support evaluations of the relative performance of providers in undertaking elective procedures such as hip replacement. As with other self-reported, voluntary health surveys, PROMs are prone to large proportions of missing data due to reasons that are associated with both the patient and the provider. However, existing official assessments of the relative provider performance that use PROMs are based solely on the complete cases.

This study is a first attempt to address the implications of missing data for the use of PROMs in the assessment of provider performance. We present a strategy for addressing the missing data using multiple imputation methods, and illustrate the impact of using such methods in reporting comparative performance. A major finding of the paper is that judgments about provider performance differ according to the assumptions made about the reasons for non-response. We found that analyses according to complete cases led to a substantially lower number of providers performing statistically above or below the national average. By ignoring the problem, this method assumes that the missing data are unrelated to both observed and unobserved values, conditional on the risk-adjustment predictors. Previous studies showed that this assumption is implausible in the context of PROMs (Hutchings *et al.*, 2012; Gutacker *et al.*, 2014; Hutchings *et al.*, 2014).

We propose a strategy for handling the missing data using MI methods. MI is particularly suitable for addressing non-response in PROMs for several reasons. Firstly, the model for the missing data is estimated separately from the analysis model. This offers particular advantages when compared, for example, with standard maximum likelihood approaches, because it allows the imputation model to include auxiliary variables that are predictive of missingness, without having to modify the pre-specified risk adjustment model. Secondly, MI provides a flexible framework for assessing the sensitivity of the results to departures from MAR. Here, we considered sensitivity analysis by re-weighting, but alternative approaches such as pattern mixture models are readily available (Carpenter and Kenward, 2013). Thirdly, missing data patterns observed in PROMs are non-monotone. Under this pattern, MI methods are preferable to alternative practical approaches such as inverse probability weighting (Carpenter *et al.*, 2006). Fourthly, the MI approach fits well with the method used for reporting comparative provider performance. It facilitates identification of provider's performance status in the funnel, given its estimated mean outcome (under MAR) and actual (observed and imputed) volume.

Our findings have important implications for policy making. Firstly, given that the proportion of observed PROMs varies considerably across providers, appropriate mechanisms for improving data collection are warranted. If the provider expects to be an underperformer, it will have little incentive to have a higher response rate since that makes them more likely to be identified as such. This is because the uncertainty around the provider effect, as reflected in their confidence intervals or, equivalently, the control limits, is an inverse function of the sample size. We have illustrated this using funnel plots, but reporting provider performance via alternative methods such as caterpillar plots would reach a similar conclusion; lower response rates makes below-average provider effects more imprecise, and hence not statistically different from zero. By imposing a 50% (or above) response rate criterion in order to qualify for a bonus payment for good performance, the BPT initiative provides a good start. However, this may not eliminate the problem as providers may be induced to collect just enough data to satisfy the requirement, still leading to a potentially large proportion of non-response and potentially unrepresentative samples. Increasing the threshold above 50% in the future can encourage providers to continually improve data collection.

Secondly, commentators suggest that provider-level characteristics should not be included in the model for estimating provider-specific outcomes if they do not constitute binding production constraints (Smith and Street, 2005). However, with missing data these variables should be carefully considered in the specification of the imputation model in order to minimise any potential bias due to differences between providers with different proportions of observed PROMs.

This paper has a number of limitations. First, we restricted our sample to the last available data cohort for hip replacement patients. Previous studies showed that non-response was higher in previous cohorts and other elective interventions (Gutacker *et al.*, 2014; Hutchings *et al.*, 2014), and so dealing with missing data would be likely to have greater implications for the inferences on provider performance.

Second, our approach was to assign all patients to the provider of care reported in the HES inpatient record. However, care is sometimes subcontracted to private providers (Independent Sector Treatment Centres) and the assignment of patients to providers may therefore be incorrect. Given that NHS providers can subcontract a proportion of their activity with multiple independent centres, it is impossible to be precise about where treatment actually took place. We therefore acknowledge that our assessment of provider response rates relies on the assumption that provider codes have been recorded correctly. However, the impact on the estimated response rates is likely to be small (Hutchings *et al.*, 2014).

Third, we have focused our analysis on the disease-specific Oxford Hip Score as this is the preferred outcome measure in the pay-for-performance BPT scheme. The impact of addressing missing data for the assessment of provider performance was very similar for the generic quality-of-life measure, the EQ-5D. Results for this outcome can be found in Appendix A2 in the online supplementary material.

Fourth, throughout we assumed the imputation model was correctly specified. We followed methodological guidance and specified an imputation model that was compatible with the analysis model, for example, by including random-effects and non-linear interactions (White *et al.*, 2011; Carpenter and Kenward, 2013). However, the use of methods that are less sensitive to the correct specification of the imputation model, such as Robust MI, may warrant consideration (Daniel and Kenward, 2012).

In conclusion, our analysis shows that inferences on the relative provider performance using PROMs differ according to the assumptions made about the missing data. Assessments based solely on observed PROMs are shown to underestimate the proportion of poorly performing providers. We provide a strategy for addressing the missing data that makes more plausible assumptions about non-response given the observed data. Considerable attention has been given to investigating the factors associated with patient non-response in health surveys. However, the reasons why health providers may differ in their ability and willingness to collect data are less well understood, and should be investigated further in the context of PROMs. In addition, future efforts are best invested in increasing provider response rates so that inferences on provider performance are less dependent on modelling assumptions.

Appendix 1

Below we describe a MCMC algorithm to impute k ordinal components by sampling from the posterior Normal distribution (using a latent Normal approach). Since each component may have a different missingness pattern, the appropriate conditional distribution will have to be derived for each component k, given all other components, Z^{K-1}, with $k = 1, \dots, K$. While the Gibbs sampler can be used (Albert and Chib, 1993), the more general Metropolis-Hastings sampler typically results in faster convergence (Cowles, 1996; Goldstein *et al.*, 2007) and is considered here. This is particularly advantageous for estimating the threshold parameters ($\boldsymbol{\alpha}$) and updating level-1 and level-2 covariance matrices ($\boldsymbol{\Omega}_e, \boldsymbol{\Omega}_u$). The algorithm proceeds as follows. At each iteration $r = 1, 2, \dots$

Step 1. The first step is concerned with generating each latent Normal variable, Z^k, given the observed ordinal component, h^k. For each component k

a) Sample category 1 from a standard Normal distribution $[-\infty, \alpha_1 - \hat{z}]$, category $m = 2, \dots, M - 1$ from $[\alpha_{m-1} - \hat{z}, \alpha_m - \hat{z}]$, and category M from $[\alpha_{M-1} - \hat{z}, \infty]$, where $\hat{z} = \hat{\beta}X + \hat{\delta}Z^{K-1} + \hat{u}$

b) Update the threshold parameters $\alpha_1, \dots, \alpha_{M-1}$. Let the component of the likelihood associated with a specific category be given by $L(\alpha_1, \dots, \alpha_M) = \prod_{i=1}^{N} \prod_{m=1}^{M} \pi_m^{I_{i,m}}$, where $I_{i,m}$ is 1 if $h_{ij} = m$, 0 otherwise, and

$$
\pi_m = \begin{cases} \displaystyle\int_{-\infty}^{\alpha_1 - \hat{z}} \varphi(t)dt, & if \ m = 1 \\ \displaystyle\int_{\alpha_{m-1} - \hat{z}}^{\alpha_m - \hat{z}} \varphi(t)dt, & if \ m = 2, \dots, M-1 \\ \displaystyle\int_{\alpha_{M-1} - \hat{z}}^{\infty} \varphi(t)dt, & if \ m = M \end{cases}
$$

For each $\alpha_1, \dots, \alpha_{M-1}$, we draw a proposal $\tilde{\alpha} \sim N(\alpha^{r-1}, v^2)$, and accept the proposal with probability $\min[1, \frac{L(\tilde{\alpha})}{L(\alpha^{r-1})}]$. If $\tilde{\alpha}$ is accepted, set $\alpha_m^* = \tilde{\alpha}$, otherwise $\alpha^r = \alpha^{r-1}$. In practice, the variance of the proposal distribution is typically set to $v^2 = 5.8/N$ (Gelman *et al.*, 2006).

Step 2. Draw the missing values conditional on the values drawn for $\boldsymbol{\alpha}^r$ and $\boldsymbol{Z}^r$; for each missing observation h_{ij}, impute $h_{ij} = m$ by finding the value of m such that $\alpha_{m-1}^* < Z_{ij}^r \leq \alpha_m^*$.

Step 3. Draw $\boldsymbol{\beta}^r$ from $\left[\left[\sum_{ij} \boldsymbol{\Lambda}^T \boldsymbol{\Omega}_e^{-1} \boldsymbol{\Lambda}\right]^{-1} \sum_{ij} \boldsymbol{\Lambda}^T \boldsymbol{\Omega}_e^{-1} (\boldsymbol{Z}_{ij} - \boldsymbol{u}_j)^T, \left[\sum_{ij} \boldsymbol{\Lambda}^T \boldsymbol{\Omega}_e^{-1} \boldsymbol{\Lambda}\right]^{-1} \right]$, where $\boldsymbol{\Lambda} = \mathbf{I}_{k \times k} \otimes X_{ij}$.

Step 4. Sample $\mathbf{u}_j^r$ from $MVN \left[[\sum_i \boldsymbol{\Omega}_e^{-1} + \boldsymbol{\Omega}_u^{-1}]^{-1} \sum_i \boldsymbol{\Omega}_e^{-1} (\boldsymbol{Z}_{ij} - \boldsymbol{\beta} X_{ij})^T, [\sum_i \boldsymbol{\Omega}_e^{-1} + \boldsymbol{\Omega}_u^{-1}]^{-1} \right]$. Level-1 residuals can be easily obtained by subtraction, $\boldsymbol{e}_{ij}^r = \boldsymbol{Z}_{ij} - \boldsymbol{\Lambda}\boldsymbol{\beta} - \boldsymbol{u}_j$

Step 5. Update the elements of $\boldsymbol{\Omega}_e^{r-1}$ and $\boldsymbol{\Omega}_u^{r-1}$, in that order and conditional on $(\boldsymbol{\alpha}^r, \boldsymbol{Z}^r, \boldsymbol{\beta}^r, \boldsymbol{u}^r)$, to obtain $\boldsymbol{\Omega}_e^r$ and $\boldsymbol{\Omega}_u^r$ (Browne, 2006).

References

Albert, J. & Chib, S. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association,* 88**,** 669-79.

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

Browne, W. 2006. MCMC algorithms for constrained variance matrices. *Computational Statistics & Data Analysis,* 50**,** 1655-1677.

Carpenter, J. & Kenward, M. 2013. *Multiple Imputation and its Application,* Chichester, UK., Wiley.

Carpenter, J. R., Goldstein, H. & Kenward, M. G. 2011. REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types. *Journal of Statistical Software,* 45**,** 1-14.

Carpenter, J. R., Kenward, M. G. & Vansteelandt, S. 2006. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society Series a-Statistics in Society,* 169**,** 571-584.

Carpenter, J. R., Kenward, M. G. & White, I. R. 2007. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Stat Methods Med Res,* 16**,** 259-75.

Cohen, G. & Duffy, J. C. 2002. Are nonrespondents to health surveys less healthy than respondents? *Journal of Official Statistics,* 18**,** 13-23.

Cowles, M. K. 1996. Accelerating Monte Carlo Markov Chain convergence for cumulative-link generalized linear models. *Statistics and Computing,* 6**,** 101-110.

Daniel, R. M. & Kenward, M. G. 2012. A method for increasing the robustness of multiple imputation. *Computational Statistics & Data Analysis,* 56**,** 1624-1643.

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

Department of Health 2012. Patient Reported Outcome Measures (PROMs) in England. A methodology for identifying potential outliers. *Department of Health policy papers,* 16340**,** 1-30.

Gale, C. P., Cattle, B. A., Moore, J., Dawe, H., Greenwood, D. C. & West, R. M. 2011. Impact of missing data on standardised mortality ratios for acute myocardial infarction: evidence from the Myocardial Ischaemia National Audit Project (MINAP) 2004-7. *Heart,* 97**,** 1926-31.

Gelman, A., Roberts, G. & Gilks, W. 2006. Efficient Metropolis jumping rules. *In:* Bernardo, J., Berger, J., Dawid, A. & Smith, A. (eds.) *Bayesian Statistics V.* Oxford: Oxford University Press.

Goldstein, H., Bonnet, G. & Rocher, T. 2007. Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics,* 32**,** 252-286.

Goldstein, H. & Spiegelhalter, D. J. 1996. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society Series a-Statistics in Society,* 159**,** 385-409.

Gomes, M., Diaz-Ordaz, K., Grieve, R. & Kenward, M. G. 2013. Multiple Imputation Methods for Handling Missing Data in Cost-effectiveness Analyses That Use Data from Hierarchical Studies: An Application to Cluster Randomized Trials. *Med Decis Making,* 33**,** 1051-63.

Green, W. H. 2003. *Econometric Analysis,* Upper Saddle River, N.J., Great Britain, Prentice Hall.

Groene, O., Kristensen, S., Arah, O. A., Thompson, C. A., Bartels, P., Sunol, R. & Klazinga, N. 2014. Feasibility of using administrative data to compare hospital performance in the EU. *Int J Qual Health Care,* 26 Suppl 1**,** 108-15.

Gutacker, N., Street, A., Gomes, M. & Bojke, C. 2014. Should English healthcare providers be penalised for failing to collect patient-reported outcome measures (PROMs)? Under submission.

Hutchings, A., Neuburger, J., Grosse Frie, K., Black, N. & Van Der Meulen, J. 2012. Factors associated with non-response in routine use of patient reported outcome measures after elective surgery in England. *Health Qual Life Outcomes,* 10**,** 34.

Hutchings, A., Neuburger, J., Van Der Meulen, J. & Black, N. 2014. Estimating recruitment rates for routine use of patient reported outcome measures and the impact on provider comparisons. *BMC Health Services Research,* 14**,** 66.

Kirkham, J. J. 2008. A comparison of hospital performance with non-ignorable missing covariates: an application to trauma care data. *Stat Med,* 27**,** 5725-44.

Little, R. J. & Rubin, D. B. 2002. *Statistical analysis with missing data,* New York, US, Wiley.
Mccall, N., Khatutsky, G., Smith, K. & Pope, G. 2004. Estimation of Non-Response Bias in the Medicare FFS HOS. *Health Care Financing Review,* 25**,** 27-41.

Monitor 2013. National Tariff Payment System - Annex 4A: Additional information on currencies with national prices: Pages 63-70. NHS England.

Murray, D. W., Fitzpatrick, R., Rogers, K., Pandit, H., Beard, D. J., Carr, A. J. & Dawson, J. 2007. The use of the Oxford hip and knee scores. *Journal of Bone and Joint Surgery-British Volume,* 89B**,** 1010-1014.

Nuttall, D., Parkin, D. & Devlin, N. 2013. Inter-Provider Comparison of Patient-Reported Outcomes: Developing an Adjustment to Account for Differences in Patient Case Mix. *Health Econ* DOI: 10.1002/hec.2999.

Perneger, T. V., Chamot, E. & Bovier, P. A. 2005. Nonresponse bias in a survey of patient perceptions of hospital care. *Med Care,* 43**,** 374-80.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys,* New York, US, Wiley.

Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G. & Cohen, A. 2006. Multiple Imputation of Missing Income Data in the National Health Interview Survey. *Journal of the American Statistical Association,* 101**,** 924-933.

Smith, P. C. & Street, A. 2005. Measuring the efficiency of public services: the limits of analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168**,** 401-417.

Spiegelhalter, D. J. 2005. Funnel plots for comparing institutional performance. *Statistics in Medicine,* 24**,** 1185-1202.

White, I. R., Royston, P. & Wood, A. M. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med,* 30**,** 377-399.