ARE COST-EFFECTIVENESS RESULTS SENSITIVE TO THE CHOICE OF IMPUTATION METHOD? A COMPARISON OF MULTIPLE IMPUTATION APPROACHES FOR CLUSTERED DATA.

KARLA DÍAZ-ORDAZ, MANUEL GOMES, MICHAEL G. KENWARD, AND RICHARD GRIEVE

ABSTRACT. Missing data are common in cost-effectiveness analysis (CEA), and methodological guidelines recommend that this issue is addressed with multiple imputation (MI). However, for MI to provide valid inferences, the imputation model must recognise the study design. Specifically, CEA tend to use data from studies with hierarchical designs e.g. cluster randomised trials (CRTs), but then ignore the multilevel structure of the missing data.

Aims To compare under missing at random assumption: (a) Multilevel MI that accounts for clustering through cluster random effects, (b) MI that includes a fixed effect for each cluster and (c) single-level MI that ignores clustering.

Methods We introduce the different MI approaches and illustrate them with a motivating example: a CEA that uses data from a CRT evaluating alternative interventions for postnatal depression (2659 participants, 100 clusters). Here, the level of clustering was relatively high for cost but lower for QALYs (intra-cluster correlation coefficient, $\text{ICC}_1 = 0.17$, and $\text{ICC}_2 = 0.04$ respectively). We then conducted a simulation study to assess the relative performance of the alternative MI methods. Informed by the methodological literature, missing data scenarios were simulated according to those factors anticipated to have an impact on the methods' performance, for example the levels of clustering, the number and size of clusters and the proportion of observations with missing values. For each scenario, we created 1000 datasets and obtained estimates of incremental cost, QALYs and incremental net benefit (INB) for each MI approach. We obtained mean levels of bias and confidence interval (CI) coverage, amongst other metrics.

Results In the case-study, the estimates (SE) of the INB were $£99.93\ (40.87)$ for the multilevel MI, $£60.11(39.73)$ for the fixed effects MI, and $£80.47(34.35)$ for the single-level MI. The simulation reported that all methods provided unbiased estimates in each scenario considered. In the simulations with low ICCs (0.01), the fixed effects MI over-estimated the SEs, resulting in CI coverage (average 98%) in excess of nominal levels, whereas with relatively high ICCs (0.2) , the single level approach resulted in low coverage (average 86%). The multilevel MI reported coverage levels of approximately 95% throughout.

Conclusion Cost-effectiveness results can differ according to the MI method. The multilevel MI approach performed better than the other MI methods considered across our settings and is appropriate for studies that have a hierarchical design.

Key words and phrases. missing data, multiple imputation, bivariate models, cost-effectiveness.
email for correspondence: karla.diaz-ordaz@lshtm.ac.uk
Address for correspondence: Department of Health Services Research and Policy,
London School of Hygiene and Tropical Medicine, 15 - 17 Tavistock Place, London WC1H 9SH, UK
Paper presented at the winter HESG 2013 .

1. INTRODUCTION

Cost-effectiveness analyses (CEA) are often used to help health policy-makers decide which health care programmes to prioritise (NICE, 2008; CADTH, 2006; PBCA, 2008). The use of randomised controlled trial data in CEA requires that such studies use appropriate statistical methods (Gold et al., 1996; Willan & Briggs, 2006; Glick et al., 2007; Gray et al., 2010). This includes the correct handling of missing data.

Missing data are common. For example, participants are lost to follow-up or fail to respond to quality-of-life (QOL) or resource use questionnaires. The main problem with missing data is that individuals with missing costs or health outcomes may be systematically different from those with complete data. This is known to lead to bias (Briggs et al., 2003). However, most CEA that use individual-level data have missing observations and report only complete-case analyses (Noble et al., 2012).

Multiple imputation (MI) has been proposed for handling missing data in CEA (Briggs et al., 2003; Blough et al., 2009; Oostenbrink & Al, 2005). MI is an attractive approach as it allows variables which are associated with the endpoints and predictive of missingness to be included in the imputation model, without modifying the analysis model. This offers an important advantage over likelihood-based approaches, because including such variables in the imputation model can help reduce bias and improve efficiency. MI can also offer substantial gains in precision when compared with generic inverse probability weighting approaches (Carpenter et al., 2006).

MI treats missing data as an explicit source of random variability and incorporates this uncertainty explicitly, by accounting for the between-imputation variability. However, to provide valid inferences, the imputation method must appropriately recognise the structure of the data. For example, in CEA that use cluster randomised trials (CRTs), the probability of missing costs or health outcomes may be more similar within than across clusters. Here, missingness may depend on individual-level characteristics, which tend to be more similar within the cluster, and on cluster-level variables, such as the size of the cluster. Failure to account for clustering appropriately when addressing the missing data may result in misleading cost-effectiveness estimates (Gomes et al., 2012a).

Multilevel MI recognises explicitly the clustering by including cluster-specific random effects, and has recently been illustrated in CEA that use CRTs (Diaz Ordaz et al., 2012; Gomes et al., 2012a). An alternative approach is to introduce each cluster as a fixed-effect, using a dummy variable for each cluster in the imputation model (Graham, 2009).

Commentators suggest that the choice between random and fixed effects is conditional on the data and context. For example, in econometric evaluation of health policies, fixed effects are often used because inferences may be at the cluster-level, for example, predicting region-specific effects on health expenditures (Wooldridge, 2002; Jones & Rice, 2011). In CEA, random effects approach is preferred because the primary interest is on the overall average treatment effect (ATE). In the missing data context, the appropriate choice of imputation model must reflect the conditional distribution of the missing data given the observed. If the substantive model

has fixed cluster effects, i.e. inferences are confined to within-cluster comparisons, then the imputation model may legitimately use fixed cluster effects. Otherwise random cluster effects are appropriate. In addition, fixed-effects may be less precise than random effects, especially when the between-cluster variability is low, as typical for health outcomes, and the size of the clusters is small (Andridge, 2011).

The aim of this paper is to compare the effect of multilevel (random-effects) MI and fixed-effects MI with a MI approach that ignores clustering (single-level MI) for handling missing hierarchical data in cost-effectiveness analyses that use random-effects models to estimate ATEs. We extend previous studies (Diaz Ordaz et al., 2012; Gomes et al., 2012a) by comparing the relative performance of the alternative MI methods using both an empirical example and a simulation study. The case study is typical of a CEA alongside of a CRT, and evaluates an intervention for postnatal depression. The methods are compared in the context of CEA alongside CRTs, but they apply more generally to other hierarchical settings such as longitudinal trials.

In the next section, we consider in some detail the handling of missing data and introduce the alternative MI methods. In Section 3, we present the motivating example and compare the results from the alternative MI strategies. In Section 4, we describe the simulation study and report a selection of results from simulated scenarios. We close with a discussion in Section 5.

2. METHODS

2.1. **Analysis model.** We focus on estimating ATEs, as these are of prime interest for policy makers (Imbens & Wooldridge, 2009), assuming linear additive treatment effects for both costs and QALYs, with no adjustment for baseline covariates. We consider previously proposed bivariate Normal multilevel models (MLM) for CEA that use CRTs (Grieve et al., 2010; Gomes et al., 2012c).

Let C_{ij} and Q_{ij} be the cost and QALY outcomes respectively from the i-th individual in cluster j of a two-arm CEA alongside a CRT. Let treatment allocation be represented by $t_j = 1$, if the cluster was allocated to intervention, and 0 otherwise. MLMs take into account the clustering by including additional cluster-level random effects, represented by the latent variables u_j^c and u_j^q. The model can be written as follows

$$\begin{aligned} C_{ij} &= \beta_0^c + \beta_1^c t_j + u_j^c + e_{ij}^c \\ Q_{ij} &= \beta_0^q + \beta_1^q t_j + u_j^q + e_{ij}^q \end{aligned} \tag{1}$$

with β_1^ℓ representing the treatment effect on the corresponding outcome $\ell = \{C, Q\}$. The error term (e_{ij}^c, e_{ij}^q) and the cluster effects are assumed to be Normally distributed:

$$\begin{pmatrix} e_{ij}^c \\ e_{ij}^q \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_q \\ \rho\sigma_c\sigma_q & \sigma_q^2 \end{pmatrix}\right] \text{ and } \begin{pmatrix} u_j^c \\ u_j^q \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_c^2 & \phi\tau_c\tau_q \\ \phi\tau_c\tau_q & \tau_q^2 \end{pmatrix}\right],$$

where σ_c, σ_q are the individual-level standard errors, ρ is the individual-level correlation between costs and outcomes and τ_c, τ_q, and ϕ are the standard errors and correlation of the two cluster random effects, respectively.

2.2. **Missing Data.** In this section, we give a brief description of multiple imputation for clustered data. Let Y_{ijk} be the partially observed response for the i-th individual in cluster j receiving treatment $k = 0, 1$ and let X_{ijk} represent the set of fully-observed covariates (including individual and cluster-level variables) and b_{jk} the random cluster-effects. Let $R_{ijk} = 1$ if the corresponding response is observed and $R_{ijk} = 0$ if it is missing (i.e. R is the non-response indicator).

We represent the missing data mechanism by

$$Pr(R_{ijk} = 0|Y_{ijk}, \theta, b_{jk}) = \pi_{ijk},$$

i.e. the conditional probability of not observing the response where θ denotes certain unknown parameters (Taljaard et al., 2008).

When data are missing, additional assumptions must be made about the missing data mechanism. One common assumption is that the data are *Missing At Random* (MAR), i.e. conditional on the observed variables such as age and gender, R_{ijk} and Y_{ijk} are independent. This is analogous to the "selection on observables" or ignorability assumption, where the fact that an individual observation is observed or missing depends only on observed factors. If R_{ijk} and Y_{ijk} are unconditionally independent, then the data are said to be *Missing Completely at Random* (MCAR). By contrast, if conditional on the observed data, the probability of missingness is associated with unobserved factors, we say that the missing data are *Missing Not at Random* (MNAR), equivalent to the non-ignorability assumption. It is impossible to rule out MNAR from the data at hand, because this depends on the existence of associations between *unobserved* information and the R_{ijk}. For example, individuals with low income may be more likely to drop out from a given study, but socio-economic variables may not have been collected.

2.3. **Multiple Imputation.** Appropriate techniques for handling missing values are based on the assumptions made about the missingness mechanism. Under MCAR and MAR, valid estimates are obtained via likelihood-based analysis without explicit assumptions about the form of the missing data process, provided all covariates associated with the missing data mechanism are included in the analysis model (Molenberghs & Kenward, 2007).

Although the analysis model used here is likelihood-based, it is unlikely that estimates based on complete-cases are valid, as there may be variables associated with the missing data mechanism which are not included in the model. Multiple imputation represents a practical solution. An important source of information that can be used to reduce possible bias is contained in observed variables that are associated with the outcome and with the missing data mechanism itself. If these variables are not part of the analysis model, they are termed *auxiliary* variables. Once auxiliary variables have been identified, it is necessary to incorporate them into the analysis. We do this through multiple imputation (MI) (Rubin, 1978). This has the advantage of retaining the original analysis model, adding to this an *imputation model*. The imputation model is a regression model

of the variables to be imputed on the auxiliary variables and all other covariates included in the analysis model.

Let $Y_{1,ijk}$ and $Y_{2,ijk}$ be two partially observed continuous variables and let $\mathbf{X_{ijk}}$ denote the matrix of all auxiliary variables (assumed to be fully observed), including individual and cluster-level variables. The imputation models compared here express $(Y_{1,ijk}, Y_{2,ijk})$ as a function of the grand mean in treatment arm k ($\nu_{\ell,0k}$, for $\ell = \{1, 2\}$), the auxiliary variables, and error terms $(e_{1,ijk}, e_{2,ijk})$.

The single-level imputation model ignores clustering (denoted SMI):

$$\begin{aligned} Y_{1,ijk} &= \nu_{1,0k} + \mathbf{X_{ijk}}\nu_{1,X} + e_{1,ijk} \\ Y_{2,ijk} &= \nu_{2,0k} + \mathbf{X_{ijk}}\nu_{2,X} + e_{1,ijk} \end{aligned} \qquad \begin{pmatrix} e_{1,ijk} \\ e_{2,ijk} \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Omega_1})$$

where $\nu_{\ell,X}$ is the vector of regression coefficients, and $\mathbf{\Omega_1}$ is the individual-level variance-covariance matrix. With single-level MI, the imputed values are drawn from the conditional distribution of the missing observations given the observed data, ignoring any dependency between observations within a cluster not explained by the cluster-level auxiliary variables included in the model. Therefore, the single-level imputation model does not properly represent the conditional distribution of the missing data given the observed data.

Two models have been proposed to account for the clustering. The fixed effects imputation model (denoted FMI):

$$\begin{aligned} Y_{1,ijk} &= \nu_{1,0k} + \mathbf{X_{ijk}}\nu_{1,X} + \sum_{j=1}^{J-1} \beta_{1,j} I_{ij} + e_{1,ijk} \\ Y_{2,ijk} &= \nu_{2,0k} + \mathbf{X_{ijk}}\nu_{2,X} + \sum_{j=1}^{J-1} \beta_{2,j} I_{ij} + e_{2,ijk} \end{aligned}$$

where I_{ij} is the indicator variable for cluster j, so that $I_{ij} = 1$ if the observation i belongs to cluster j and the error term $(e_{1,ijk}, e_{2,ijk})$ is assumed to be bivariate normal as before. This model allows a different intercept for each cluster within treatment group k. Missing outcomes will be imputed from the conditional normal distribution given the other endpoint, if observed, and the auxiliary variables, which must all be at the individual level, with a mean determined by the fixed-effect for that cluster.

The random-effects imputation model (denoted MMI):

$$\begin{aligned} Y_{1,ijk} &= \nu_{1,0k} + \mathbf{X_{ijk}}\nu_{1,X} + b_{1,j} + e_{1,ijk} \\ Y_{2,ijk} &= \nu_{2,0k} + \mathbf{X_{ijk}}\nu_{2,X} + b_{2,j} + e_{2,ijk} \end{aligned} \qquad \begin{pmatrix} b_{1,j} \\ b_{2,j} \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Omega_2})$$

where $\mathbf{\Omega_2}$ is the cluster-level variance-covariance matrix and the individual-level residuals $(e_{1,ijk}, e_{2,ijk})$ are assumed normally distributed independently of $(b_{1,j}, b_{2,j})$, the cluster random-effects.

2.4. **MI procedure.** Given the analysis and imputation models, conventional MI procedures can be followed. These are set out in detail in many references, including Little & Rubin (2002) and

Molenberghs & Kenward (2007). The overall MI procedure, done separately in each treatment arm, is as follows.
(1) The imputation model is fitted to the observed data and Bayesian draws are taken from the posterior distribution of the model parameters.
(2) Missing data are imputed from the imputation model, using the parameters drawn in step (1).
(3) The analysis model is fitted (here using maximum likelihood) to the data set that has been *completed* using the imputations from step (2), producing parameter estimates and their estimated covariance matrix.
(4) Steps (1)-(3) are repeated a fixed M number of times.
(5) The M sets of parameter and covariance estimates from step (3) are then combined using Rubin's formulae (Rubin, 1987) to produce a single MI estimate of the substantive model parameters and associated covariance matrix.

Under the MAR assumption, this will produce consistent estimators and in the absence of auxiliary variables, is asymptotically (as M increases) equivalent to maximum likelihood.

Sampling from the approximate predictive distribution of the missing data as described above can be performed in a variety of widely available MI software. The algorithms used in different packages represent two very different theoretical and computational arguments; the first approach jointly models variables subject to missingness, by sampling from the underlying predictive distribution under multivariate normality (Schafer, 1997; Carpenter et al., 2011). The second approach, called fully-conditional specification (FCS), also known as chained equations (White et al., 2011), approximates the joint distribution with a variable-by-variable approach, by specifying an imputation model per variable.In this work, the MICE package in R (van Buuren & Groothuis Oudshoorn, 2011), based on FCS, is used for the fixed-effects and single-level imputations. The multilevel imputations are performed with the PAN package, which is based on joint multivariate normal modelling. Details on the MCMC imputation method used in PAN can be found in Schafer & Yucel (2002). As both implementations assume that continuous variables are drawn from a multivariate normal distribution, it is advised to transform skewed variables to approximate normality (Briggs et al., 2003; Yu et al., 2007) before imputation and then transform back before fitting the analysis model. This has been shown in simulations to help obtain reliable estimates of associations with that (skewed) variable and to avoid bias in inferences for other variables (Lee & Carlin, 2010).

We obtain $M = 10$ multiply imputed sets under each of the MI approaches, as recommended by Schafer (1999).

3. MOTIVATING EXAMPLE: THE PONDER TRIAL

The PONDER study (psychological interventions for post-natal depression trial and economic evaluation) was a CRT evaluating an intervention for preventing postnatal depression, (Morrell et al., 2009). It included 2659 patients who attended 101 primary care providers in the UK (general practices). Clusters were randomly allocated to provide either usual care (control) or an

TABLE 1. Percentages of missing outcome data and baseline covariates in the PONDER case study, by treatment group.

			Control group (Total n=911)		Intervention group (Total n=1730)	
Outcome variables	type	symbol	Missing n	%	Missing n	%
Cost	continuous	C_{ij}	402	41.1	460	26.6
QALY	continuous	Q_{ij}	39	4.3	59	3.4
Fully-observed baseline variables						
Edinburgh Postnatal Depression Scale	continuous	epds_{ij}				
Ethnicity	binary	eth_{ij}				
Economic status	binary	eco_{ij}				
Age	continuous	age_{ij}				

intervention delivered by a health visitor (treatment). The intervention comprised health visitor training to identify and manage patients with postnatal depression. As is common, the PONDER CRT had an imbalanced design; the number of patients per cluster varied widely (from 1 to 101 in the control group and from 1 to 81 in the treatment group).

Patients were followed up for 18 months with costs (£ sterling) and health-related quality of life (QOL) recorded at six monthly intervals. This paper considers costs and QOL reported at six months. These QOL data were used to adjust life years and present quality-adjusted life years (QALYs) over six months. Let C_{ij} and Q_{ij} denote the cost and QALYs of the i individual in the j cluster, respectively. Intra-cluster correlation coefficients (ICCs) were high for costs ($\text{ICC}_1 = 0.17$) but moderate for QALYs ($\text{ICC}_2 = 0.04$). While QALYs were approximately normally distributed, costs were right-skewed.

Table 1 reports the percentage of observations with missing outcome data, by treatment group. For simplicity, we only consider here fully-observed covariates as potentially auxiliary variables (also Table 1). There were 31 clusters without any observed cost data (15 in the control arm).[1]

The CEA presents incremental QALYs δ_Q, and costs δ_C, as the differences in means, between the treatment and control groups. Cost-effectiveness is then reported as incremental net monetary benefits $\text{INB}(\lambda) = \lambda\delta_Q - \delta_C$, for $\lambda = £20000$, the willingness to pay for a unit of health gain. Its standard error can be calculated from the estimated variances and covariances of $\hat{\delta}_C$ and $\hat{\delta}_Q$ in the usual way (Willan & Briggs, 2006).

3.1. **Multiple imputation for the case-study.** To identify potential auxiliary variables from amongst those observed, it is useful to investigate the associations between these variables and the missing-data indicator $R_{\ell,ijk}$, using logistic regression. This has been done separately here for cost and QALYs and also for each treatment group ($k = 0, 1$). In addition to the patient-level covariates described, we added the cluster-level variable *cluster size,* n_j, defined as the number

[1]There was one cluster that withdrew from the study for which imputation was not carried out.

TABLE 2. Imputation models used for the cost and QALY endpoints in the PONDER case study, by treatment group.

Model	Control Group	Intervention Group
SMI	$(\log C_{ij}, Q_{ij}) = \text{epds}_{ij} + \text{eco}_{ij} + \text{eth}_{ij} + n_j$	$(\log C_{ij}, Q_{ij}) = \text{epds}_{ij} + \text{eco}_{ij} + \text{age}_{ij} + n_j$
FMI	$(\log C_{ij}, Q_{ij}) = \text{epds}_{ij} + \text{eco}_{ij} + \text{eth}_{ij}$	$(\log C_{ij}, Q_{ij}) = \text{epds}_{ij} + \text{eco}_{ij} + \text{age}_{ij}$
MMI	$(\log C_{ij}, Q_{ij}) = \text{epds}_{ij} + \text{eco}_{ij} + \text{eth}_{ij} + n_j$	$(\log C_{ij}, Q_{ij}) = \text{epds}_{ij} + \text{eco}_{ij} + \text{age}_{ij} + n_j$

of participants randomised in each cluster. Previous studies suggest that cluster size may be associated with costs or health outcomes (Gomes et al., 2012b). We also consider that the number of participants recruited in each cluster may be associated with missingness.

Costs are log-transformed before imputation to improve normality. Thus, to simplify the exposition, we restricted our analyses to those individuals with positive costs, by excluding 18 observations with zero costs (15 in the treatment group). We take an inclusive approach, adding all auxiliary variables found to be associated with either endpoint and their missingness, and modelling the two outcomes simultaneously. As the effects of cluster-level variables cannot be estimated simultaneously with cluster fixed-effects, the fixed-effects imputation model does not include cluster size, though effectively all cluster-level characteristics have been accounted for. The imputation models chosen are summarised in Table 2.

3.2. **Results for the case study.** Table 3 shows the multiple imputation estimates for incremental cost and incremental QALYs as well as INB. Notice that for QALYs, which had a very low ICC, all MI methods report similar estimates. However, the choice of MI method has an impact on the estimates for the incremental cost, and therefore on INB.

In particular, fixed-effects imputation results in an estimated positive incremental cost. This may be due to the fact that there were a considerable number of clusters were all costs were missing. A fixed-effect imputation uses the information contained in the individual-level covariates included in the imputation model across all clusters, but estimates the fixed-effect for cluster based only on the other endpoint to estimate the predictive posterior distribution of the variable of interest. It ignores information about other clusters' mean cost and variability, and this may have an impact especially on those clusters where there is no information on cost available.

As expected, single-level MI results in smaller standard errors, symptomatic of an underestimation of the variance. Point-estimates for cost, and hence for INB also differ markedly. This is possibly due to the fact that single-level models weight individuals differently from multilevel models (Grieve et al., 2010), resulting in the imputation models obtaining different estimates of the parameters from which the data are imputed (step (2) in Section 2.4).

4. SIMULATION STUDY

A simulation study was conducted to assess the performance of the MI methods in order to investigate which factors are most influential for the choice of MI method. This study followed a full factorial design (Cox & Reid, 2000).

Based on previous methodological literature (Andridge, 2011; Taljaard et al., 2008; van Buuren, 2012) and the case study, bivariate endpoint data, say cost and health outcome, were simulated according to those factors anticipated to have an impact on the MI methods' performance. Different values of ICCs were included to reflect the levels of clustering seen in practice (Campbell et al., 2005). These were allowed to differ between the two endpoints. Other factors included: the non-response rate (Andridge, 2011) and the strength of association between the covariates and non-response indicator (White & Carlin, 2010). The case study results provided motivation for testing the impact of including cluster-level auxiliary variables, such as cluster size, particularly if they are associated with missingness and endpoints.

We also allow the number and size of cluster to vary (van Buuren, 2012), while maintaining the same overall sample size ($S = 500$) . We consider three main types of CRT design: (i) large number of clusters ($J = 50$) and few individuals per cluster ($n_j = 10$); (ii) small number of clusters ($J = 10$) and large cluster size ($n_j = 50$); (iii) moderate number of clusters (30) and variable number of individuals per cluster (mean $n_j = 20$, with coefficient of variation 0.5).[2] The factors involved in the simulation are summarized in Table 4. The correlation between each

[2]Cluster size n_j was assumed to follow a Gamma distribution according to a mean and a coefficient of variation $cv = \frac{SD(n)}{E(n)}$, where SD and E denote the standard deviation and mean of the cluster size respectively.

TABLE 3. Incremental cost ($\pounds$) and QALYs at 6 months using bivariate normal substantive model and different MI strategies for PONDER.

Estimates	SMI	FMI	MMI
Incremental cost	-13.26 (21.39)	8.29 (29.03)	-29.43 (28.79)
Incremental QALY	0.003 (0.001)	0.003 (0.001)	0.003 (0.001)
INB	80.47 (34.35)	60.11 (39.73)	99.93 (40.87)

TABLE 4. Factors and their chosen levels varying across the different scenarios

Factor	Levels
Missing data mechanism	Individual level covariate logit $\pi^{\ell}_{ijk} = \alpha_0 + \eta_X X_i$
	Cluster level covariate logit $\pi^{\ell}_{ijk} = \alpha_0 + \eta_W W_j$
	Both logit $\pi^{\ell}_{ijk} = \alpha_0 + \eta_X X_i + \eta_W W_j$
Cluster design	$J = 50$ clusters, $n_j = 10$ individuals each
	$J = 10$ clusters, $n_j = 50$ individuals each
	$J = 30$ clusters, variable size (coef variation=0.5)
Association between covariates and non-response indicator	$\eta = 1$
	$\eta = 2$
Non-response rate	20% each endpoint
	30% for $Y_{1,ij}$ and 10% for $Y_{2,ij}$
ICC_1 and ICC_2	(0.01, 0.01)
	(0.20, 0.05)
	(0.20, 0.20)
	(0.60, 0.01)

outcome and covariates X_i and W_j was fixed at 0.5, while the correlation between the endpoints $Y_{1,ijk},\ Y_{2,ijk}$ (conditional on covariates) was assumed constant across all scenarios (0.1).

For each subject i in cluster j, standard normal covariates X_i and W_j were generated at the individual and cluster-level respectively. Bivariate normal outcome $(Y_{1,ijk}, Y_{2,ijk})$ was then generated depending on the covariates, as specified in the corresponding scenario and separately in each treatment arm $k = 0, 1$. For example, the outcome depending on the individual-level covariate was simulated according to

$$\begin{aligned} Y_{1,ijk}|u_{1,j}, X_i &\sim N(\mu_{1,k} + \beta_X X_i, \sigma_1^2), \\ \mathrm{E}[Y_{2,ijk} \mid Y_{1,ijk}, u_{1,j}, u_{2,j}] &= \gamma_X X_i + \gamma_1(Y_{1,ijk} - \mu_{1,k}) + u_{2,j}, \end{aligned} \tag{2}$$

where γ_1 is the corresponding regression coefficient of Y_2 on the the first enpoint. Hence, conditioned on the cluster random effects and X, (Y_1, Y_2) is bivariate normal.

Data for other scenarios were simulated by replacing X_i in equation (2) for W_j and finally including both, X_i and W_j. We simulated the data such that $Y_{1,ij}$ had mean $\mu_{1,0} = 100$ in the control arm and $\mu_{1,1} = 120$ in the intervention, with level-1 standard deviation $\sigma_1 = 40$; the corresponding parameters for $Y_{2,ij}$ were $\mu_{2,0} = 50,\ \mu_{2,1} = 60$ and $\sigma_2 = 20$ respectively. The level 2 variances were calculated so as to achieve the level of clustering (measured by the ICC) specified in the corresponding scenario.

We assumed that only the endpoints had missing values, while all other covariates were fully-observed, though the formulation above allows for partially-observed covariates to be imputed simultaneously.[3] Non-response indicators $R_{\ell,ijk}$ for each outcome were independently drawn from a Bernoulli distribution with probabilities $\pi_{\ell,ijk}$ as specified in Table 4 and choosing the intercept α_0 so that the non-response rates were as specified in Table 4. Missing values were then generated to create the respondent data set, all under the MAR assumption. The data generating process was done separately by treatment arm, but we assumed the same parameters in both arms, except for the mean of the endpoints. This means that variables associated with missingness are not associated with treatment effect.

For each scenario, $N = 1000$ datasets were obtained, and each MI method was used to obtain $M = 10$ multiply imputed sets. Then, the analysis model equation (**??**) was applied to each multiply imputed dataset to estimate incremental cost Y_1, incremental health outcome Y_2 and INB. These multiply imputed estimates are then combined using Rubin's rules (Rubin, 1987). Finally, given the true incremental cost, health outcome and INB, we calculated bias, SE of the bias, 95% confidence interval (CI) coverage and root mean square error for each method. These metrics are defined in the Appendix.

4.1. **Results from the simulation.** Across all scenarios, all three MI methods under comparison resulted in estimates for incremental cost and health outcome with negligible bias (not reported). Table 5 reports the results for the base-case scenarios, defined by $\mathrm{ICC}_1, \mathrm{ICC}_2 = 0.01, \eta_X, \eta_W = 1$, with balanced clusters $J = 50$, of size $n_j = 10$, and the non-response rate is 20% in each endpoint.

[3]Partially observed covariates can be included in the imputation model as dependent variables.

TABLE 5. Confidence interval (CI) coverage (%), mean CI width and rMSE obtained with each MI method for estimating incremental cost and health outcomes in a moderate scenario (base case) ($ICC_1 = 0.01,\ ICC_2 = 0.01$, $\eta = 1$, balanced clusters $J = 50,\ n_j = 10$, non-response rate=20%); based in 1000 simulations.

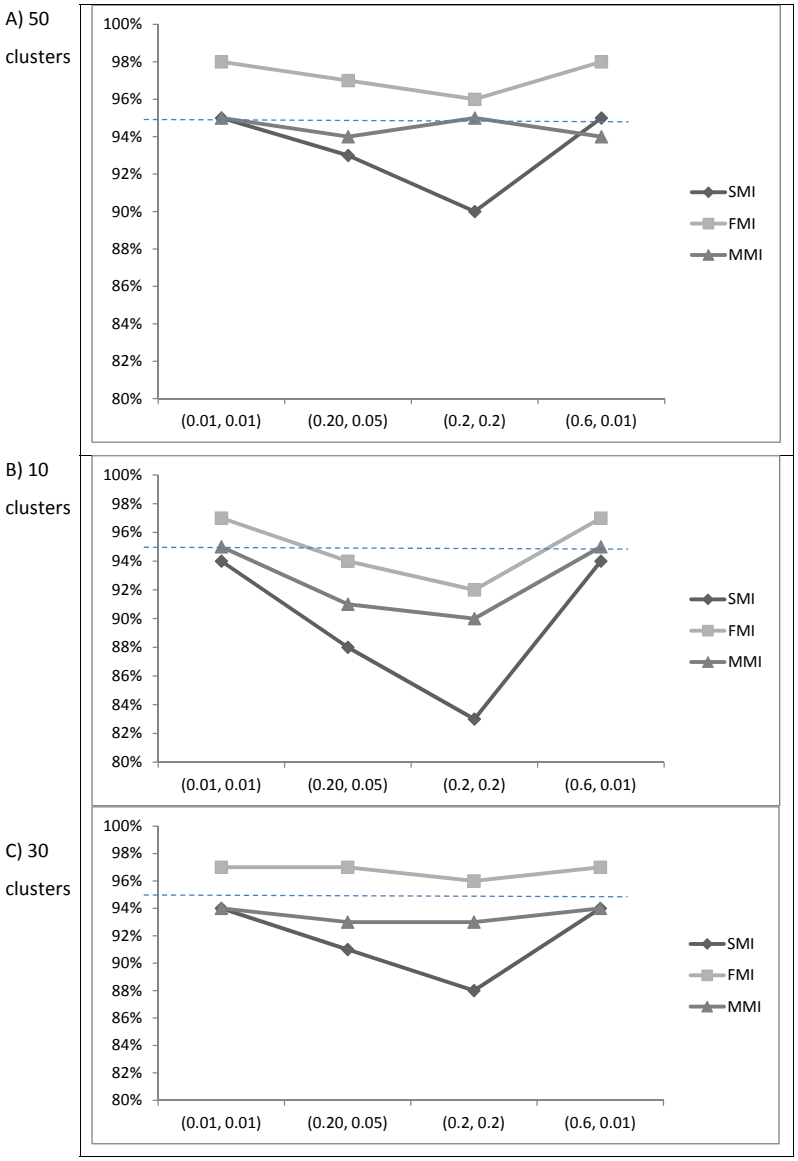
		Incremental cost			Incremental health outcome		
Missingness mechanism	Metric	SMI	FMI	MMI	SMI	FMI	MMI
Individual-level covariate	coverage	94.5	97.8	95.3	94.8	97.3	95.8
	CI width	16.74	19.69	16.88	8.54	10.09	8.66
	rMSE	4.20	4.24	4.20	2.13	2.16	2.14
Cluster-level covariate	coverage	95.0	98.3	94.9	95.4	98.6	96.0
	CI width	17.78	23.95	17.75	8.96	12.00	9.07
	rMSE	4.48	4.69	4.48	2.29	2.42	2.29
Both	coverage	95.0	98.7	94.7	94.6	98.8	95.2
	CI width	17.00	22.07	17.03	8.67	11.16	8.79
	rMSE	4.21	4.37	4.21	2.21	2.29	2.20

TABLE 6. CI coverage (in %) and Mean CI width and rMSE of the incremental cost and health outcome across a range of scenarios under MAR depending on both individual and cluster-level covariates and maintaining all other factor levels fixed with respect to the base-case.

		Incremental cost			Incremental health outcome		
Factor	Metric	SMI	FMI	MMI	SMI	FMI	MMI
Higher ICCs (0.2, 0.2)	coverage	89.2	96.8	92.7	88.9	96.2	93.5
	CI width	23.86	31.06	26.60	12.27	15.91	13.96
	rMSE	7.00	7.08	6.99	3.75	3.81	3.73
Few large clusters	coverage	93.2	97.0	93.6	94.3	97.1	94.8
($J = 10,\ n_j = 50$)	CI width	18.35	22.43	18.91	9.53	11.70	10.04
	rMSE	4.97	5.07	4.91	2.46	2.52	2.48
Unequal cluster size ($cv = 0.5$)	coverage	94.0	97.9	94.6	94.3	97.6	94.7
	CI width	16.14	21.01	16.24	8.21	10.63	8.46
	rMSE	4.25	4.50	4.23	2.14	2.26	2.14
% missing data differ	coverage	95.3	99.1	95.3	94.8	97.0	95.0
by endpoint (0.3; 0.1)	CI width	18.12	25.89	18.15	8.15	9.40	8.21
	rMSE	4.47	5.08	4.52	2.05	2.07	2.04
Higher association of predictors	coverage	95.3	98.8	95.4	94.8	98.9	95.2
with missingness ($\eta = 2$)	CI width	18.72	27.80	18.64	9.32	13.94	9.49
	rMSE	4.58	5.25	4.60	2.38	2.68	2.36

In all three scenarios, each method resulted in similar rMSE for incremental cost and health outcome. However, FMI resulted in over-coverage and has consistently wider CIs, indicating it is an inefficient method. Table 6 shows the results of changing different factors (one at a time) with respect to the base-case under a MAR missing data mechanism depending on both individual and cluster level covariates (last three rows reported in Table 5). In particular, higher ICCs (0.20 for both outcomes) resulted in substantial under-coverage for the single-level MI. The number of clusters and size of clusters appears to be more problematic. For scenarios where the

FIGURE 1. CI coverage for INB with each MI method across alternative levels of ICC and No. of clusters. Nominal level is 95%.



number of clusters is relatively low ($J = 5$ per arm), MMI and SMI have low CI coverage for incremental cost, while for incremental health outcome, coverage is acceptable. FMI has over-conservative coverage in both endpoints. Across all methods and scenarios shown, the accuracy is similar (comparable rMSE) but FMI is systematically inefficient (wider CI for similarly unbiased estimates).

Figure 1 shows the mean coverage of the 95% CI for the INB for the different levels of ICC and for each of the cluster designs considered, marginalising over all other factors. The figure suggest that MMI coverage is nearer to the nominal value in most cases. When both outcomes have similar low levels of clustering, the coverage for FMI is too high, especially when there are

many clusters ($J = 50$) of small size (98%). Single-level MI has acceptable coverage when at least one ICCs is low, but has poor coverage when both ICCs are somewhat high, especially when there are few clusters (83%). In this case, all methods have low coverage. This could be because there is too little information when there are few clusters and observations are very correlated within clusters, and FMI appears to perform better.

Finally, the expected CI coverage of different combination of levels of the main factors, namely, cluster design (large number of small clusters, a few large clusters or unbalanced design) and levels of ICC, across the three missing data mechanism under consideration, is reported in Table 7. We marginalised over the non-response rate and η. It is clear that FMI coverage is always larger than the coverage of the other two methods. This over-estimation of variability is worse for small ICCs. The best method is MMI, even though it too resulted in low coverage when there are only a few clusters and both endpoints have moderate to high ICCs.

5. DISCUSSION

This study compares the relative performance of multilevel and fixed-effects MI with single-level MI for handling missing data in CEA that use cluster trials. The case study illustrates that both point estimates and uncertainty are sensitive to the choice of MI method. The simulation finds that each MI method results in low bias and similar rMSE throughout the scenarios included. This may be expected as, by design, there was no differential missing data mechanism by treatment arm and the auxiliary variables where not associated with treatment effect. The case-study results suggest that when there is such associations, there may be bias. Over and underestimation of variance was studied by examining the coverage of 95% CIs. The single-level MI approach provides low ($< 90\%$) CI coverage across most scenarios, in particular when the ICCs are moderate to high (≥ 0.05) and there are few clusters ($J = 10$). While fixed-effects MI recognises the clustering, it results in conservative CI coverage ($> 96\%$), especially in circumstances where the ICCs are small (as with health outcomes) and there are a moderate to high number of clusters ($J \geq 30$). Multilevel MI gives CI coverage consistently close to the nominal level (95%), except in settings with few clusters, where fixed-effects MI perform slightly better.

This is the first study to test the performance of two alternative MI methods, fixed-effects and multilevel MI, which account for the clustering when handling missing data in CEA that use hierarchical data. This paper highlights that MI approaches that ignore the clustering underestimate the uncertainty. Although in a different context, this corroborates previous findings which showed that single-level MI can increase Type-I error rates (Taljaard et al., 2008). A potential approach to recognise the clustering in the imputation model is to include the cluster as a fixed effect (White et al., 2011; Graham, 2009). However, this imputation model represents the limiting case where the proportion of variability at the cluster-level tends to one and does not properly capture the conditional distribution of the missing data given the observed. It cannot be used to impute cluster-level variables and there may be an issue with estimation, as the number of parameters that need to be estimated by the imputation model to obtain the predictive distribution from

TABLE 7. Marginal mean coverage (%) of the true values by the 95% CI for incremental costs and incremental health outcome estimates under the multiple imputation methods under comparison, averaged over the non-response rate and the strength of association between covariates and non-response indicator.

	Factor			Cost			Health outcome	
Missingness mechanism	Cluster design	ICC	SMI	FMI	MMI	SMI	FMI	MMI
Individual-level covariate	$J = 50$, $n_j = 10$	(0.01,0.01)	95.1	97.9	95.0	94.8	97.1	95.1
		(0.20,0.05)	91.0	96.7	93.8	93.2	96.4	94.0
		(0.20,0.20)	90.7	96.7	93.0	90.6	95.2	93.2
		(0.60,0.01)	86.7	95.2	94.4	95.1	96.9	95.0
	$J = 10$, $n_j = 50$	(0.01,0.01)	92.9	96.3	93.0	94.5	96.7	95.3
		(0.20,0.05)	80.2	92.2	89.6	89.1	93.9	91.8
		(0.20,0.20)	80.6	92.3	89.7	84.0	91.7	90.3
		(0.60,0.01)	78.9	91.6	91.1	94.6	96.5	95.6
	$J = 30$, variable n_j	(0.01,0.01)	93.4	97.4	93.3	94.0	96.6	94.2
		(0.20,0.05)	86.1	96.2	93.1	91.5	95.9	92.7
		(0.20,0.20)	86.0	96.1	92.9	89.2	95.3	92.8
		(0.60,0.01)	83.0	94.7	94.1	94.1	96.3	94.3
Cluster-level covariate	$J = 50$, $n_j = 10$	(0.01,0.01)	94.7	98.8	95.0	95.2	98.3	95.4
		(0.20,0.05)	88.9	97.9	93.1	93.0	97.9	94.1
		(0.20,0.20)	89.1	97.9	92.7	90.8	97.2	93.6
		(0.60,0.01)	84.6	96.0	94.7	95.3	98.3	95.6
	$J = 10$, $n_j = 50$	(0.01,0.01)	93.1	98.4	93.1	93.6	97.8	94.9
		(0.20,0.05)	80.3	93.8	89.1	87.8	94.7	91.0
		(0.20,0.20)	80.8	93.9	89.1	83.4	92.1	90.0
		(0.60,0.01)	75.9	91.6	91.0	94.4	97.6	94.8
	$J = 30$, variable n_j	(0.01,0.01)	94.7	98.8	94.4	94.2	98.1	94.8
		(0.20,0.05)	84.3	96.6	91.1	91.8	97.4	93.4
		(0.20,0.20)	84.9	96.4	90.6	87.3	96.1	92.9
		(0.60,0.01)	79.9	94.0	93.3	94.7	98.1	95.1
Both	$J = 50$, $n_j = 10$	(0.01,0.01)	95.3	98.9	95.1	94.7	98.3	95.1
		(0.20,0.05)	89.0	97.8	94.1	93.0	97.9	94.4
		(0.20,0.20)	89.3	97.8	93.3	89.8	96.4	93.8
		(0.60,0.01)	84.8	95.8	95.0	95.1	98.3	95.3
	$J = 10$, $n_j = 50$	(0.01,0.01)	93.4	97.9	93.6	93.9	97.2	95.0
		(0.20,0.05)	78.2	92.6	89.4	88.2	94.0	91.3
		(0.20,0.20)	78.7	92.8	89.3	82.3	91.9	90.0
		(0.60,0.01)	74.6	91.3	90.8	93.6	97.1	95.3
	$J = 30$, variable n_j	(0.01,0.01)	93.9	98.5	94.3	94.0	97.5	94.3
		(0.20,0.05)	83.9	96.3	92.6	91.0	96.6	92.5
		(0.20,0.20)	83.9	96.3	92.3	86.8	95.4	92.9
		(0.60,0.01)	79.4	94.3	93.7	94.4	96.9	94.5

which the missing data are to be imputed, increases with the number of clusters in the study. In addition, if we have small cluster sizes, the estimated cluster-effects may be unreliable. A previous simulation study considering a single endpoint (Andridge, 2011) found that this MI approach can overestimate the variance especially when ICCs are small and with small clusters sizes.

By contrast, multilevel MI recognises that there is little information for these clusters and "shrinks" their estimated mean towards the overall mean. It properly models the multilevel structure of the data by including cluster-level random-effects, and it is recommended more generally for the analysis of missing data in hierarchical settings (Goldstein et al., 2009; Carpenter & Goldstein, 2004).

A previous study generated several missing data scenarios from a fully-observed CEA alongside a CRT, and suggested that multilevel MI led to both point estimates and standard errors consistently close to the fully-observed estimates (Gomes et al., 2012a). We extended this study in two ways: we considered fixed-effects MI as well as multilevel MI, and we tested the methods across a wider range of circumstances using a full factorial simulation study.

The multilevel MI considered here is compatible with the bivariate multilevel models used as analysis models, which have been proposed for handling clustering in CEA that use CRTs (Grieve et al., 2010; Gomes et al., 2012c), but could be extended to other multilevel structures such as longitudinal studies. In some areas such as the econometric evaluation of health programmes, the use of random-effects approaches has raised two main concerns. First, the method may be sensitive to distributional assumptions of the random-effects (Wooldridge, 2002; Greene, 2003; Jones & Rice, 2011). A simulation study by Yucel & Dermitas (2010) investigated the impact of misspecifying the random-effects distribution of the multilevel MI. They found that when the imputation model has sufficient auxiliary variables, inferences are insensitive to non-normal random-effects, unless the non-response rate is high ($> 50\%$). Second, the random-effects approach assumes no correlation between the unobserved cluster-specific effect and the observed covariates (Greene, 2003; Jones & Rice, 2011). In the context of a randomised trial, we argue that the assumption that clusters-specific effects are not correlated with individual-level variables predictive of missingness is more plausible.

Ultimately, the choice between fixed and random effects models is linked to the target of inference and thus to the substantive research question. If a fixed-effect analysis model has been used for the hierarchical data at hand, then a fixed-effects MI may be appropriate. This is rarely the case for cluster trials but may be the case for other designs such as multinational CEA. Finally, a random-effects imputation model is compatible with a fixed effects analysis model but not vice versa except in the limiting case, as the between-cluster variance increases.

This study has some limitations. Firstly, we have assumed that all variables in the imputation model follow a multivariate normal distribution. While results have been shown in simulations to be robust to departures from normality (Lee & Carlin, 2010; Yucel & Dermitas, 2010), both costs and health outcomes may exhibit complex distributions. For example, the endpoint may be semi-continuous, and normalising transformations may not help (Yu et al., 2007). Similarly, bivariate normal random-effects models were used for the analysis. These have been shown in simulations to performed well even costs were highly skewed (Gomes et al., 2012c).

Secondly, we have assumed that missing data mechanism is MAR throughout. However, the probability that costs and health outcome are missing may be conditional on unobserved information, so that the MAR assumption does not hold (Briggs et al., 2003; Imbens & Wooldridge, 2009). In many situations, it is therefore difficult, or even impossible, to know what mechanism is responsible for the missingness and the best we can hope for is that MAR approximately holds, and that any further dependence on the unobserved data has limited impact on our conclusions. The impact of departures from MAR can be explored through MI (Carpenter et al., 2007), and in principle, standard procedures should apply without much modification.

Another possible concern is the use of different implementations of the sampling mechanism for MI. We used fully-conditional specification models (van Buuren & Groothuis Oudshoorn, 2011) for fixed-effects and single level MI, while multilevel MI was implemented using multivariate normal joint models (Schafer & Yucel, 2002). However, results by Lee & Carlin (2010) show that these approaches lead to similar MI estimates across a range of settings.

While we propose a general approach to handling missing values in clustered data, our study does not represent all the circumstances faced by health econometric evaluations that use hierarchical data. Firstly, as the results from the case study show, the choice of MI method can lead to differences in the point estimates, while in our simulation we observed no bias. In the case-study, there were variables associated with missingness and treatment effect, while in our simulations this was not the case. Future studies will explore scenarios were the missing data mechanism is different by treatment arm and is associated with average treatment effect. In addition, there may be circumstances when the data display quite different structures to those considered here, for example the endpoints may have a semin-continuous distribution (Yu et al., 2007), or health outcomes with highly irregular distributions (Basu & Manca, 2012). In addition, there may be many auxiliary variables available, including post-randomisation variables.

Finally, while here we combine multilevel MI with a multilevel model estimated by maximum likelihood, there may be circumstances in which it would be advantageous to combine multilevel MI with multilevel models estimated by MCMC (Lambert et al., 2005), or indeed adopt a fully Bayesian approach to handling the missing data and specify an analytical model which can simultaneously accommodate MNAR mechanisms (Mason et al., 2012).

This paper suggests some new directions for further research. Future studies could explore other circumstances, e.g. missing data mechanisms to differ by treatment group, or informative cluster size, and allow for non-normal distributions. Further research will also consider circumstances where missing data are MNAR.

Acknowledgements. We are grateful to Jane Morrell (PI) and Simon Dixon for permission to use, and for providing access to, the PONDER data. This work was partly funded by the UK Medical Research Council grant (RG and KDO).

REFERENCES

Andridge, R. R. (2011). 'Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials.'. *Biometrical journal. Biometrische Zeitschrift* **53**(1):57–74.

Basu, A. & Manca, A. (2012). 'Regression estimators for generic health-related quality of life and quality-adjusted life years'. *Medical Decision Making* **32**:56–69.

Blough, D. K. , et al. (2009). '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* **18**:91–101.

Briggs, A. , et al. (2003). 'Missing... presumed at random: cost-analysis of incomplete data.'. *Health economics* **12**(5):377–92.

Burton, A. , et al. (2006). 'The design of simulation studies in medical statistics'. *Stat Med* **25**(24):4279–92.

CADTH (2006). *Guidelines for the Economic Evaluation of Health Tecnologies*. Canadian Agency for Drugs and Technologies in Health. Ottawa, Canada.

Campbell, M. K. , et al. (2005). 'Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research'. *Clinical Trials* **2**(2):99–107.

Carpenter, J. R. & Goldstein, H. (2004). 'Multiple imputation in MLwiN.'. *Multilevel Modelling Newsletter* **16**:9–18.

Carpenter, J. R. , et al. (2011). 'REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types'. *Journal of Statistical Software* **45**(5):1–14.

Carpenter, J. R. , et al. (2006). 'A comparison of multiple imputation and inverse probability weighting for analyses with missing data'. *Journal of the Royal Statistical Society, Series A* **169**:571–584.

Carpenter, J. R. , et al. (2007). 'Sensitivity analysis after multiple imputation under missing at random: a weighting approach.'. *Statistical Methods in Medical Research* **16 (3)**:259–275.

Cox, D. R. & Reid, N. (2000). *The theory of the design of experiments. Monographs on statistics and applied probability (86)*. Chapman & Hall. CRC.

Diaz Ordaz, K. , et al. (2012). 'Handling missing values in cost-effectiveness analyses that use data from cluster randomised trials'. *Submitted* .

Glick, H. A. , et al. (2007). *Economic evaluation in clinical trials*. Oxford University Press, Oxford, UK.

Gold, M. , et al. (eds.) (1996). *Cost-effectiveness in health and medicine*. Oxford University Press.

Goldstein, H. , et al. (2009). 'Multilevel models with multivariate mixed response types'. *Statistical Modelling* **9**(3):173–197.

Gomes, M. , et al. (2012a). 'Multiple imputation methods for handling missing data in CEA: an application to cluster randomized trials.'. *Med Decis Making.* **submitted**.

Gomes, M. , et al. (2012b). 'Methods for Covariate Adjustment in Cost-Effectiveness Analysis That Use Cluster Randomised Trials'. *Health Econ* .

Gomes, M. , et al. (2012c). 'Developing Appropriate Methods for Cost-Effectiveness Analysis of Cluster Randomized Trials'. *Medical Decision Making* **32**(2):350–361.

Graham, J. W. (2009). 'Missing Data Analysis: Making It Work in the Real World'. *Annual Review of Psychology* **60**:549–576.

Gray, A. , et al. (2010). *Applied Methods of Cost-effectiveness Analysis in Healthcare*. Oxford University Press.

Greene, W. H. (2003). *Econometric analysis*. Prentice Hall, Upper Saddle River, N.J., Great Britain, 5th edn.

Grieve, R. , et al. (2010). 'Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials.'. *Medical decision making : an international journal of the Society for Medical Decision Making* **30**(2):163–75.

Imbens, G. W. & Wooldridge, J. M. (2009). 'Recent Developments in the Econometrics of Program Evaluation'. *Journal of Economic Literature* **47**(1):5–86. 427VM Times Cited:50 Cited References Count:311.

Jones, A. & Rice, N. (2011). *Econometric Evaluation of Health Policies*. Oxford University Press, Oxford, UK.

Lambert, P. , et al. (2005). 'How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS.'. *Statistics in medicine* **24**:2401–28.

Lee, K. J. & Carlin, J. B. (2010). 'Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation'. *American Journal of Epidemiology* **171**(5):624–632.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data (Second Edition)*. Chichester: Wiley.

Mason, A. , et al. (2012). 'Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods.'. *Journal of Official Statistics* **28**:279–302.

Molenberghs, G. & Kenward, M. G. (2007). *Missing Data in Clinical Studies.* Wiley, Chichester.

Morrell, C. J. , et al. (2009). 'Clinical effectiveness of health visitor training in psychologically informed approaches for depression in postnatal women: pragmatic cluster randomised trial in primary care'. *British Medical Journal* **338**:3045.

NICE (2008). *Methods for Technology Appraisal.* National Institute for Health and Clinical Excellence, London, UK.

Noble, S. , et al. (2012). 'Missing data in trial-based cost-effectiveness analysis: the current state of play.'. *Health Econ* **21(2)**:187–200.

Oostenbrink, J. B. & Al, M. J. (2005). 'The analysis of incomplete cost data due to dropout'. *Health Econ* **14**(8):763–76.

PBCA (2008). *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee.* Australian Government - Department of Health and Ageing., Camberra, Australia.

Rubin, D. (1978). 'Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse'. *Proceedings of the Survey Research Methods Section of the American Statistical Association* pp. 20–34.

Rubin, D. (1987). 'Multiple Imputation for Nonresponse in Surveys'. *J. Wiley Sons, New York* .

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall. London.

Schafer, J. & Yucel, R. (2002). 'Computational strategies for multivariate linear mixed-effects models with missing values.'. *Journal of Computational and Graphical Statistics* **11**:421–442.

Schafer, J. L. (1999). 'Multiple imputation: a primer'. *Statistical Methods in Medical Research* **8**:3–15.

Taljaard, M. , et al. (2008). 'Imputation strategies for missing continuous outcomes in cluster randomized trials.'. *Biometrical journal. Biometrische Zeitschrift* **50**(3):329–45.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall. CRC Press.

van Buuren, S. & Groothuis Oudshoorn, K. (2011). 'mice: Multivariate Imputation by Chained Equations in R'. *Journal of Statistical Software* **45**(3):1–67.

White, I. R. & Carlin, J. B. (2010). 'Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values.'. *Statistics in Medicine* **29(28)**:2920–2931.

White, I. R. , et al. (2011). 'Multiple imputation using chained equations: Issues and guidance for practice.'. *Statistics in medicine* **30**(4):377–99.

Willan, A. & Briggs, A. (2006). *Statistical Analysis of Cost-Effectiveness Data.* John Wiley & Sons Ltd.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press, Cambridge, Mass.

Yu, L. M. , et al. (2007). 'Evaluation of software for multiple imputation of semi-continuous data'. *Stat Methods Med Res* **16**(3):243–58.

Yucel, R. & Dermitas, H. (2010). 'Impact of non-normal random effects on inference by multiple imputation: A simulation assessment'. *Comput Stat Data Anal* **54**(3):790–801.

APPENDIX

We describe here briefly the metrics used to assess relative performance of the MI methods in the simulation study. The mean square error (MSE) provides a measure of accuracy, it incorporates bias and variability. The square root of the MSE transforms the MSE back onto the same scale as the parameter.The CI coverage is the proportion of times that the obtained CI contains the true parameter of interest. The coverage should be approximately equal to the nominal coverage, 95%. Over-coverage, i.e. rates above 95%, suggests that the results are too conservative (over-estimation of the variance). By contrast, under-coverage (lower than 95%) indicates over-confidence in the estimates (the precision is over-estimated). Acceptable coverage should not fall outside of approximately two SEs of the nominal coverage probability, so for 95% CI based on

a 1000 simulated sets, between 936 and 964 of the CIs should include the true value. The average width of the 95% CI provides a measure of efficiency and power, as for estimates which are relatively unbiased, a narrower CI implies more precise estimates (Burton et al., 2006).

These metrics are defined in Table 8.

TABLE 8. Performance measures for estimates $\hat{\theta}_k$ of a given parameter of interest θ, with N simulated data.

Measure	Definition	Best performance
Bias	$B = \frac{\sum_{k=1}^{N} \hat{\theta}_k - \theta}{N}$	$B = 0$
SE of bias	$\text{SE}(B) = \frac{\text{SD}(\hat{\theta})}{\sqrt{N}}$ where $\text{SD}(\hat{\theta}) = \sqrt{\frac{\sum_{k=1}^{N} (\hat{\theta}_k - \bar{\hat{\theta}})^2}{N-1}}$	Lowest $\text{SD}(B)$
rMSE	$\sqrt{\frac{\sum_{k=1}^{N} (\hat{\theta}_k - \theta)^2}{N}}$	Lowest rMSE
CI coverage	$\frac{1}{N} \sum_{k=1}^{N} I_{\text{CI}(\hat{\theta}_k)}(\theta)$	Nearest to the nominal level
Mean CI width	$\frac{1}{N} \sum_{k=1}^{N} \lvert\text{CI}(\hat{\theta}_k)\rvert$	Smallest Mean CI width

where SE = standard error; SD = standard deviation; rMSE = root mean square error; $\text{CI}(\hat{\theta}_k)$ is the 95% confidence interval for estimate $\hat{\theta}_k$, $I_{\text{CI}(\hat{\theta}_k)}(x)$ is an indicator function, taking the value 1 if x belongs to the interval and $\lvert \circ \rvert$ represents the length of an interval.