

Multiple imputation to deal with missing EQ-5D-3L data – should we impute individual domains or the actual index?

Running head: Multiple imputation to deal with missing EQ-5D-3L data

Keywords: EQ-5D-3L, missing data, multiple imputation, missing data pattern, quality of life

Abstract

Purpose

Missing data is a well-known and widely documented problem in cost-effectiveness analyses alongside clinical trials using individual patient-level data. Current methodological research recommends multiple imputation (MI) to deal with missing health outcome data, but there is little guidance on whether MI for multi-attribute questionnaires, such as the EQ-5D-3L, should be carried out at domain or at summary score level. In this paper we evaluated the impact of imputing individual domains versus imputing index values to deal with missing EQ-5D-3L data using a simulation study, and developed recommendations for future practice.

Methods

We simulated missing data in a patient-level dataset with complete EQ-5D-3L data at one point in time from a large multinational clinical trial ($n=1,814$). Different proportions of missing data were generated using a missing at random (MAR) mechanism and three different scenarios were studied. The performance of using each method was evaluated using root mean square error (RMSE) and mean absolute error (MAE) of the actual versus predicted EQ-5D-3L indices.

Results

In large sample sizes ($n>500$) and a missing data pattern that follows mainly unit non-response, imputing domains or the index produced similar results. However, domain imputation became more accurate than index imputation with pattern of missingness following an item non-response. For smaller sample sizes ($n<100$), index imputation was more accurate. When MI models were misspecified, both domain and index imputations were inaccurate for any proportion of missing data.

Conclusions

The decision between imputing the domains or the EQ-5D-3L index scores depends on the observed missing data pattern and the sample size available for analysis. Analyst conducting this type of exercises should also evaluate the sensitivity of the analysis to the missing at random assumption and whether the imputation model is correctly specified.

1. Introduction

Missing data is a well-known and widely documented problem in cost-effectiveness analyses alongside clinical trials using individual patient-level data. Several authors have compared and evaluated different statistical methods on how to deal with missing data in cost-effectiveness studies and have also made recommendations [1-7]. In general, researchers recommend that statistical methods such as multiple imputation (MI) [8], which incorporate uncertainty around the estimated missing values, may be implemented when replacing missing values in economic evaluation studies. Although the evidence about the use of MI to deal with missing cost and cost-effectiveness data is quite robust, the uptake of such statistical technique by health economists in applied studies is still slow as reported in a recent literature review [9].

So far, methodological research on handling missing resource use, cost, or a summary measure of cost-effectiveness (e.g. net benefits) has been widely conducted [1-7]. Dealing with missing health outcomes data in the economic evaluation of health care technologies, on the other hand, has not received the same level of attention. Reviews on how missing outcome data have been handled in clinical trials also suggest that missing outcome data are very often inadequately managed. Researchers mostly use complete case analysis to avoid missing items or missing total scores from disease-specific or generic health outcome instruments [10, 11].

In economic evaluations, a measure of health-related quality of life (HRQoL) is often used to obtain utility values for the calculation of the quality-adjusted life years (QALYs), which is the most commonly applied 'currency' to express overall health benefits in the analyses. HRQoL data is usually collected using preference-based instruments such as the EuroQol EQ-5D [12, 13] or the Health Utilities Index (HUI) [14]. The responses from these questionnaires are

transformed into a utility index using valuation sets that express preferences from the general public, patients or experts. The HUI Procedures Manual provides some guidance on how to handle missing data when respondents have missing items [14]. Nevertheless, although the EQ-5D is the preferred measure of HRQoL by several major health technology agencies, including the National Institute for Health and Care Excellence (NICE) [15], the developers of the instrument have not provided recommendations or guidance on how to deal with missing EQ-5D data.

The relevant literature on this is also limited. Thus far, we have only found one article exploring the impact of missing data in the EQ-5D-3L in a longitudinal study of liver transplant patients [16]. In such study, the authors explored different adjustments for informative dropouts and non-responders¹ and compared the results with an unadjusted base case analysis scenario. The authors evaluated the impact of 1) not adjusting, 2) adjusting for informative dropouts and non-responders in isolation, and 3) adjusting for both informative dropouts and non-responders at the same time in the EQ-5D-3L index. The results of the study suggested that adjusting for informative dropouts and non-responders yielded different results to the base case scenario of no adjustment for the post-transplant data. The authors concluded that researchers should consider carefully allowing for missing data in their analysis of EQ-5D-3L data, ideally using appropriate methods such as MI. The use of MI to impute missing quality of life scores has also recently been suggested in another study that

¹ Missing data due to dropouts for informative reasons are cases when the participant fails to complete a questionnaire as a result of their severity of their illness, death or other known reason. Non-responders occurs when the participant does not respond to a questionnaire at one or multiple time points, which creates different missing data patterns in a dataset. In cross-sectional studies the main missing data patterns are **unit non-response** when the participant fails to complete all the items within a questionnaire, and **item non-response** when the participant fails to complete some of the items within a questionnaire. In longitudinal studies the participant may drop out before the end of the study and do not return creating a **monotone** missing data pattern.

explored the impact of using different imputation methods for missing Quality of Well Being (QWB) scores in patients with lung-volume-reduction surgery [17].

Although the study by Ratcliffe and colleagues provided important guidance for applied researchers dealing with missing EQ-5D-3L index data, it did not address the question of missing data at the individual EQ-5D-3L domain levels (mobility, self-care, usual activities, pain/discomfort and anxiety/depression). A recent study describing a guide to handle missing data in cost-effectiveness analysis alongside patient-level data in randomised trials, reported that QALYs can be imputed at the HRQoL domain level, at the index score level or as QALYs, but not particular guidance on which option was preferable was included [7].

As any missing EQ-5D index is the result of missing responses at the domain level, the aim of the current study was to explore whether there is any advantage if imputation was carried out at the domain level and not at the aggregated index level. Our starting hypothesis was that imputing missing responses at the domain level could yield more precise EQ-5D-3L index scores estimates if complete information for some of the remaining domains is available. Such complete information have the potential to be included in an MI model acknowledging any correlations between domains improving hence the index score estimation. In this study we report a simulation exercise of missing at random (MAR) EQ-5D-3L data using a case study from a large randomised clinical trial. We explored MI at domain level versus index level at different levels of missing data (5%-40%) in relation to the proportion of unit non-response in the dataset, the available sample size and when the imputation model was misspecified with omitted relevant variables.

The rest of the paper proceeds as follows: the Methods section describes the study hypothesis, the case study and the design of the simulation exercise; the Results section

presents the findings from the base case and scenario analyses of the simulation exercise; and in the final Discussion section we examine the implications of the findings for best practices along with the limitations of the study.

2. Methods

2.1 The relationship between EQ-5D-3L index and EQ-3D-3L domains and study hypothesis

The mathematical relationship between the EQ-5D-3L index and domains is characterised by the country-specific value set used to calculate utility values. Most studies conducted in the literature have estimated value sets similar to the original UK valuation exercise [18], which is defined as follows:

$$EQ5D_{index} = 1 - \left(\hat{\beta}_{constant} + \sum_{i=1}^5 \sum_{m=2}^3 \hat{\beta}_{y_i=m} x_{y_i=m} + \hat{\beta}_{N3} N3 \right)$$

where each EQ-5D-3L domain (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) is indicated by y_i with $i = 1, 2, \dots, 5$ and can take m levels with $m = 2, 3$; $x_{y_i=m}$ represents ten dummy variables that indicate the presence of either a level 2 or a level 3 answer and $\hat{\beta}_{y_i=m}$ the associated coefficients; $\hat{\beta}_{constant}$ is the EQ-5D-3L coefficient corresponding to the constant in the UK value set and indicates deviations from full health; and $N3$ is a dummy that indicates that there is at least one dimension at level 3.

The above expression indicates a linear relationship between the EQ-5D-3L index and domains. Therefore, the index calculation retrieves a missing value if there are missing elements in the term $\sum_{i=1}^5 \sum_{m=2}^3 \hat{\beta}_{y_i=m} x_{y_i=m}$ in the right hand side of the expression.

Therefore, imputing index or domain may yield different results depending on the available missing data pattern. If the missing data pattern follows an item non-response (i.e. participants choose not to respond to one or more questions of the EQ-5D-3L

questionnaire), differences between index and domain imputation may be observed because partial domain information is available that could potentially inform an imputation model for the remaining missing domain responses. If the missing data pattern follows a unit non-response (i.e. all domain responses missing), imputing the domains or the index should yield similar results as no additional information is available at the domain level to inform an imputation model.

2.2 The design of the simulation exercise

A flowchart of the simulation exercise conducted is presented in Figure 1. The simulation started with a dataset with complete information on the EQ-5D-3L instrument, additional relevant health outcomes, some demographics and resource use covariates. Using this dataset, a pre-defined proportion of missing data was simulated for the EQ-5D-3L information using an algorithm based on missingness patterns observed for the EQ-5D-3L data in five large randomised controlled trials (Table 1). Two different imputation models were then estimated either using the EQ-5D-3L index or the domains as dependent variables. Imputed datasets were combined using Rubin's rule [8] to obtain estimated EQ-5D-3L indices. The estimated index values were then compared with the actual EQ-5D-3L index from the original complete case dataset. The results from the comparison between the actual and predicted indices were stored in a matrix. The whole process was repeated 1,000 times to minimise Monte Carlo error. Different proportions of simulated missing data at 5%, 10%, 20% and 40% were evaluated. The simulation exercise was conducted using the statistical software Stata version 12 [19].

<< Insert Figure 1 around here>>

<< Insert Table 1 around here >>

2.3 Case Study

Data for this study was obtained, with permission, from the International Subarachnoid Aneurysm Trial (ISAT). ISAT was a large prospective randomised controlled trial that compared neurosurgical clipping with endovascular coiling for patients with aneurysmal subarachnoid haemorrhage [20]. For this methodological study, a cross-sectional sample of 1,814 participants from the one year follow-up data was used. In addition to the HRQoL EQ-5D-3L, the complete case dataset included relevant demographics such as trial centre size, treatment arm, age, and gender; health outcomes such as the Glasgow Outcome Scale (GOS) [21] and the modified Rankin Scale (mRS) [22]- two widely used disease-specific instruments in stroke; and resource use variables such as operation duration, ITU days and hospital length of stay.

2.4 The EQ-5D-3L instrument

The EQ-5D-3L is a generic, multi-attribute HRQoL scale developed in the early 1990s to assist health economists conducting economic evaluations of health care technologies [12, 13]. It is designed to be completed by patients through postal surveys or through face-to-face interviews and is currently available in more than 100 languages [23]. In addition, it is the preferred measure of HRQoL in adults by the National Institute for Health and Care Excellence (NICE) in the United Kingdom [15].

The EQ-5D-3L has five domains (mobility, self-care, usual activities, pain/discomfort and anxiety/depression), and each domain has three levels corresponding to no problems (normally assigned the value 1), some/moderate problems (2) and extreme problems (3). Therefore, there are 243 (3^5) potential individual health states in the EQ-5D-3L and each state can be expressed as a five-digit number. For example, the health state 21221 corresponds to someone with some problems in mobility, usual activities and pain/discomfort, but no

problems with self-care and anxiety/depression. Each of the 243 health states can be converted into an index using a country-specific value set. In this study, we used the UK value set originally estimated from preferences of a representative British general population sample in 1993.

2.5 Missing data simulation (Base case)

When dealing with missing data, it is necessary to make some plausible assumptions about the nature of the missingness. The framework presented in the seminal work by Little and Rubin [8] is widely used to classify the missing data as being missing completely at random (MCAR - the probability of the data being missing does not depend on the observed or unobserved data), missing at random (MAR - the probability of the data being missing does not depend on the unobserved data, conditional on the observed data), and missing not at random (MNAR - the probability of the data being missing does depend on the unobserved data, conditional on the observed data). For this study, we created MAR data based on the algorithm by van Buuren et al. [24] assuming that the probability of the data being missing was related to the trial centre size, age, gender, and the stroke disease-specific health outcomes such as GOS and mRS. The relevant step-by-step guide is presented in the online appendix 1. We evaluated the effect of simulating different proportions of missing data at levels of 5%, 10%, 20% and 40%.

In order to ensure that the missing data patterns selected for the algorithm reflected patterns found in real practice, we explored missing EQ-5D-3L patterns in five large randomised controlled trials [20, 25-28]. Twenty-five different patterns of missingness across the EQ-5D-3L domains were identified in these trials. Table 1 summarises the seven most frequent patterns of missingness that were also used in our simulation of missing data. Based on these,

unit non-response (i.e. missing data in all domains) accounted for almost 90% of the replicated missing data patterns.

2.6 Multiple Imputation (MI)

MI is a statistical technique to deal with missing data that has become popular, particularly due to recent software developments [6, 29-31]. The method has the distinctive feature that missing values are not replaced by a unique value, but are replaced instead for a number of plausible values creating multiple datasets. These plausible values are estimated using a model based on the distribution of the observed data. Each dataset is analysed individually but using similar methods to obtain a parameter of interest (e.g. mean or regression coefficient). Finally, the estimates are combined using appropriate methods to obtain the overall parameter of interest, and associated measures of variability. When MI is implemented correctly, it produces asymptotically unbiased estimates and standard errors and is asymptotically efficient. A clear description of the implementation of MI and its different stages can be found in a recent MI tutorial by White and colleagues [31]. We used multiple imputation by chained equations (MICE) in Stata [19] using the command `mi impute chained` and created 50 imputed datasets. Rubin's rules were implemented using the command `mi estimate` also in Stata when imputing the EQ-5D-3L index. When imputing domains, EQ-5D-3L index scores were calculated using the UK value set and variances adjusted using Rubin's rule formulas [31].

2.7 Imputing the EQ-5D-3L index

The index was estimated using an ordinary least squares (OLS) linear model. The imputation model included the EQ-5D-3L index (with simulated missing values), trial centre size, age, gender, randomisation group, procedure time, ITU days, hospital length of stay, GOS and the

mRS. As OLS estimation can result in predictions which can be outside the possible EQ-5D-3L range (-0.594 and 1.000), we used prediction mean matching. With prediction mean matching, the predicted missing value from the imputation model is replaced with the closest value from the observed complete EQ-5D-3L index.

2.8 Imputing the EQ-5D-3L domains

Information from the EQ-5D-3L domains is categorical data and regression models such as ordered and multinomial logit can be used. We encountered convergence problems when estimating multinomial logit models even when the simulated missing data was at the 5% level. This was not the case when using ordered logit and it was hence the selected model to impute the domain data. The imputation model included the EQ-5D-3L domains (mobility, self-care, usual activities, pain/discomfort and anxiety/depression with simulated missing values), trial centre size, age, gender, randomisation group, procedure time, ITU days, hospital length of stay, GOS and the mRS. The imputed domains in each imputed dataset were used to calculate the EQ-5D-3L index using the UK tariff value set.

2.9 Assessing performance

Performance was evaluated comparing actual versus predicted EQ-5D-3L indices between each imputation model. For each simulation, a combined estimated mean estimate and Rubin-adjusted associated standard error across imputed datasets were calculated. Root mean-squared error (RMSE) and mean absolute error (MAE) were then computed to assess performance of models. The RMSE is defined as the squared root of the mean of squared differences between the actual EQ-5D-3L index and the EQ-5D-3L index estimated from the imputation model, whilst the MAE is the mean of absolute differences between actual and estimated EQ-5D-3L indices. Estimated means, and the corresponding standard errors, RMSE

and MAE were averaged across the 1,000 simulations. Comparisons between the actual and simulated frequency distributions of domain responses for each proportion of missing data were compared using Chi-squared test controlling for false discovery rate [32]. Given the number of multiple comparisons conducted and to minimise type I error, we selected a 1% significance level.

2.10 Scenario analyses

We explored three separate scenario analyses to evaluate the robustness of the base case results and obtain additional information necessary to improve guidance for future practice.

The first scenario analysis evaluated the impact of estimating missing data using different simulated patterns of missingness and was informed by our hypothesis that if the missing data pattern follows an item non-response differences between index and domain imputation may be observed (see section 2.1). The proportion of missing unit non-response in Table 1 was reduced to 60%, 40%, 20% and 0% to create different scenarios with higher proportions of item non-response missing data.

In a second scenario analysis, the base case simulation exercise was replicated reducing the total sample size from 1,814 to 900, 500, 200, 100 and 50 patients. Sample sizes were selected using a bootstrap resample while keeping the original trial allocation [33].

Recent guidance suggests that variable selection is important and that analysts should include in the imputation model all variables that are to be included in the final analysis model [31].

In the third scenario analysis, the simulation exercise was repeated but assuming the imputation model was not correctly specified and relevant variables were omitted in the specification of the model. To achieve this, we removed GOS and mRS that we originally

assumed were related to the probability of being missing from the imputation model (see section 2.5). Therefore, the final imputation model in both the index and domain imputations included only trial centre size, age, gender, randomisation group, procedure time, days in ITU and hospital length of stay plus the EQ-5D-3L data.

3. Results

3.1 Summary statistics of the complete case cohort

Descriptive statistics of the selected covariates for the 1,814 patients included in the complete case cohort are shown in Table 2. The table presents the covariates used as predictors of quality of life for the imputation model and the covariates used in the missing data simulation. The cohort represents a sample of patients who had an aneurysmal subarachnoid haemorrhage and a mean (SD) age of 51 (11.32) years with a proportion of 63% being females. The mean (SD) HRQoL measured with the EQ-5D-3L was estimated to be 0.721 (0.305) indicating some level of disability a year post treatment (mean EQ-5D-3L estimate in the UK general population is 0.860 [34]). A large proportion of patients in the complete case cohort (32.30%) had an EQ-5D-3L index value of 1 (corresponding to full health) with only 0.22% having the lowest EQ-5D-3L index value of -0.594 (corresponding to worst health). The distribution of responses across domains in the EQ-5D-3L instrument is also presented in Table 2. Compared with mobility and self-care, considerably more participants experienced extreme problems (level 3 answers) in the usual activities (8.0%), pain/discomfort (4.2%) and anxiety/depression domains (7.0%).

<< Insert Table 2 around here >>

3.2 Base case analysis

Table 3 presents the comparisons of distributions between the actual and the estimated responses from the domain imputation model. At 5%, 10% and 20% of missing data, there were no evidence of differences between the actual and estimated domain scores ($P > 0.01$). However, at 40% missing data, there was an evidence of discrepancies between the two distributions for the pain/discomfort ($P < 0.01$) and anxiety/depression ($P < 0.01$) domains. Table 4 shows the base case results of estimated index scores when imputation was carried out at domain level and index scores respectively. The accuracies of estimating the actual EQ-5D-3L were almost identical for both imputation models and this was maintained for all proportion of simulated missing data.

<< Insert Table 3 around here >>

<< Insert Table 4 around here >>

3.3 Scenario analyses

Table 5 presents the results of the scenario analysis when the proportion of unit non-response was varied. Although the RMSE remained similar between domain and index imputation for any proportion of missing data, the results of the MAE suggests that when reducing the proportion of unit non-response (increasing the proportion of item non-response missing pattern) domain imputation seems to predict the actual EQ-5D-3L better than index imputation, especially at higher percentages of missing data.

<< Insert Table 6 around here >>

Figure 2 shows the results of using different sample sizes in the simulation exercise. It was not possible to estimate domain imputation due to convergence issues for a sample size of 50 for all proportion of missing data and a sample size of 100 for 40% missing data. For sample sizes

over 500 and any proportion of missing data, index and domain imputation produced similar predictions. For sample sizes between 100 and 500, index and domain imputation were similar at 5% and 10% proportion of missing data in terms of prediction accuracy, but index imputation was more accurate for 20% and 40% proportion of missing data.

<< Insert Figure 2 around here >>

Online Appendix 2 reports the results of the simulation exercise when the MI model was misspecified. For all proportions of missing data, domain imputation predicted the EQ-5D-3L index more accurately than index imputation. However, for all proportions of missing data, the prediction errors were higher compared to the simulation exercise results using a correctly specified model (Table 4). For instance, the RMSE and MAE obtained for the 5% missing data in online Appendix 2 are similar to the estimated figures reported in Table 4 for the 40% proportion of missing data.

4. Discussion

To date, EQ-5D-3L is one of the most commonly used health outcome instrument to estimate utility values when conducting economic evaluations of health care technologies. However, researchers are often unsure whether handling of missing data should be carried out at the domain level or at the index score level. In this simulation study we have demonstrated that the effects of imputing missing values using MI at either level were similar when missing data patterns were dominated with unit non-responses and a large sample size dataset was available. In these cases, analysts may find domain imputation superior given the additional information it is able to produce for each individual EQ-5D-3L domain.

The missing data patterns used in the missing data generation in the base case had around 90% unit non-response missing data and were based on patterns observed in real practice across five large randomised controlled trials [20, 25-28]. Similar missing data patterns have been observed in other studies [35, 36]. This reflects that item non-response missingness in EQ-5D-3L data is less common in clinical trials. However, more research is needed from other trials, before this can be safely generalised. Our scenario analysis confirmed that if the pattern of missingness has considerable amount of item non-response, domain imputation is likely to produce more accurate predictions than imputing the index.

Sample size was shown to be an important aspect when deciding about imputing at the index or the domain level. Our scenario analysis suggested that for proportions of missing data of 5% and 10% and a sample size between 100 and 500 both methods yielded similar results. However, for proportions of missing data of 20% and 40%, index imputation was more accurate. This implies that unless a dataset with a large sample size (over 500) is available or the proportion of missing data is relatively small, domain imputation may not be as accurate as index imputation. With smaller sample sizes, domain imputation experienced convergence problems. In fact, we could not estimate ordered logit models for sample sizes below 100. This was expected as it is known that categorical dependent regression may struggle with reduced samples and low frequencies in the dependent variable [37].

Complete information was assumed on all the predictor variables used in the MI model and missing information was only simulated in the EQ-5D-3L domains. However, in real practice missing data will be also present for other variables in the MI model. This can pose problems when selecting covariates for the imputation model, particularly if the trial has not collected enough potential variables. This automatically translates into a problem of misspecification

in the MI model. We showed in the scenario analysis that omitting relevant variables in the model can lead to serious bias predictions for both index and domain imputation.

Unfortunately, researchers rarely have access in the context of clinical trials to a range of covariates complete enough to rule out potential omitted variables in the specification of the MI model. Therefore, it is recommended that MI models are tested for misspecification (at least for omitted variables and functional form) using available econometric tests such as Ramsey-reset test [38].

In the current study, the missing data patterns were imposed using a MAR assumption and the resulting missingness treated as non-informative. However, this assumption may not be representative of real life as there may be systematic, unobservable factors affecting whether EQ-5D-3L domains are missing. For example, those who are very sick are more likely not to complete the EQ-5D-3L questionnaire leading to results being skewed towards more healthy individuals leading to missing values that are NMAR [16]. In these cases, MI inferences are likely to be biased. The analysis of data with missing values that are NMAR are complex because it is necessary to include in the MAR imputation model reasons for dropouts. Recent guidance suggests that researchers should evaluate the sensitivity of the analysis to the MAR assumption using methods such as the weighting or pattern mixture approaches [7, 39]. Therefore future users imputing missing index or domain EQ-5D-3L data should evaluate how the results vary when inferences are drawn under a NMAR mechanism using one of these methods.

The ordered logit model used for the imputation of the domains relies heavily on the proportional odds assumption, which does not hold completely for the EQ-5D-3L. In the UK value set, a decrement for a level 2 response is less than half of a decrement for a level 3

response for each of the domains. Although we also tested a multinomial logit model to overcome this problem, the multinomial logit model did not converge even for a 5% of simulated missing data and was computationally not possible to execute. This is likely to be the result of the low prevalence of level 3 responses which is a common feature of real life missing EQ-5D-3L data. When using the ordered logit model, 43 out of the 1,000 simulations in the 40% missing data scenario did not converge. These cases were excluded from the analysis and the simulation was continued until all 1,000 converged cases were reached. Overall, convergence was not an important issue when using the ordered logit model in our cross-sectional dataset. However, it is expected that convergence may play a much more important role for this model when the imputation model includes longitudinal data. In addition missing data patterns observed in longitudinal studies are also substantially monotone and future studies should evaluate how convergence and a monotone missing pattern may affect our results.

This study concentrates on the EQ-5D-3L measure and our results may not be generalizable to the more recent five-level version of the EQ-5D (EQ-5D-5L) [40] or other multi-attribute outcome measures. In the EQ-5D-5L, domain imputation may experience further convergence issues and will be more computationally demanding, which raises the question of whether the additional computational time could offset any improvement in the estimation of parameters. HRQoL measures with more additional domains or levels than the EQ-5D-3L are subject to specific psychometric properties and subject to different missing data patterns in comparison to those evaluated in this study and will need to be assessed in separate methodological evaluations. Nevertheless, our results are likely to be

generalisable to instruments similar to the EQ-5D-3L in terms of number of domains and levels and with patterns of missingness similar to the ones observed in our study.

5. Conclusions and recommendations

In summary, we have shown that the decision between imputing the domains or the EQ-5D-3L index scores greatly depends on the observed missing data pattern and the sample size available for analysis. In large sample sizes ($n > 500$) with a primary missing data pattern of unit non-response, imputing domains or the index produces similar results. Domain imputation becomes more accurate when the pattern of missingness is dominantly item non-response, whereas index imputation is likely to be more accurate for smaller samples. Analyst conducting this type of exercises should also evaluate the sensitivity of the analysis to the missing at random assumption and whether the imputation model is correctly specified.

Acknowledgment

We are indebted to the International Subarachnoid Aneurysm Trial (ISAT) Collaborative Group for providing the data for this methodological work. ISAT was supported by grants from: The Medical Research Council, UK; Programme Hospitalier de Recherche Clinique 1998 of the French Ministry of Health (AOM 98150) sponsored by Assistance Publique-Hôpitaux de Paris (AP-HP); the Canadian Institutes of Health Research; and the Stroke Association of the UK. An early version of this paper was presented in the 83rd Health Economists' Study Group (HESG) at the University of Warwick and we are grateful to Lazaros Andronis for discussing the manuscript and providing feedback and useful suggestions. Claire Simons is supported by a NIHR Research Methods Fellowship (ref. MET-12-15).

Ethical Standards

All human studies have been approved by the appropriate ethics committee and have therefore been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. All persons gave informed consent prior to their inclusion in the ISAT study.

References

1. Briggs, A., T. Clark, J. Wolstenholme, and P. Clarke. Missing....presumed at random: Cost-analysis of incomplete data. *Health Economics*, 2003. **12**(5): 377-392.
2. Manca, A. and S. Palmer. Handling missing data in patient-level cost-effectiveness analysis alongside randomised clinical trials. *Appl Health Econ Health Policy*, 2005. **4**(2): 65-75.
3. Burton, A., L.J. Billingham, and S. Bryan. Cost-effectiveness in clinical trials: using multiple imputation to deal with incomplete cost data. *Clin Trials*, 2007. **4**(2): 154-61.
4. Grieve, R., J. Cairns, and S.G. Thompson. Improving costing methods in multicentre economic evaluation: the use of multiple imputation for unit costs. *Health Economics*, 2010. **19**(8): 939-954.
5. Oostenbrink, J.B. and M.J. Al. The analysis of incomplete cost data due to dropout. *Health Economics*, 2005. **14**(8): 763-776.
6. Yu, L.M., A. Burton, and O. Rivero-Arias. Evaluation of software for multiple imputation of semi-continuous data. *Stat Methods Med Res*, 2007. **16**(3): 243-58.
7. Faria, R., M. Gomes, D. Epstein, and I.R. White. A Guide to Handling Missing Data in Cost-Effectiveness Analysis Conducted Within Randomised Controlled Trials. *Pharmacoeconomics*, 2014. **Online First**.
8. Little, R.J. and D.B. Rubin. *Statistical Analysis with Missing Data*. 2nd ed. Wiley Series in Probability and Statistics. 2002, Hoboken, NJ: Wiley.
9. Noble, S.M., W. Hollingworth, and K. Tilling. Missing data in trial-based cost-effectiveness analysis: the current state of play. *Health Econ*, 2012. **21**(2): 187-200.
10. Wood, A.M., I.R. White, and S.G. Thompson. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials*, 2004. **1**(4): 368-76.
11. Eekhout, I., R.M. de Boer, J.W.R. Twisk, H.C.W. de Vet, and M.W. Heymans. Missing Data: A Systematic Review of How They Are Reported and Handled. *Epidemiology*, 2012. **23**(5): 729-732.
12. EuroQol, G. EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy*, 1990. **16**: 199-208.
13. Brooks, R. EuroQol: the current state of play. *Health Policy*, 1996. **37**(1): 53-72.
14. Horsman, J., W. Furlong, D. Feeny, and G. Torrance. The Health Utilities Index (HUI): concepts, measurement properties and applications. *Health Qual Life Outcomes*, 2003. **1**: 54.
15. National Institute for Health and Care Excellence. *Guide to the methods of technology appraisal*. 2013, London: National Institute for Health and Care Excellence.

16. Ratcliffe, J., T. Young, L. Longworth, and M. Buxton. An assessment of the impact of informative dropout and nonresponse in measuring health-related quality of life using the EuroQol (EQ-5D) descriptive system. *Value Health*, 2005. **8**(1): 53-8.
17. Blough, D.K., S. Ramsey, S.D. Sullivan, and R. Yusen. The impact of using different imputation methods for missing quality of life scores on the estimation of the cost-effectiveness of lung-volume-reduction surgery. *Health Economics*, 2009. **18**(1): 91-101.
18. Szende, A., M. Oppe, and N. Devlin. EQ-5D value sets: inventory, comparative review and user guide, ed. A. Szende, M. Oppe, and N. Devlin. 2007, Dordrecht: Springer.
19. StataCorp. Stata Statistical Software. 2011, Stata Press: College Station, TX: StataCorp LP.
20. Molyneux, A., R. Kerr, I. Stratton, P. Sandercock, M. Clarke, J. Shrimpton, and R. Holman. International Subarachnoid Aneurysm Trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: a randomised trial. *Lancet*, 2002. **360**(9342): 1267-1274.
21. Jennett, B. and M. Bond. Assessment of outcome after severe brain damage. *Lancet*, 1975. **1**(7905): 480-484.
22. van Swieten, J.C., P.J. Koudstaal, M.C. Visser, H.J. Schouten, and J. van Gijn. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke*, 1988. **19**(5): 604-7.
23. EuroQol Group. (2014). Available from <http://www.euroqol.org/>. [Accessed 14 September 2014].
24. Van Buuren, S., J.P.L. Brand, C.G.M. Groothuis-Oudshoorn, and D.B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 2006. **76**(12): 1049-1064.
25. Fairbank, J., H. Frost, J. Wilson-MacDonald, L.M. Yu, K. Barker, R. Collins, and G. Spine Stabilisation Trial. Randomised controlled trial to compare surgical stabilisation of the lumbar spine with an intensive rehabilitation programme for patients with chronic low back pain: the MRC spine stabilisation trial. *BMJ*, 2005. **330**(7502): 1233.
26. K. A. T. Trial Group, L. Johnston, G. MacLennan, K. McCormack, C. Ramsay, and A. Walker. The Knee Arthroplasty Trial (KAT) design features, baseline characteristics, and two-year functional outcomes after alternative approaches to knee replacement. *J Bone Joint Surg Am*, 2009. **91**(1): 134-41.
27. Rivero-Arias, O., A. Gray, H. Frost, S.E. Lamb, and S. Stewart-Brown. Cost-Utility Analysis of Physiotherapy Treatment Compared With Physiotherapy Advice in Low Back Pain. *Spine*, 2006. **31**(12): 1381-1387.
28. Kendrick, T., L. Simons, L. Mynors-Wallis, A. Gray, J. Lathlean, R. Pickering, S. Harris, O. Rivero-Arias, K. Gerard, and C. Thompson. Cost-effectiveness of referral for generic care or problem-solving treatment from community mental health nurses, compared with usual general practitioner care for common mental disorders: Randomised controlled trial. *Br J Psychiatry*, 2006. **189**: 50-9.

29. Horton, N.J. and K.P. Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat*, 2007. **61**(1): 79-90.
30. Harel, O. and X.H. Zhou. Multiple imputation: review of theory, implementation and software. *Stat Med*, 2007. **26**(16): 3057-77.
31. White, I.R., P. Royston, and A.M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*, 2011. **30**(4): 377-99.
32. Benjamini, Y. and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1995. **57**(1): 289-300.
33. Efron, B. 1977 Rietz Lecture - Bootstrap Methods - Another Look at the Jackknife. *Annals of Statistics*, 1979. **7**(1): 1-26.
34. Kind, P., G. Hardman, and S. Macran. UK population norms for EQ-5D. 1999, Centre for Health Economics, University of York, UK.
35. Janssen, M.F., A.S. Pickard, D. Golicki, C. Gudex, M. Niewada, L. Scalone, P. Swinburn, and J. Busschbach. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Qual Life Res*, 2013. **22**(7): 1717-27.
36. Konig, H.H., A. Born, O. Gunther, H. Matschinger, S. Heinrich, S.G. Riedel-Heller, M.C. Angermeyer, and C. Roick. Validity and responsiveness of the EQ-5D in assessing and valuing health status in patients with anxiety disorders. *Health Qual Life Outcomes*, 2010. **8**: 47.
37. Long, J.S. Regression models for categorical and limited dependent variables. 1997, London: Sage.
38. Ramsey, J.B. Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 1969. **31**(2): 350-371.
39. Carpenter, J.R., M.G. Kenward, and I.R. White. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Stat Methods Med Res*, 2007. **16**(3): 259-75.
40. Herdman, M., C. Gudex, A. Lloyd, M. Janssen, P. Kind, D. Parkin, G. Bonsel, and X. Badia. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*, 2011. **20**(10): 1727-36.

Figure 1: The design of the simulation study

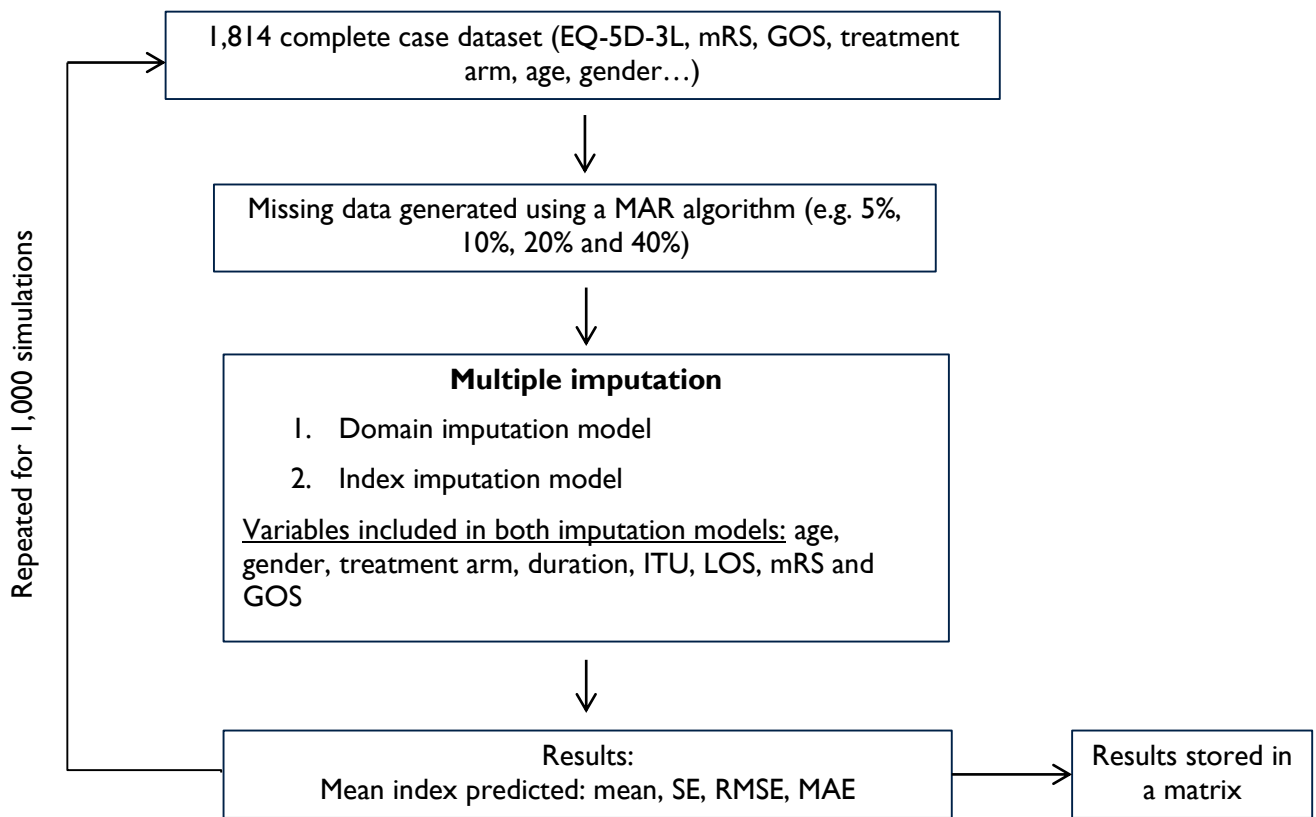


Table 1: Missing data patterns used in the missing data generation

Missing Pattern	Total	True %	% Used in simulation	Cumulative %
● ● ● ● ●	4673	88.65	88.70	88.7
- - - - ●	175	3.32	3.30	92.0
- - - ● -	104	1.97	2.20	94.2
- - ● - -	95	1.80	2.00	96.2
- ● - - -	75	1.42	1.70	97.9
● - - - -	60	1.14	1.20	99.1
- - - ● ●	39	0.74	0.90	100.0

● indicates that a domain is missing, - indicates that a domain has been completed.

Table 2: Complete case cohort summary statistics (n = 1,814)

Variable	Mean	SD	Minimum	Maximum
Age	51.33	11.32	18.30	86.79
Procedure time (minutes)	173.54	79.76	0.00	595.00
ITU days	3.55	6.60	0.00	63.00
Hospital length stay	18.98	15.44	4.00	315.00
EQ-5D-3L index	0.721	0.305	-0.594	1.000
	n	%		
Gender				
Male	675	37.2%		
Female	1,139	62.8%		
Treatment arm				
Endovascular	930	51.3%		
Neurosurgery	884	48.7%		
Centre Size				
Small	76	4.2%		
Medium	387	21.3%		
Large	1,351	74.5%		
Glasgow Outcome Score				
1	990	54.6%		
2	471	26.0%		
3	296	16.3%		
4	57	3.1%		
Modified Rankin Score				
0	424	23.4%		
1	566	31.2%		
2	470	25.9%		
3	232	12.8%		
4	66	3.6%		
5	56	3.1%		
EQ-5D-3L				
<i>Mobility</i>				
1	1,263	69.6%		
2	519	28.6%		
3	32	1.8		
<i>Self-care</i>				
1	1,507	83.1%		
2	250	13.8%		
3	57	3.1%		
<i>Usual activities</i>				
1	1,047	57.7%		
2	623	34.3%		
3	144	8.0%		
<i>Pain/discomfort</i>				
1	959	52.9%		
2	778	42.9%		
3	77	4.2%		

Anxiety/Depression

1	902	49.7
2	785	43.3
3	127	7.0

Table 3: EQ-5D-3L domain prediction accuracy after multiple imputation for 1,000 simulations*

Mobility			Self-Care		Usual Activities		Pain/ Discomfort		Anxiety/ Depression	
Actual	Estimated		Actual	Estimated	Actual	Estimated	Actual	Estimated	Actual	Estimated
5% missing data										
1	1,263	1,267	1,507	1,506	1,047	1,042	959	953	902	896
2	519	516	250	252	623	631	778	789	785	798
3	32	31	57	56	144	141	77	72	127	120
χ ² ,p	0.06, p=0.98		0.03, p=0.98		0.19, p=0.99		0.52, p=0.98		0.64, p=0.98	
10% missing data										
1	1,263	1,271	1,507	1,506	1,047	1,038	959	946	902	890
2	519	513	250	253	623	639	778	801	785	812
3	32	30	57	55	144	137	77	67	127	112
χ ² ,p	0.25, p=0.98		0.11, p=0.98		0.83, p=0.98		2.16, p=0.76		2.86, p=0.60	
20% missing data										
1	1,263	1,278	1,507	1,505	1,047	1,028	959	932	902	878
2	519	509	250	255	623	654	778	824	785	839
3	32	27	57	54	144	132	77	58	127	97
χ ² ,p	1.15, p=0.98		0.26, p=0.98		2.89, p=0.60		8.17, p=0.07		11.44, p=0.02	
40% missing data										
1	1,263	1,292	1,507	1,504	1,047	1,013	959	905	902	853
2	519	499	250	261	623	679	778	869	785	894
3	32	23	57	49	144	122	77	40	127	67
χ ² ,p	3.97, p=0.46		1.61, p=0.89		9.50, p=0.04		31.46, p<0.01		46.14, p<0.01	

Bold figures identify significant differences between actual and predicted. Chi-square test controlled for false discovery rate.

*Frequencies estimated averaging frequencies for each domain across the 10 imputed datasets for each patient; and then averaging the resulting frequencies for each domain and patient across the 1,000 simulations.

Table 4: Base case results and imputation model comparison between imputing the EQ-5D-3L domain and index

	Mean	SE	RMSE	MAE
Actual EQ-5D-3L index	0.7212	0.0072		
5% Missing				
Domain imputation	0.7212	0.0073	0.0073	0.0009
Index imputation	0.7213	0.0073	0.0073	0.0009
10% Missing				
Domain imputation	0.7212	0.0074	0.0074	0.0012
Index imputation	0.7214	0.0073	0.0073	0.0013
20% Missing				
Domain imputation	0.7210	0.0076	0.0076	0.0019
Index imputation	0.7214	0.0076	0.0076	0.0021
40% Missing				
Domain imputation	0.7205	0.0084	0.0084	0.0035
Index imputation	0.7219	0.0082	0.0082	0.0039

SE: Standard error; RMSE: root mean-squared error; MAE: mean absolute error

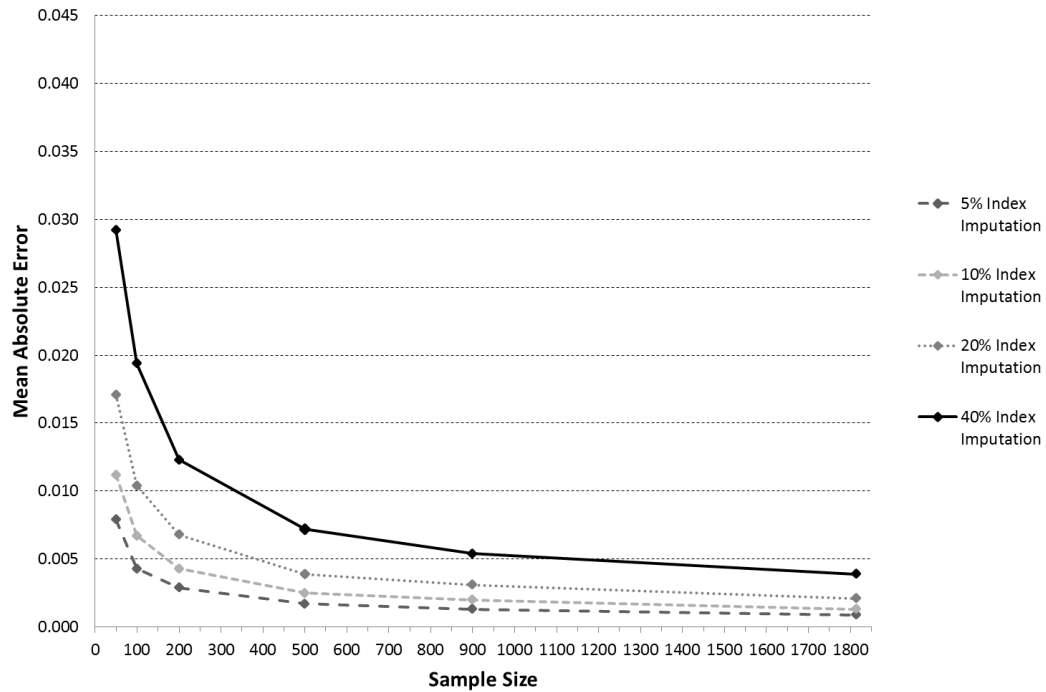
Table 5: Scenario analysis on the impact of reducing unit non-response in the missing data pattern generation

Proportion of unit non-response	5% Missing		10% Missing		20% Missing		40% Missing	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Base Case (~90%)								
Domain imputation	0.0073	0.0009	0.0074	0.0012	0.0076	0.0019	0.0084	0.0035
Index imputation	0.0073	0.0009	0.0073	0.0013	0.0076	0.0021	0.0082	0.0039
60%								
Domain imputation	0.0072	0.0008	0.0073	0.0011	0.0075	0.0017	0.0080	0.0027
Index imputation	0.0072	0.0009	0.0073	0.0014	0.0076	0.0021	0.0082	0.0037
40%								
Domain imputation	0.0072	0.0007	0.0073	0.0010	0.0074	0.0014	0.0076	0.0022
Index imputation	0.0073	0.0009	0.0073	0.0014	0.0076	0.0020	0.0082	0.0038
20%								
Domain imputation	0.0072	0.0006	0.0072	0.0008	0.0073	0.0011	0.0075	0.0018
Index imputation	0.0072	0.0010	0.0073	0.0013	0.0076	0.0021	0.0082	0.0037
0%								
Domain imputation	0.0072	0.0004	0.0072	0.0006	0.0073	0.0009	0.0074	0.0013
Index imputation	0.0073	0.0009	0.0073	0.0014	0.0076	0.0021	0.0082	0.0038

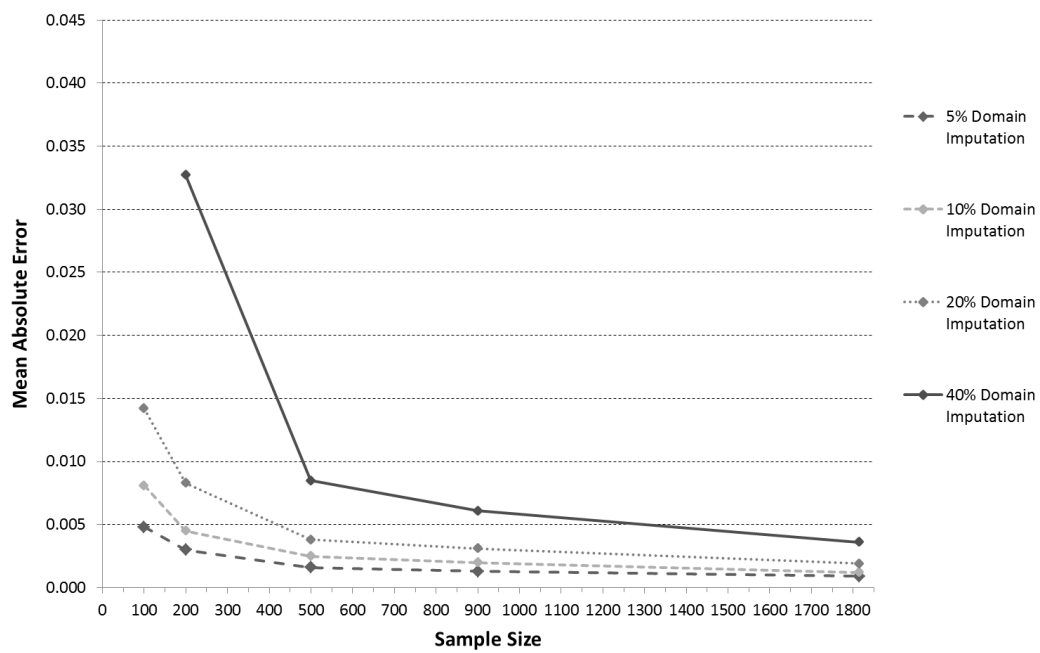
RMSE: root mean-squared error; MAE: mean absolute error

Figure 2: Mean absolute error of different scenarios of sample size and simulated missing data for index (a) and domain imputation (b).

(a)



(b)



Online appendix 1 – Algorithm for simulating missing at random data

The following step-by-step guidance is based on the implementation by van Buuren *et al.* [24] subsequently also used in Yu *et al.* [6].

1. A random variable, $X \sim U[0,1]$ was assigned to each patient and compared against the cumulative probability of being in each of the seven patterns ($MPAT_i = 1, \dots, 7$).
2. A linear score was calculated for each patient using the variables thought to influence whether a patient had missing data. The weights for the linear score were calculated as the coefficients of an OLS regression of the variables on each of the five EQ-5D domains in turn. If a missing pattern included more than one missing domain the weights were a sum of the regression coefficients of the domains missing in that pattern.
3. The sample of the cases allocated to the i^{th} response pattern $MPAT_i = 1, \dots, 7$ were divided into three subgroups $j=1,2,3$ by their linear scores with the 33% and 66% quantiles of the linear scores being the cut-off points, giving N_j cases in each subgroup.
4. Increasing odds (O_j) of having the i^{th} response pattern $MPAT_i$ for the three subgroups were specified as 1, 2 or 3.
5. The probability of having the i^{th} response for each case was calculated as

$$prob(miss) = \frac{1,814 \times O_j \times P \times f_i}{N_i \sum_{j=1}^3 O_j}$$

With O_j , P , N_i being as previous and f_i the proportion of patients having missing patterns i .

6. Another random variable, $Y \sim U[0,1]$ was assigned to each patient and compared against the probability calculated in step five. Those patients with $Y < prob(miss)$ had their observations set to be missing according to the pattern of missingness allocated in step one.

Online appendix 2: Scenario analysis assuming MI model is not correctly specified (omitted relevant variables) in the simulation exercise

	Mean	SE	RMSE	MAE
Actual EQ-5D-3L index	0.7212	0.0072		
5% Missing				
Domain imputation	0.7243	0.0073	0.0079	0.0031
Index imputation	0.7249	0.0073	0.0082	0.0037
10% Missing				
Domain imputation	0.7275	0.0075	0.0098	0.0063
Index imputation	0.7289	0.0074	0.0107	0.0077
20% Missing				
Domain imputation	0.7350	0.0078	0.0158	0.0137
Index imputation	0.7383	0.0077	0.0188	0.0171
40% Missing				
Domain imputation	0.7558	0.0086	0.0356	0.0345
Index imputation	0.7673	0.0082	0.0467	0.0460

SE: Standard error; RMSE: root mean-squared error; MAE: mean absolute error