ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2016)
179, *Part* 3, *pp.* 683–706

A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies

Oya Kalaycioglu,

University College London, UK, and Abant Izzet Baysal University, Bolu, Turkey

Andrew Copas

Medical Research Council Clinical Trials Unit, London, UK

and Michael King and Rumana Z. Omar

University College London, UK

[Received June 2014. Final revision June 2015]

Summary. Multiple-imputation (MI) methods for imputing missing data in observational health studies with repeated measurements were evaluated with particular focus on incomplete time varying explanatory variables. Standard and random-effects imputation by chained equations, multivariate normal imputation and Bayesian MI were compared regarding bias and efficiency of regression coefficient estimates by using simulation studies. Flexibility of the methods in handling different types of variables (binary, categorical, skewed and normally distributed) and correlations between the repeated measurements of the incomplete variables were also compared. Multivariate normal imputation produced the least bias in most situations, is theoretically well justified and allows flexible correlation for the repeated measurements. It can be recommended for imputing continuous variables. Bayesian MI is efficient and may be preferable in the presence of categorical and non-normally distributed continuous variables. Imputation by chained equations approaches were sensitive to the correlation between the repeated measurements. The moving time window approach may be used for normally distributed continuous variables with auto-regressive correlation.

Keywords: Bayesian imputation; Imputation by chained equations; Missing data; Multilevel data; Multiple imputation; Multivariate normal imputation

1. Introduction

Missing data commonly occur in longitudinal observational studies where data on both outcome and explanatory variables are collected repeatedly at several time points. In particular, studies on vulnerable patient groups such as patients suffering from cancer or mental health problems can experience considerable missing data, both for the outcome and for the time varying explanatory variables (King *et al.*, 2008; Paskett *et al.*, 2007; Peyre *et al.*, 2011). These data are typically analysed by using regression models that account for the repeated measurements, such as multilevel models. One can conduct an available case (AC) analysis (Fitzmaurice, 2003) by restricting the analysis to include only time points where patients have complete information

Address for correspondence: Oya Kalaycioglu, Department of Econometrics, Abant Izzet Baysal University, 14280 Golkoy, Bolu, Turkey.
E-mail: kalaycioglu.oya@gmail.com

 0964–1998/16/179683

for all variables. When the missing values of the outcome depend on the explanatory variables and/or values of the outcome observed at previous time points, estimates of the regression coefficients from the AC analysis are unbiased provided that

(a) the explanatory variables are fully observed and
(b) the variables that are related to both missingness and outcome are included in the analysis model (Molenberghs and Kenward, 2007).

However, the estimates may not be unbiased when the explanatory variables are missing at random (Goldstein *et al.*, 2009). Moreover, if there is a considerable amount of missing data, the estimates may be inefficient owing to the reduced sample size. In such situations the use of multiple-imputation (MI) methods may alleviate the problem of bias and efficiency of the estimates (Goldstein *et al.*, 2009).

This work was motivated by a cohort study of 199 cancer patients, investigating the association between their perceived continuity of care and their level of health needs. Both outcome and explanatory variables were measured at five time points, resulting in considerable missing data. Although MI methods have increasingly been used for handling missing data in recent years, their use is not straightforward for repeated measurement studies. The imputation methods need to account for the correlation between the repeated measurements of the incomplete variables and to make specific assumptions regarding the correlation. Whereas several researchers have compared MI methods for studies where variables are measured at a single time point (van Buuren, 2007; Yu *et al.*, 2007; Lee and Carlin, 2010), a detailed comparison of the imputation methods for repeated measurement studies has not been done. The aim of this paper is to evaluate the MI methods for the analysis of observational repeated measurement studies with particular focus on missing time varying explanatory variables. Simulation studies are used to investigate the bias, coverage and efficiency of the estimates of the regression coefficients when using MI for different types of variable (binary, categorical and normal and non-normal continuous), and different correlation structures for the repeated measurements of the incomplete variables. Recommendations are made for the use of MI methods in practice. The layout of the paper is as follows. The data which motivated this work and are used as a basis for comparing the MI methods and AC analysis are introduced in Section 2. In Section 3, the AC analysis and the various imputation methods for handling missing data are described. The results from the case-study and the simulation study are presented in Sections 4 and 5 respectively. Section 6 provides a discussion.

2. Data

The cohort study of 199 cancer patients investigated the association between the perceived continuity of care of these patients (defined by a score) and their physical, psychological and health system needs, satisfaction with care, mental health status and the type of cancer (King *et al.*, 2008). The outcome and the explanatory variables were measured for each patient at recruitment and every 3 months throughout a 12-month follow-up period, providing a total of 995 measurements. The variables are summarized in Table 1. Approximately 48% of the measurements had missing observations; 10% had outcome fully observed and at least one explanatory variable missing, 27% had both outcome and all of the time varying explanatory variables missing because of dropout of patients, 5% had sporadic missed visits and 6% had only outcome missing. The correlations between the repeated measurements of both the explanatory variables and the outcome appeared to have an unstructured form.

Table 1. Summary of the variables in the continuity-of-care data

Variable	*Description*	*Mean*	*Standard deviation*	*Range*	*% of missing values*	*Skewness*	*Kurtosis*
Outcome (measured repeatedly)							
concare	Continuity of cancer care score	50.6	9.4	7–73	38.2	—	—
Explanatory variables (measured repeatedly)							
phys	Physical need score	10.7	4.7	5–25	33.2	0.58	2.5
psyc	Psychological need score	19.6	8.6	10–50	33.8	0.96	3.3
hltsys	Health system need score	19.2	6.7	11–55	34.1	1.25	5.9
satis	Satisfaction score	41.5	7.8	10–50	39.0	0.57	17.5
Explanatory variables (measured at baseline)							
site	Cancer site (breast, lung or colorectal)	—	—	0, 1 or 2	0.0	—	—
trtphase	Treatment phase (initial diagnosis, completion of first treatment, remission, relapse or specialist palliative care)	—	—	0, 1, 2, 3 or 4	0.0	—	—

3. Methods for analysis

3.1. Notation

The univariate outcome for the ith patient at time j is represented as Y_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, t)$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{it})^{\mathrm{T}}$ is defined as the $(t \times 1)$-dimensional vector of planned measurements of the outcome, the time varying kth explanatory variable $(k = 1, \ldots, p)$ for the same patient is represented as X_{ijk} and hence the $t \times p$ covariate matrix of p explanatory variables is $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip})$. Parameters β and γ denote the regression coefficients in the analysis and missingness models (which represent how the probability of missingness depends on other variables) respectively.

3.2. Available case analysis

A simple multilevel model for the analysis can be defined as

$$Y_{ij} = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ijk} + u_i + e_{ij} \tag{1}$$

where β_0 and the β_ks are the fixed coefficients for the explanatory variables and $u_i \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$ represent the random coefficients at the patient and measurement time point levels respectively. This model assumes an exchangeable correlation structure between the repeated measurements of the outcome. Other types of correlation for the outcome were also considered for the case-study. In AC analysis, model (1) is fitted to the completely observed measurement time points by maximum likelihood.

3.3. Multiple-imputation methods

The MI method proposed by Rubin (1987, 2008) uses *imputation models* to replace the missing values of the incomplete variables with plausible values based on the distribution of the observed data. The imputation models use the relationship between the incomplete variables and the other variables in the data and typically include variables that are associated with missingness

as predictors. Markov chain Monte Carlo (MCMC) or Gibbs sampling estimation techniques (generally with non-informative priors) are used to fit the models. The missing values are replaced with m random draws, resulting in m distinct complete data sets which are analysed by maximum likelihood using the *analysis model* (e.g. model (1)). There is no need for the variables that are used in the analysis and the imputation models to match. If there are auxiliary variables in the data which are not of direct interest for the analysis but could be related to the missing values, these may also be used as predictors in the imputation model. For single-time-point data, it has been recommended that the number of imputed data sets should be equal to the proportion of missing observations (White *et al.*, 2011). The estimated regression coefficients from these m analysis models are combined by using Rubin's (1987) rule to obtain a single estimate of each regression coefficient and its standard error. For repeated measurement studies, including data from time points in the analysis where the outcome was imputed may add noise to the estimates and reduce their efficiency. von Hippel (2007) thus recommended the inclusion of data from all time points in the imputation model but the exclusion of time points with imputed outcomes from the analysis model. This method is referred to as the multiple-imputation–deletion (MID) approach. However, when the proportion of time points with missing outcome values is large, the MID approach has the disadvantage of excluding a substantial amount of measurements from the analysis model and hence may produce inefficient estimates.

There are two main approaches for MI based on different modelling assumptions:

(a) multivariate normal imputation (Schafer, 1997) and
(b) imputation by chained equations (ICE) (van Buuren and Oudshoorn, 2000).

These approaches are described below in the context of repeated measurements data and the imputation models that were used by each method are summarized in Table 2.

Table 2. Imputation models and their assumptions used by MI methods

Method	*Imputation model*	*Predictor variables used in the imputation model*	*Variable types that can be imputed*	*Correlations assumed between time points of the incomplete variables*
Multivariate normal imputation	Multivariate normal random-effects model	Fully observed	Continuous† Binary or categorical‡	Unstructured
ICE(FE-MTW)	Fixed effects regression model(s)	Fully observed and incomplete	Continuous† Binary or categorical‡	Auto-regressive
ICE(FE-FS)	Fixed effects regression model(s)	Fully observed and incomplete	Continuous† Binary or categorical‡	Does not assume a specific correlation structure
ICE(RE)	Linear random-effects regression model(s)	Fully observed and incomplete	Continuous†	All coefficients in the imputation model are specified to be random
Bayesian MI	Random effects regression model(s)	Fully observed explanatory variables and outcome§	Continuous† Binary or categorical‡	Any correlation structure can be defined by the analyst

†Continuous variables are assumed to be normally distributed.
‡Binary or categorical variables are imputed assuming an underlying latent normal distribution.
§If the outcome is incomplete, the feedback from the analysis model to the imputation models should be prevented by using the cut function.

3.3.1. Multivariate normal imputation

In the multivariate normal imputation approach a three-level random-effects model is specified where the incomplete variables are treated as multivariate normal responses and the fully observed variables are included as predictors. The lowest level or level 1 of this model is used to define the multivariate incomplete variables and thus no variation is specified at this level. Level 2 represents the measurement time points with level 3 representing the patients. If X_{ij1} and X_{ij2} are two incomplete continuous explanatory variables and the remaining variables Y_{ij}, $X_{ij3}, \ldots, X_{ijp}$ are fully observed, then X_{ij1} and X_{ij2} are treated as bivariate normal responses and the multivariate imputation model can be written as

$$\begin{aligned} X_{ij1} &= \gamma_{11} Y_{ij} + \sum_{k=2}^{p-r-1} \gamma_{k1} X_{ij(k+1)} + \sum_{k=1}^{r} u_{ik1} Z_{ijk} + e_{ij1}, \\ X_{ij2} &= \gamma_{12} Y_{ij} + \sum_{k=2}^{p-r-1} \gamma_{k2} X_{ij(k+1)} + \sum_{k=1}^{r} u_{ik2} Z_{ijk} + e_{ij2}, \end{aligned} \tag{2}$$

where γ_{k1} and γ_{k2} $(k = 1, \ldots, p-r-1)$ are the fixed effects parameters for the fully observed predictors and u_{i1} and u_{i2} represent the patient level residuals for the first and the second responses (incomplete variables) respectively. $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{ir})$ is a subset of $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip})$ with r elements, which contains the predictor variables for the patient level random effects. The terms e_{ij1} and e_{ij2} represent the variability of the repeated observations within a patient for the corresponding responses. The patient and measurement time point level residuals are assumed to be independent, with separate unstructured covariance matrices (Carpenter *et al.*, 2011) and they are sampled from the inverse Wishart distribution (see Goldstein *et al.* (2009)).

This method assumes an underlying latent normal distribution (Goldstein *et al.*, 2009) for imputing missing binary or categorical variables. A multilevel multivariate normal response model is fitted and continuous imputed values for the latent normal variables are generated. The imputed values are then converted back into appropriate categories on the original scales, using a threshold value for each category.

3.3.2. Multiple imputation by chained equations

Rather than constructing a joint distribution for the incomplete variables, the ICE approach uses a separate imputation model based on conditional densities (Molenberghs and Kenward, 2007; van Buuren and Oudshoorn, 2011) to impute each incomplete variable. This approach is advocated as an alternative to multivariate normal imputation when no suitable joint distribution of the incomplete variables is found and there are many incomplete variables (van Buuren and Oudshoorn, 2011). However, a requirement for ICE is that the univariate conditional densities should correspond to a genuine joint multivariate distribution of the incomplete variables (Molenberghs and Kenward, 2007). The ICE approach replaces the missing values in X_{ijk} by estimates obtained from regressing X_{ijk} on all the other variables in the data as well as the outcome. Different types of incomplete variables (i.e. continuous, binary or categorical) can be handled using appropriate regression models, e.g. linear or logistic regression.

3.3.2.1. Fixed effects imputation by chained equations. The fixed effects ICE approach uses standard regression models as imputation models, with all predictors specified to have fixed coefficients. The correlation between the repeated measurements of an incomplete variable is accounted for by including its values measured at previous and subsequent time points as fixed effects in the imputation model when imputing the variable at a specific time point (White and

Carlin, 2010). Consequently, this approach assumes an auto-regressive correlation between the repeated measurements of the incomplete variables. However, the inclusion of several measurement time points and time varying variables in the imputation models may result in overfitting and collinearity problems (Nevalainen *et al.*, 2009). To avoid this problem the moving time window approach (ICE(FE-MTW)) (Nevalainen *et al.*, 2009) was proposed where a variable at a specific time point X_{ij1} is imputed using information from all time points for that variable and the outcome but only information from the previous, current and next time points for the other explanatory variables:

$$X_{ij1} = \gamma_0 + \gamma_1^y Y_{i1} + \ldots + \gamma_t^y Y_{it} + \gamma_1^{x_1} X_{i11} + \ldots + \gamma_{j-1}^{x_1} X_{i(j-1)1}^{x_1} + \gamma_j^{x_1} X_{i(j+1)1} + \ldots + \gamma_{t-1}^{x_1} X_{it1} + \sum_{k=2}^{p} \gamma_1^{x_k} X_{i(j-1)k} + \sum_{k=2}^{p} \gamma_2^{x_k} X_{ijk} + \sum_{k=2}^{p} \gamma_3^{x_k} X_{i(j+1)k} + e_{ij}. \quad (3)$$

Alternatively a stepwise forward selection approach (ICE(FE-FS)) (Nevalainen *et al.*, 2009) may be used to select the predictors in the imputation model on the basis of the strength of their association (p-value) with the variable being imputed or its contribution to the model R^2. This approach does not impose any definite structure for the correlation between the measurements of the incomplete variables.

3.3.2.2. Random-effects imputation by chained equations. Random-effects ICE is an extension of the ICE approach where linear mixed models with random coefficients are used to account for the correlation between the repeated measurements in the imputation model (van Buuren and Oudshoorn, 2011). The implementation of this method in currently available software requires all predictors in the imputation model to be specified as random effects, thus allowing the predictors to have a different effect on the imputed variable for each patient. The two-level random-coefficients imputation model can be expressed as

$$X_{ij1} = \gamma_0 + u_{i0} + (\gamma_1 + u_{i1}) Y_{ij} + \sum_{k=2}^{p} (\gamma_k + u_{ik}) X_{ijk} + e_{ij},$$
$$\begin{pmatrix} u_{i0} \\ u_{i1} \\ \vdots \\ u_{ip} \end{pmatrix} \sim N(0, \Omega_u), \qquad e_{ij} \sim N(0, \Omega_e), \quad \Omega_e = \sigma_e^2 \quad (4)$$

where u_{i0} represents the random intercept and the u_{ik} ($k = 1, \ldots, p$) represent the random coefficients for the predictors of the ith patient. The terms u_{i0} and u_{ik} ($k = 1, \ldots, p$) are assumed to be independent of each other; hence Ω_u is a diagonal matrix. The terms e_{ij} represent the random variation between observations and are independent across patients.

3.4. Bayesian multiple imputation

For continuous incomplete variables, the MI methods that were described above require the assumption of normality. To impute non-normal continuous incomplete variables, it has been recommended that appropriate transformations are used to achieve normality along with predictive mean matching (von Hippel, 2009). Predictive mean matching involves sampling only from the observed values to avoid the generation of out-of-range values via back-transformation. However, when the actual range of the variable is wider than the observed range in the data, predictive mean matching may not be suitable. Moreover, transformations may not preserve the original correlation structure between the variables (Demirtas *et al.*, 2008) and generate outliers. Furthermore, it may not be possible to find a suitable transformation particularly when

non-normality is due to reasons other than skewness (e.g. non-zero peakedness and kurtosis) (Demirtas *et al.*, 2008). A flexible alternative is to use the Bayesian implementation of the hierarchical imputation models which allows other distributions such as gamma, log-normal and half-normal to be specified for non-normal continuous incomplete variables.

In the Bayesian MI approach, both the analysis and the imputation models require specification of priors for the fixed and random parameters (Lunn *et al.*, 2009). When there is no prior knowledge or no consensus on the prior distribution of the parameter it is common to use non-informative priors. The Bayesian random-effects analysis model can be formulated in a similar way to model (1), using non-informative priors on $\beta_0, \beta_1, \ldots, \beta_p \sim N(0, 10^6)$ and $1/\sigma_u^2, 1/\sigma_e^2 \sim$ gamma(0.001,0.001). Similarly, incomplete explanatory variables can be imputed by using random-effects imputation models, in which, for example, non-informative $N(0, 10^6)$ priors are given to the fixed effects parameters and gamma(0.001,0.001) priors are given to the inverse of the variance of the random-effects parameters. In Bayesian MI, the analysis and imputation model parameters are updated at each MCMC iteration. An MCMC iteration is completed when all missing values and the model parameters have been estimated. The subsequent iterations are based on these estimates and imputed missing values. It is therefore not necessary to specify the number of imputations as the missing values are imputed at each MCMC iteration and the final estimates are the summary (e.g. mean and median) of all the estimates from the MCMC iterations. In contrast, in standard MI approaches the final estimate is the summary of the estimates from m randomly drawn imputed data sets. If a substantial amount of missingness occurs in explanatory variables, then the imputations for these variables are likely to be strongly influenced by the 'feedback' from the outcome which is estimated by the analysis model (Lunn *et al.*, 2009). A 'cut' function in WinBUGS (Lunn *et al.*, 2009; Carpenter and Kenward, 2005; Carrigan *et al.*, 2007) is thus used to prevent the feedback from the analysis model to imputations so that there is no influence of the estimated outcome values on the imputed values of the explanatory variables at a particular MCMC iteration. The MID approach is not relevant for Bayesian MI since outcomes are not imputed as the analysis model serves the same purpose. The imputation and analysis models can be specified to have exchangeable, independent, auto-regressive or unstructured correlations.

The imputation model for the Bayesian MI can either be a multivariate response model as a counterpart of multivariate normal imputation, or it can consist of separate models for each incomplete variable similar in a sense to ICE. It is important to note that in Bayesian MI incomplete explanatory variables are not used in the imputation models and when there are no fully observed variables in the data the latter approach may suffer because of the lack of information in the imputation models.

Table 2 summarizes the imputation models that are used by the various MI methods.

4. Results from the case-study

The data from the continuity-of-care study were analysed and the results obtained from applying the various MI methods and the AC analysis were compared. The data suggest that the correlations between the repeated measurements of the outcome have an unstructured form; however, when a multilevel analysis model with an unstructured correlation was fitted to the data it did not converge. Therefore, for illustration a random-intercept model is fitted with concare as outcome and $X_1 =$ phys, $X_2 =$ psyc, $X_3 =$ hltsys, $X_4 =$ satis, $X_5 =$ site:lung, $X_6 =$ site:col and $X_7 =$ period as the explanatory variables. The MLwiN MI macros were used for multivariate normal imputation (Carpenter and Goldstein, 2005) and Stata 12.0's ICE routine (Royston, 2005; StataCorp, 2011) for the fixed effects ICE approaches. Currently, Stata 13.0

supports fixed effects ICE through the `mi impute chained` command. It became available after this work had been done. Using this new command for the analysis did not change the results. The random-effects ICE model ICE(RE) was fitted with the R `mice` package using the `2l.norm` option (van Buuren and Oudshoorn, 2011; R Development Core Team, 2008). 50 imputations were performed for each method. Imputed data sets were analysed both including and excluding the time points with imputed outcome (MID) for comparison. WinBUGS 1.4 (Lunn *et al.*, 2000) was used for Bayesian MI with 25 000 MCMC iterations and 5000 iterations for the burn-in.

Since there were five measurement time points in this study the fixed effects ICE resulted in 25 imputation models (five incomplete variables times five time points) with 26 predictors in each model. Some of these imputation models did not converge, possibly because of overfitting and collinearity problems, and thus this method could not be used to analyse the real data. For model ICE(FE-FS) a p-value of less than 0.1 was used to select the predictors in the imputation model and the outcome at all available time points was forced to remain in the models.

Some of the incomplete continuous variables in these data showed substantial departures from normality with skewed distribution and non-zero peaks either at the maximum or the minimum values (see Fig. 1 in the on-line appendix). Transformations such as the Box–Cox and shifted-log-transformations did not help to achieve normality and using predictive mean matching produced values that were similar to that obtained from model ICE(FE) (the results are not presented). When using the ICE(RE) approach, the imputed values of the non-normally distributed continuous variables were constrained to be between 0 and the maximum value that they could take by using the `squeeze` option within the approach, to avoid imputing out-of-range values.

When applying Bayesian MI, three different imputation model structures were considered.

(a) *Bayesian MI–univariate gamma*: univariate random-intercept gamma imputation models were used for imputing positively skewed variables (phys, psyc and hltsys) defined as (e.g. for psyc)

$$\begin{aligned} \text{psy}_{ij} &\sim \text{gamma}(\mu_{ij}^{\text{imp}^2}\tau, \mu_{ij}^{\text{imp}}\tau), \\ \log(\mu_{ij}^{\text{imp}}) &= \gamma_{0i} + \gamma_1 \,\text{concare}_{ij} + \gamma_2 \,\text{site:lung}_i + \gamma_3 \,\text{site:col}_i + \gamma_4 \text{period}_{ij}, \end{aligned} \tag{5}$$

where an $N(\mu_{\gamma 0}^{\text{imp}}, \sigma_{\gamma 0}^{\text{imp}^2})$ prior is specified for the random intercept γ_{0i}. The covariance structure is imposed on the precision parameter τ, which is the inverse of the variance. Non-informative $N(0, 10^6)$ priors were specified for the regression coefficients γ_k $(k = 1, 2, 3, 4)$ and $\mu_{\gamma 0}^{\text{imp}}$ and a gamma(0.01, 0.01) prior for the precision parameter τ. A truncated normal distribution was used for the negatively skewed variable satis by truncating it at its maximum observed value of 50.

(b) *Bayesian MI–univariate truncated normal*: univariate random-intercepts truncated normal imputation models were used for imputing all the incomplete explanatory variables. The positively skewed variables were truncated between 0 and 100 and satis was again truncated at 50. $N(0, 10^6)$ and gamma(0.01, 0.01) priors were specified for the regression coefficients of the imputation models, and the precision parameters respectively.

(c) *Bayesian MI–multivariate truncated normal*: the Bayesian extension of multivariate normal imputation uses a hierarchical multivariate truncated normal imputation model which treats the incomplete variables as multivariate responses of the imputation model. Bayesian MI does not require an imputation model for the study outcome: hence the number of multivariate responses is reduced to 4 when compared with the multivariate normal imputation model given in equation (2).

Table 3. Estimates of the regression coefficients and their standard errors obtained with the random-intercepts analysis model using different imputation methods for handling the missing data

Method	*Results for the following variables:*						
	phys†	*psyc*†	*hltsys*†	*satis*†	*site:lung*	*site:col*	*period*
AC analysis	−0.099	−0.198	−0.100	0.503	−0.480	−0.139	−0.658
	(0.084)	(0.045)‡	(0.058)	(0.046)‡	(0.888)	(0.963)	(0.200)‡
Bayesian MI–univariate gamma	−0.121	−0.223	−0.064	0.571	−0.431	−0.646	−0.603
	(0.077)	(0.039)§	(0.055)	(0.045)§	(0.881)	(0.895)	(0.184)§
Bayesian MI–univariate truncated normal	−0.139	−0.169	−0.077	0.484	−0.485	−0.393	−0.553
	(0.079)	(0.041)§	(0.056)	(0.046)§	(0.912)	(0.952)	(0.192)§
Bayesian MI–multivariate truncated normal	−0.151	−0.156	−0.094	0.401	−0.576	−0.210	−0.623
	(0.083)§	(0.047)§	(0.059)	(0.050)§	(1.013)	(1.024)	(0.204)§
MI including time points with incomplete outcome							
Multivariate normal imputation	−0.175	−0.160	−0.107	0.436	−0.519	−0.342	−0.723
	(0.084)‡	(0.049)‡	(0.059)	(0.050)‡	(1.005)	(1.010)	(0.206)‡
ICE(FE-MTW)	−0.136	−0.170	−0.079	0.513	−1.098	−0.246	−0.559
	(0.098)	(0.055)‡	(0.061)	(0.048)‡	(0.923)	(0.993)	(0.231)‡
ICE(FE-FS)	−0.115	−0.164	−0.143	0.531	−0.757	−0.282	−0.672
	(0.085)	(0.047)‡	(0.061)‡	(0.047)‡	(0.911)	(0.972)	(0.204)‡
ICE(RE)	−0.198	−0.141	−0.138	0.745	−0.383	−1.114	−0.119
	(0.136)	(0.082)	(0.111)	(0.112)‡	(1.197)	(1.447)	(0.306)
ICE(RE-squeeze)	−0.183	−0.159	−0.113	0.712	−0.527	−0.863	−0.293
	(0.132)	(0.078)	(0.103)	(0.109)‡	(1.054)	(1.411)	(0.301)
MI using MID approach							
Multivariate normal imputation	−0.146	−0.207	−0.060	0.559	−0.465	−0.731	−0.604
	(0.084)	(0.049)‡	(0.058)	(0.049)‡	(0.924)	(0.993)	(0.204)‡
ICE(FE-MTW)	−0.116	−0.225	−0.040	0.567	−0.454	−0.699	−0.595
	(0.087)	(0.051)‡	(0.060)	(0.048)‡	(0.908)	(0.992)	(0.202)‡
ICE(FE-FS)	−0.117	−0.214	−0.054	0.573	−0.481	−0.701	−0.609
	(0.080)	(0.044)‡	(0.059)	(0.045)‡	(0.890)	(0.976)	(0.204)‡
ICE(RE)	−0.176	−0.200	−0.024	0.674	−0.520	−0.677	−0.576
	(0.093)	(0.053)‡	(0.065)	(0.054)‡	(0.954)	(1.028)	(0.203)‡
ICE(RE-squeeze)	−0.151	−0.213	−0.028	0.591	−0.497	−0.683	−0.582
	(0.091)	(0.052)‡	(0.064)	(0.051)‡	(0.946)	(1.002)	(0.203)‡

†Incomplete explanatory variables.
‡The p-value of the coefficient is less than 0.05 and the 95% confidence interval does not include 0.
§The Bayesian credible interval does not include 0.

The various methods sometimes produced substantially different estimates of the regression coefficients and their standard errors (Table 3). The agreement was somewhat better, in particular between model ICE(RE) and the other methods, when the MID approach was used as the exclusion of time points with missing outcomes could have resulted in less noise in the estimation. Using univariate or multivariate imputation models within the Bayesian MI also resulted in different estimates. The main difference between the standard MI and the MID approaches was observed for β_3 (hltsys). Investigation of the imputed data showed that the hltsys-score was higher on average for patients with complete outcomes compared with those with incomplete outcomes. Therefore, when the time points with incomplete outcome are excluded from the analysis model, the smaller values of hltsys are not used in the analysis. This may have resulted in the underestimation of the negative association between concare and hltsys. However, the numbers are too small to draw any definite conclusions. Model ICE(RE) which required eight

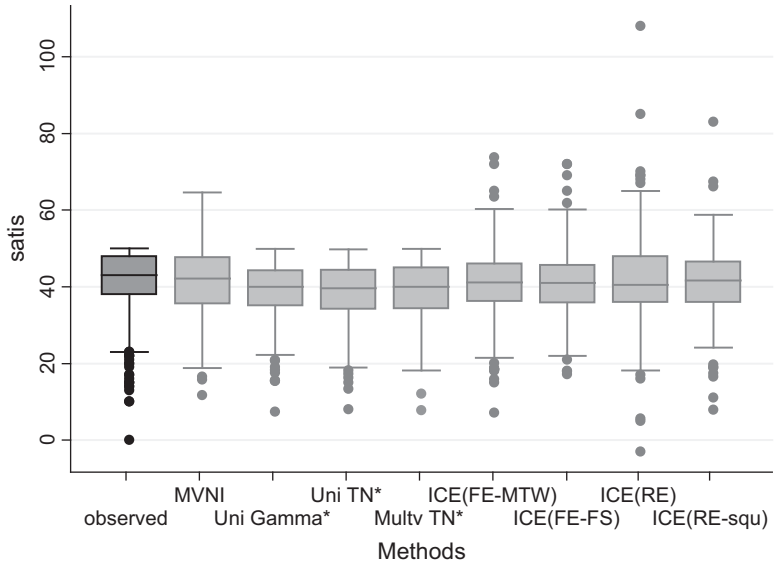

Fig. 1. Boxplots of observed and imputed values of satis based on a randomly selected imputed data set with the various MI approaches (∗, the Bayesian MI method using different imputation models)

random parameters to be estimated by the imputation model produced relatively large standard errors.

The observed and the imputed values of the incomplete variables were compared to investigate whether the imputed values appeared to be reasonable. For the skewed continuous variables phys, psyc, hltsys and satis, the distributions of the imputed values produced by the multivariate normal imputation and ICE methods were considerably different from their observed values. The agreement was good for Bayesian MI. The imputed values for satis are presented in Fig. 1.

5. Simulations

A simulation study based on the cancer care data was conducted to compare the performance of the MI methods for both dropout and item missingness by using 500 simulations. The bias $\Sigma_{q=1}^{500}(\hat{\beta}_{qk} - \beta_k^{\text{true}})/500$, percentage relative bias $(\text{bias}_{\beta_k}/\beta_k^{\text{true}}) \times 100$, coverage of nominal 95% confidence intervals (CIs) (the proportion of times that the 95% CIs include β_k^{true}, for $i = 1, \ldots, 500$), model-based standard errors $\Sigma_{q=1}^{500}\text{SE}(\hat{\beta}_{qk})/500$, empirical standard errors $\sqrt{\{(1/499)\,\Sigma_{q=1}^{500}(\hat{\beta}_{qk} - \hat{\beta}_k)^2\}}$ and the relative mean-squared errors as given below were calculated for the regression coefficients for each imputation method:

$$\text{RMSE}_{\text{MI/AC}} = \frac{\text{SE}(\hat{\beta}_k^{\text{MI}})^2 + (\text{bias}_{\beta k}^{\text{MI}})^2}{\text{SE}(\hat{\beta}_k^{\text{AC}})^2 + (\text{bias}_{\beta k}^{\text{AC}})^2} \times 100.$$

All missing variables in the real data were initially imputed using the model ICE(FE-MTW) method to provide a complete data set for the simulation. The explanatory variables were standardized and a random-intercept model (as in equation (1)) was fitted to the data. Data sets were generated by simulating new outcomes for each of these data sets, assuming that this model was true with $u_i$and e_{ij}drawn from $N(0, 0.4^2)$ and $N(0, 0.7^2)$ and combining these with the values of the explanatory variables from the complete data.

5.1. Simulations scenario 1: incomplete non-normal continuous explanatory variables, item missingness

Since MI is known to be most useful for imputing missing explanatory variables at measurement time points with observed outcome, this simulation scenario was constructed by keeping the outcome and the explanatory variable phys complete at all time points. All the other time varying variables were complete only at time point 1. Missing values were generated for three of the explanatory variables (psyc, hltsys, satis) at subsequent time points. The missingness probability for each incomplete variable at time point j for patient i was defined as follows:

$$\begin{aligned} p_{ij}^{\text{psyc}} &= P(\text{psyc}_i \text{is missing at time} j) \\ &= \text{invlogit}(\delta_{ij0}^{\text{psyc}} + \delta_{ij1}^{\text{psyc}} \text{phys}_{ij} + \delta_{ij2}^{\text{psyc}} \text{concare}_{i(j-1)}), \end{aligned} \tag{6}$$

$$\begin{aligned} p_{ij}^{\text{hltsys}} &= P(\text{hltsys}_i \text{ is missing at time} j) \\ &= \text{invlogit}(\delta_{ij0}^{\text{hltsys}} + \delta_{ij1}^{\text{hltsys}} \text{phys}_{ij} + \delta_{ij2}^{\text{hltsys}} \text{concare}_{i(j-1)}), \end{aligned} \tag{7}$$

$$\begin{aligned} p_{ij}^{\text{satis}} &= P(\text{satis}_i \text{ is missing at time } j) \\ &= \text{invlogit}(\delta_{ij0}^{\text{satis}} + \delta_{ij1}^{\text{satis}} \text{phys}_{ij} + \delta_{ij2}^{\text{satis}} \text{concare}_{i(j-1)}), \end{aligned} \tag{8}$$

where $\delta_{ij1}^{x} > 0$ and $\delta_{i2}^{x} < 0$, implying that the patients with higher physical needs and lower continuity-of-care score at a previous time point are more likely to have missing data. Specifically the probabilities of missingness for these variables at a particular time point were allowed to depend on the value of phys measured at the same time point and the value of concare measured at a previous time point. The methods were compared under different degrees of missingness by setting 50% and 20% of the time points to have missing values on one of the explanatory variables, i.e. psyc, hltsys and/or satis. The percentage of time points with missing explanatory variables but complete outcome was considerably higher in the simulated data compared with the real data. This allowed the assessment of the relative gain in using the MI methods over the AC analysis as, when considerable information is available from the outcome, the MI methods are expected to be superior to the AC analysis. The correlation between the repeated measurements of the incomplete variables was unstructured, following the correlation patterns observed in the real data.

The biases of the regression coefficients are presented in Figs 2 and 3, and more detailed results on the nominal coverage of the 95% CIs, mean-squared errors (MSEs) and standard errors can be found in Tables 1 and 2 in the on-line appendix. For the incomplete variables, multivariate normal imputation and Bayesian MI using multivariate truncated normal models produced the least bias for the estimates of the regression coefficients and showed reasonable performance for imputing non-normal continuous variables, as also found by Demirtas *et al.* (2008), even if the proportion of missingness was as large as 50%. In terms of the coverage of the 95% CIs, all methods except for models ICE(FE-FS) and ICE(RE) performed well. The model-based standard error and empirical standard errors were close to each other. The gains in MSEs in incomplete variables with MI approaches compared with the AC analysis were more significant with Bayesian MI because of the smaller standard errors. When the proportion of missingness was 50%, the gain in MSE was more obvious. AC analysis always produced greater bias compared with multivariate normal imputation. AC analysis also produced bias for the complete time varying variables. MI uses information from the observed variables at the incomplete time points, whereas this does not happen with AC analysis. Therefore, although the AC analysis adjusts for phys, the results do not fully account for the reasons of missingness; for example the values of phys are excluded from the analysis at the incomplete time points. For the

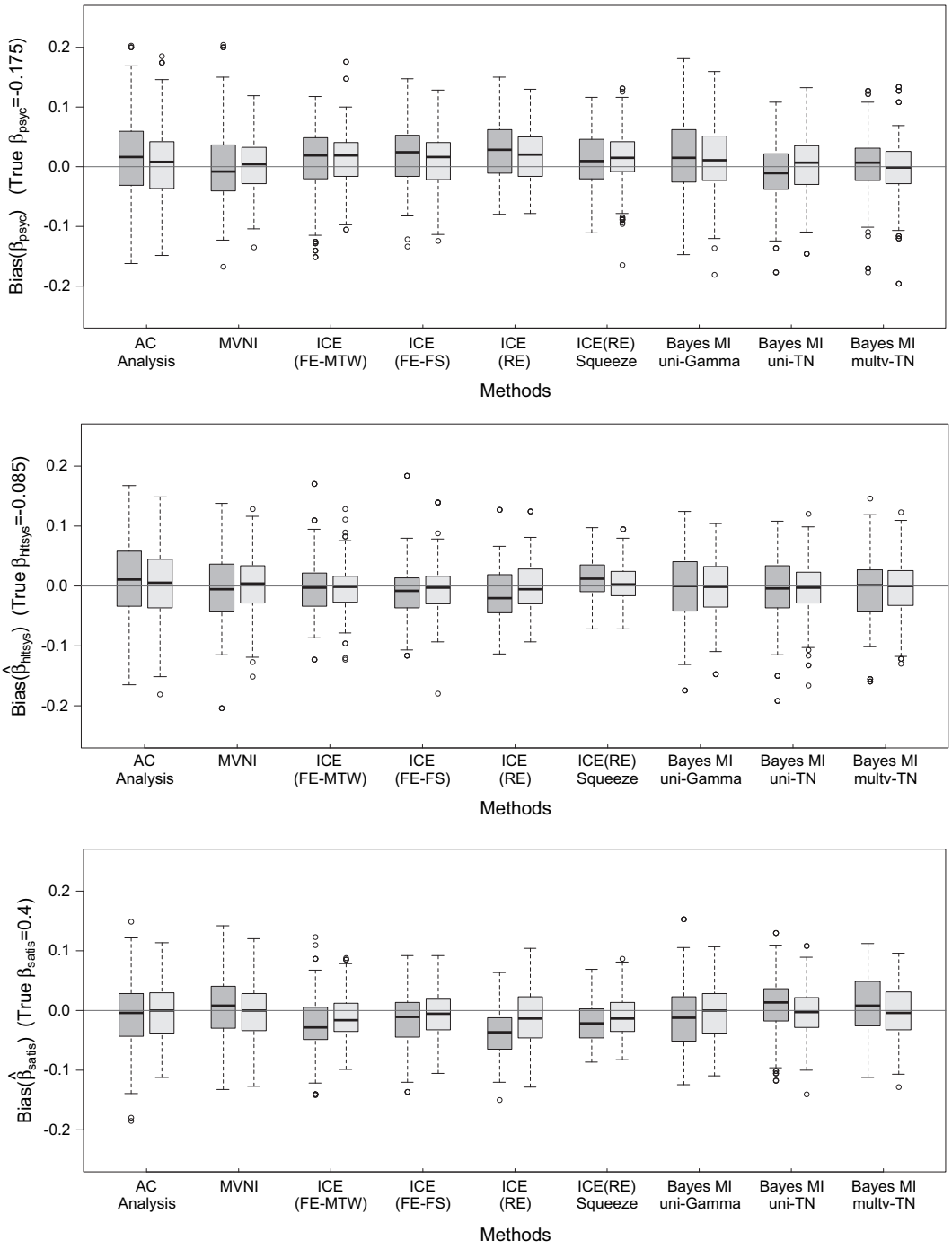

Fig. 2. Simulation scenario 1 (incomplete explanatory variables, with fully observed outcome: boxplots of bias($\hat{\beta}_k$) for incomplete variables (■, 50% missingness; □, 20% missingness)

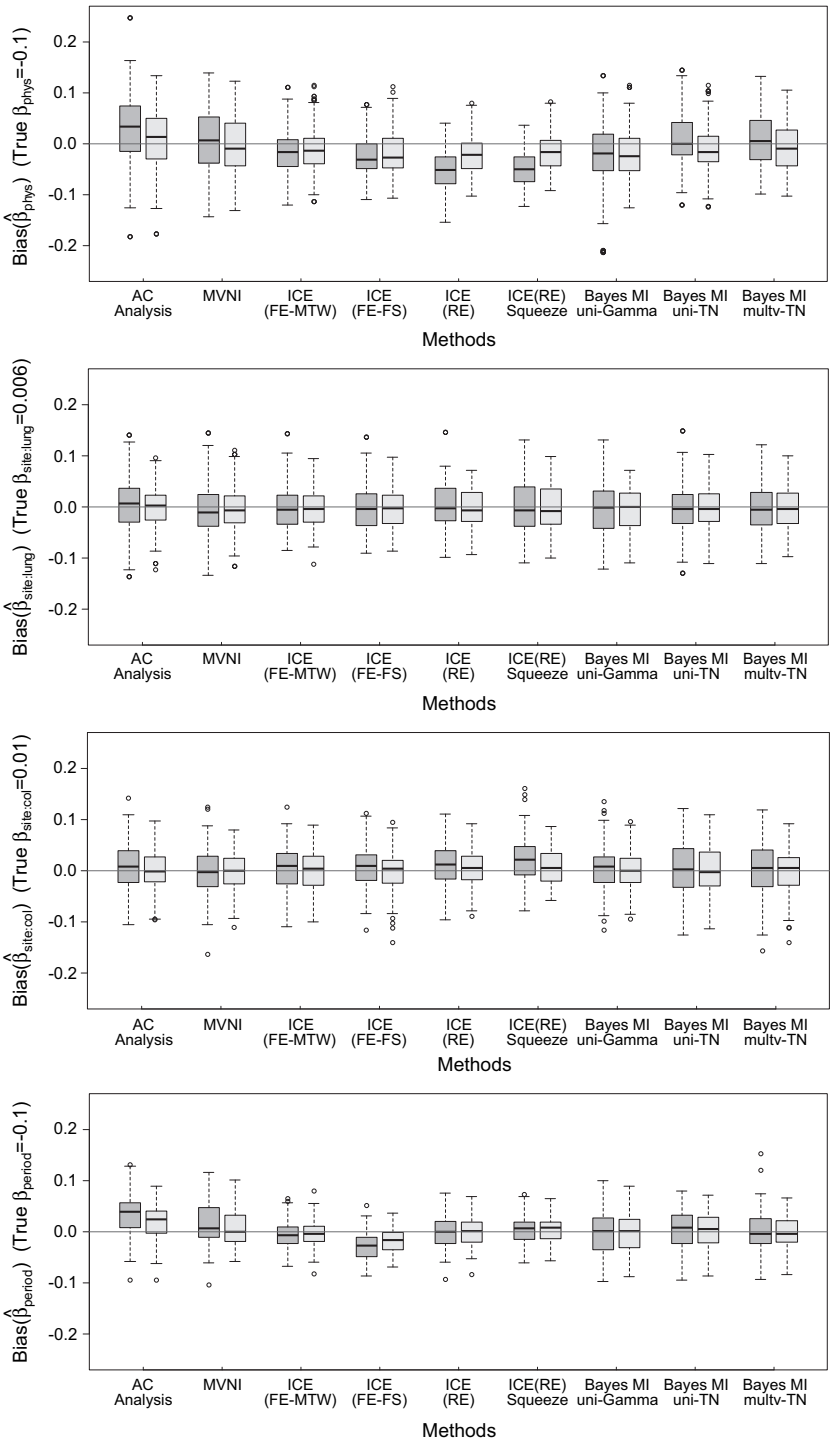

Fig. 3. Simulation scenario 1 (incomplete explanatory variables, with fully observed outcome: boxplots of bias($\hat{\beta}_k$) for fully observed variables (■, 50% missingness; □, 20% missingness)

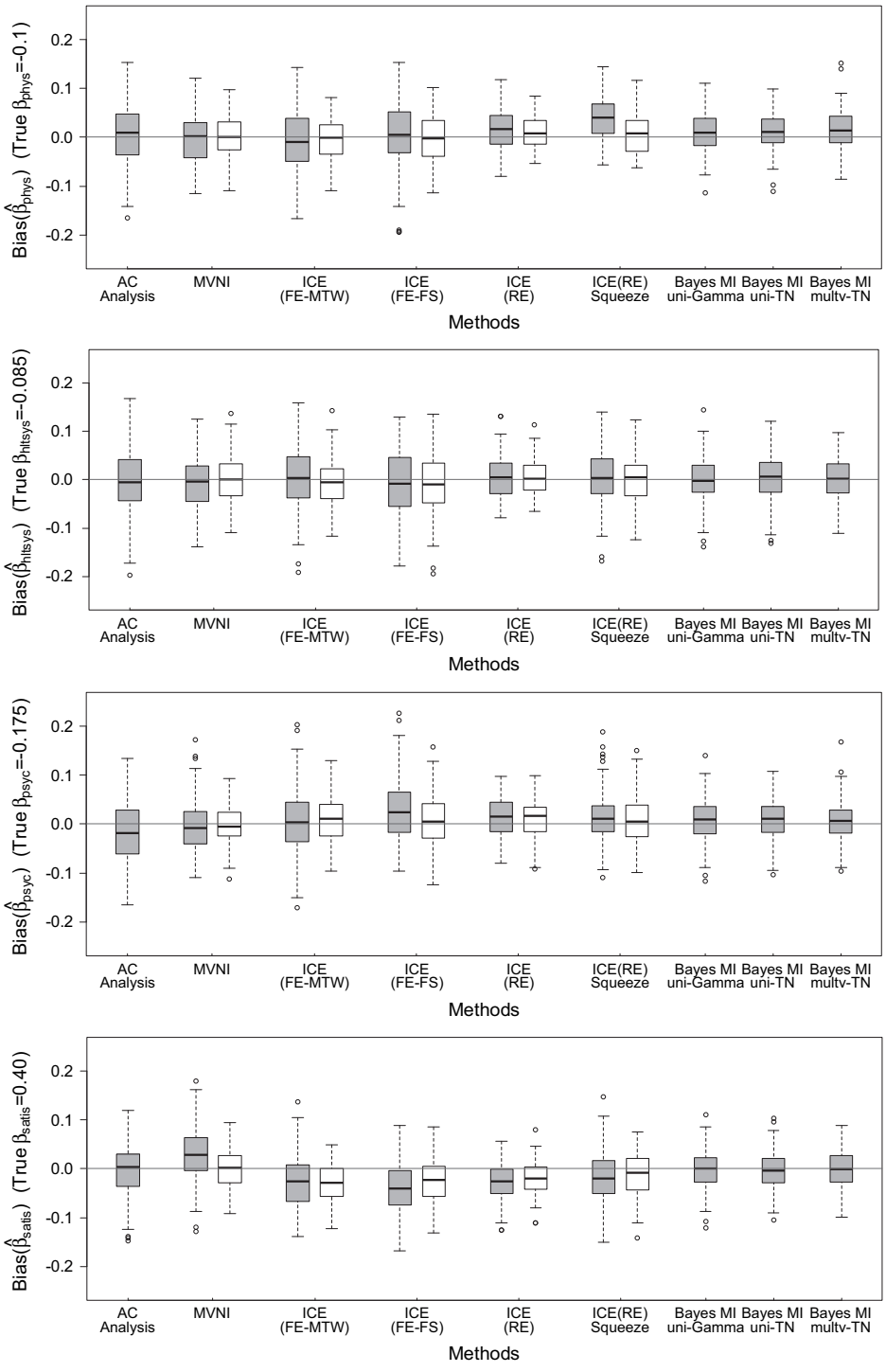

Fig. 4. Simulation scenario 2 (incomplete outcome and explanatory variables): boxplots of bias($\hat{\beta}_k$) for incomplete variables (■, time point with missing outcome used in the analysis; □, MID approach)

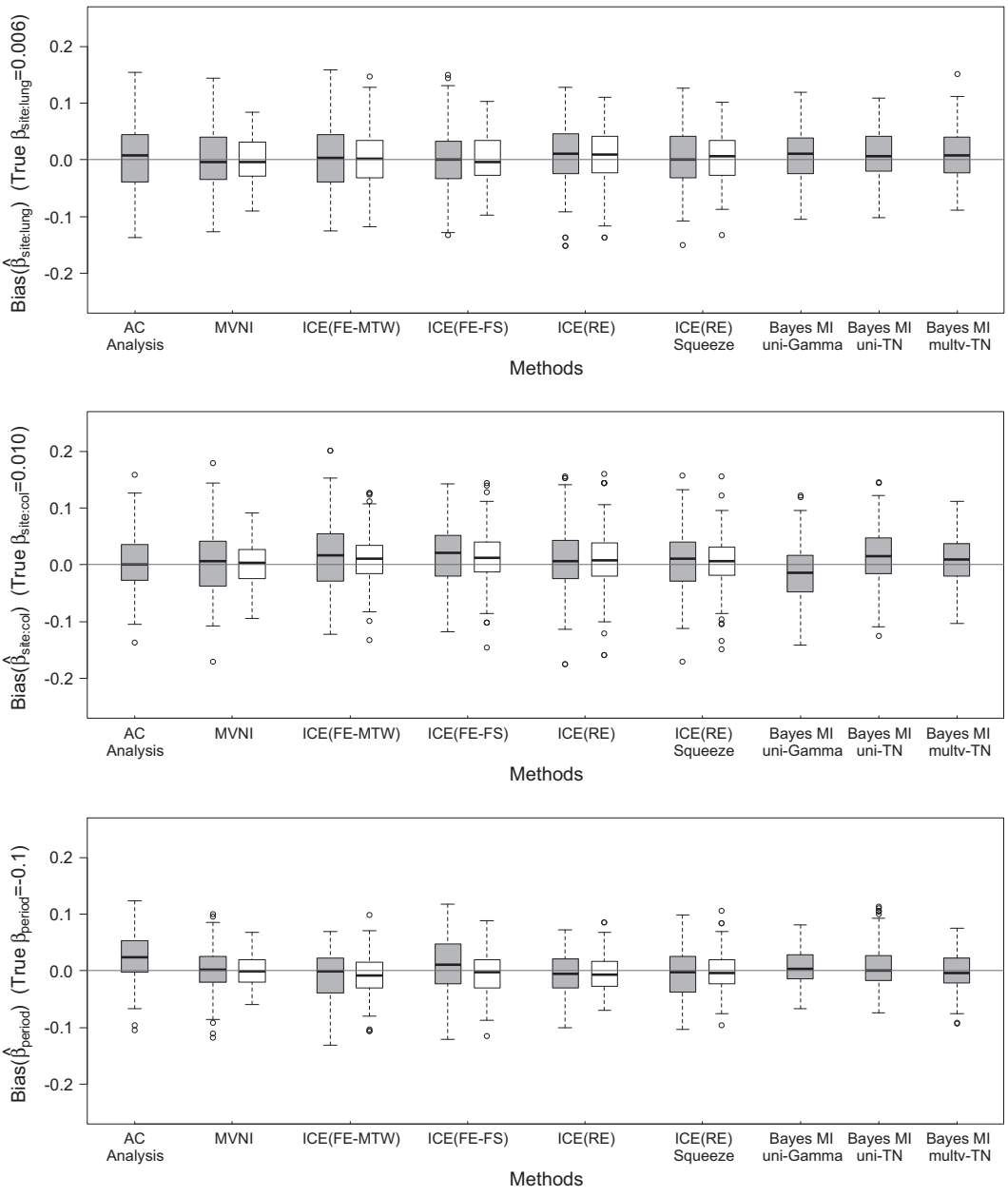

Fig. 5. Simulation scenario 2 (incomplete outcome and explanatory variables): boxplots of bias($\hat{\beta}_k$) for fully observed variables (■, time points with missing outcome used in the analysis; □, MID approach)

complete variables, multivariate normal imputation and Bayesian MI produced similar root-mean-squared-error and performed better than the ICE approaches. The relatively worse performance for model ICE(FE-MTW) could be due to the violation of the model assumptions such as the misspecification of the correlation structure in the imputation models and assumptions of normality for the incomplete variables. This is also a problem for model ICE(FE-FS) which

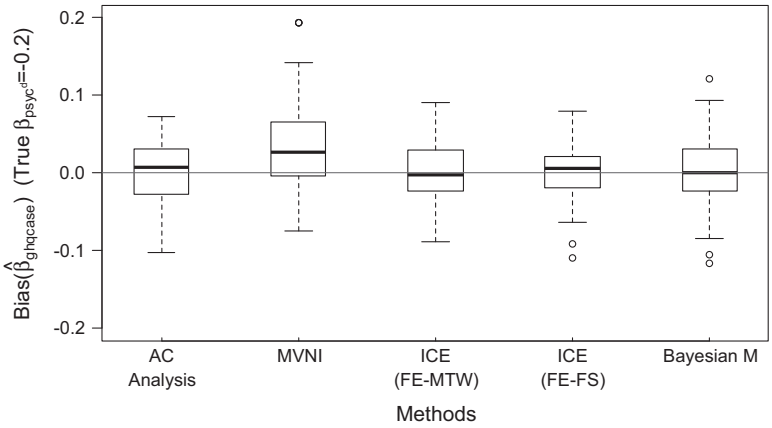

Fig. 6. Simulation scenario 3 (non-normally distributed continuous and binary variables): boxplots of bias($\hat{\beta}_{ghqcase}$)

does not assume any specific correlation between the repeated measurements of the incomplete variables; moreover it is prone to variable selection problems (Harrell, 2001). In contrast multivariate normal imputation assumes unstructured correlation, thus representing the true correlation structure. Model ICE(RE) performed the worst for both the complete and the incomplete variables; however, when the imputations were constrained to be between the minimum and maximum values that the variables can take, estimates obtained with model ICE(RE) substantially improved in terms of bias and coverage.

The Bayesian MI was sensitive to the assumption of the distribution of incomplete variables. Use of gamma imputation models to impute right-skewed variables resulted in bias for the regression coefficient of the variable psyc, though it was smaller than that observed for models ICE(RE) and ICE(FE-FS). The multivariate truncated normal imputation model performed the best. When univariate imputation models were used, Bayesian MI also did not correctly specify the correlation structure in the imputation model. When the correct unstructured correlation was specified by including time points as dummy variables and treating these as random in the univariate imputation models, it did not improve the bias and the efficiency was reduced (the results are not shown).

To assess the effect of reducing the sample size and the planned number of measurement time points per patient,

(a) 100 patients were selected from the complete data set and
(b) the number of time points, j, reduced to 3 by excluding the last two time points of each patient, in turn.

The relative biases that were estimated with different methods did not substantially change, but the standard errors of bias increased for all the methods as expected. Specifically, the standard errors that were obtained with AC analysis increased the most. Thus the gain in the precision by using MI methods compared with AC analysis was greater when the sample size is small. The imputation method which was most affected because of small sample size and fewer measurement time points was model ICE(FE-MTW). It provided larger standard errors of estimates and hence larger MSEs because it uses many predictors in the imputation models and when these predictors are not fully observed the variability between the imputation models increases more with smaller sample sizes.

Table 4. Simulation scenario 3: incomplete non-normally distributed continuous, binary and categorical variables, with fully observed outcome†

Variable	*Results for the following analysis methods:*														
	Relative bias (%)			*Coverage (%)*			*RMSE (%)‡*			*Model-based standard error*			*Empirical standard error*		
	AC analysis	*ICE(FE-FS)*	*Bayesian MI§*	*AC analysis*	*ICE(FE-FS)*	*Bayesian MI§*	*AC analysis*	*ICE(FE-FS)*	*Bayesian MI§*	*AC analysis*	*ICE(FE-FS)*	*Bayesian MI§*	*AC Analysis*	*ICE(FE-FS)*	*Bayesian MI§*
Incomplete variables															
psyc (−0.175)	−4.5	−4.3	3.7	96.2	95.8	96.4	100.0	94.9	69.4	0.055	0.050	0.034	0.053	0.047	0.032
hltsys (−0.085)	3.8	11.6	2.9	94.6	92.2	95.6	100.0	108.3	52.4	0.051	0.046	0.033	0.048	0.042	0.033
satis (0.4)	2.2	−4.7	−1.3	94.0	95.4	95.6	100.0	88.3	59.1	0.046	0.042	0.033	0.044	0.040	0.032
ghqcase															
= 2 (−0.2)	6.5	5.6	3.2	93.0	92.4	95.2	100.0	91.7	62.4	0.088	0.076	0.073	0.087	0.074	0.070
trtphase															
= 2 (−0.1)	−6.4	8.1	−4.6	93.4	96.4	96.0	100.0	84.5	58.7	0.111	0.105	0.080	0.106	0.099	0.076
= 3 (−0.15)	−5.5	−4.2	−3.7	91.6	93.8	96.6	100.0	85.0	61.9	0.104	0.099	0.079	0.100	0.095	0.077
= 4 (−0.2)	−9.7	−5.8	−5.0	92.6	95.2	95.0	100.0	94.6	59.2	0.106	0.098	0.078	0.102	0.094	0.077
= 5 (−0.2)	−12.6	−13.0	−7.2	93.4	93.8	95.8	100.0	88.0	57.2	0.107	0.101	0.081	0.101	0.092	0.079
Fully observed variables															
phys (−0.1)	−14.2	6.5	−5.8	92.2	96.0	95.6	100.0	83.4	49.1	0.058	0.048	0.039	0.056	0.044	0.037
site: lung (0.006)	12.4	41.3	−17.8	93.4	94.2	96.0	100.0	84.5	75.7	0.105	0.096	0.094	0.099	0.093	0.095
site: col (0.01)	8.1	30.3	−17.7	98.4	95.2	97.8	100.0	84.9	64.3	0.111	0.100	0.090	0.101	0.094	0.091
period (−0.1)	−18.9	7.9	7.5	88.2	96.4	96.4	100.0	65.3	59.0	0.042	0.034	0.030	0.041	0.031	0.028

†Maximum Monte Carlo standard error 0.004.
‡RMSE is the percentage relative MSE of the imputation methods compared with AC analysis.
§Univariate truncated normal imputation models were used to impute psyc, hltsys and satis, a univariate hierarchical logistic regression model was used to impute ghqcase and a univariate hierarchical multinomial model was used to impute trtphase.

The main gain in using MI was in precision as it uses the observed information from all time points via imputations, whereas AC analysis discards all time points with incomplete observations from the analysis, even time points with only one explanatory variable missing.

5.2. Simulation scenario 2: incomplete outcome and non-normally distributed continuous explanatory variables, mixed patterns of missingness

In the continuity-of-care data, mixed patterns of missingness were observed. The majority of the patients dropped out from the study and did not come back. Item missingness was also observed for some of the patients. To compare the methods when all repeatedly measured variables have missing values as in the real data, in this simulation scenario 20% of the time points are defined to be missing due to dropout of the patients and 30% of the time points have missing values due to the item missingness for the four non-normally distributed explanatory variables.

In addition to the item missingness models defined in models (6)–(8), to satisfy the missingness at random assumption for the dropouts, the probability of dropout for patient i at time j was allowed to depend on their previously observed value of the outcome concare and the explanatory variable phys:

$$p_{ij}^{D} = P(\text{patient } i \text{ drops out at time point } j) = \text{invlogit}(\delta_{ij0} + \delta_{ij}\,\text{phys}_{i(j-1)} + \delta_{ij2}\,\text{concare}_{i(j-1)}) \tag{9}$$

where $\delta_{ij1} > 0$ and $\delta_{ij2} < 0$.

The results for the bias of the regression coefficients are presented in Figs 4 and 5, and Tables 3 and 4 in the on-line appendix present the more detailed results. Using the MID approach reduced the bias and MSE in general for all the MI methods except for model ICE(RE). Multivariate normal imputation using the MID approach was the least biased method. Bayesian MI approaches produced some bias. When 20% of the outcome data were missing, using the MID approach appeared to be a better strategy than including imputed outcomes in the analysis model. However, if there are many time points with missing outcome, not including these time points in the analysis stage may cause the analysis to be inefficient.

5.3. Simulation scenario 3: incomplete non-normal continuous and categorical variables

In the original data set, in addition to the four time varying skewed variables there were two time varying categorical variables: ghqcase (mental health score, binary) and trtphase (treatment phase, categorical, five categories: 1, initial diagnosis (prevalence = 14.1); 2, completion of first treatment (prevalence = 16.1); 3, remission (prevalence = 46.2); 4, relapse (prevalence = 10.3); 5, specialist palliative care (prevalence = 13.3)). The software implementation of the MI methods which can handle incomplete time varying categorical variables is limited. Therefore earlier in this paper these categorical variables were excluded to compare all the available MI methods which can handle incomplete time varying continuous variables. In this section, incomplete categorical variables are included in the simulations in addition to the three incomplete skewed variables. The proportion of the missing values in each variable is defined such that the proportion of time points with missing values is 50%.

The previous simulation scenarios showed that multivariate normal imputation gave promising results for the imputation of both normal and non-normal continuous variables. However, multivariate normal imputation assumes latent multivariate normality for categorical variables (Goldstein *et al.*, 2009). Using rounding schemes within the continuous imputations as in multivariate normal imputation may lead to serious bias, especially for nominal explanatory variables (Vermunt *et al.*, 2008). In contrast, fixed effects ICE and Bayesian MI approaches, which use

univariate logistic or multinomial regression models, may be used to impute the binary or categorical variables. Model ICE(RE) was excluded from this comparison as the available software can impute continuous variables only.

5.3.1. Incomplete non-normal continuous and binary variables

Missingness is additionally defined for the binary variable ghqcase, which had a prevalence of 35%. The simulation study showed that multivariate normal imputation which assumes multivariate normality for a mixture of continuous and binary variables produces biased estimates and poor coverage of 95% CIs for the coefficient of the incomplete binary variable (Fig. 6). When multivariate normality is assumed for only incomplete continuous variables and ghqcase is imputed with a univariate hierarchical logistic regression model, as in the Bayesian MI method, the coefficient of the incomplete binary variable is estimated with the least bias. The fixed effects ICE approaches and Bayesian MI provided approximately 40–50% gains in efficiency compared with the AC analysis and multivariate normal imputation (see Table 5 in the on-line appendix).

5.3.2. Incomplete non-normal continuous, binary and categorical variables

Multivariate normal imputation using MLwiN macros is not suitable for imputing nominal variables as it assumes that these variables are normally distributed. Instead, REALCOM-Impute software (Carpenter *et al*., 2011) can be used to apply multivariate normal imputation in conjunction with the multilevel software MLwiN or Stata. REALCOM-Impute can handle categorical as well as normal variables with multilevel structures (Carpenter *et al*., 2011). However, running simulations with REALCOM is not straightforward as each analysis needs to be done individually. Therefore multivariate normal imputation is not included in this simulation scenario, but the results from an analysis of a single data set with incomplete binary and categorical variables with multivariate normal imputation using REALCOM-Impute in conjuction with Stata is presented in the appendix C in the on-line appendix. For applying the fixed effects ICE methods, multinomial logistic regression is used to impute the categorical variable trtphase. Among model ICE(FE) approaches when the moving time window approach was used there was a problem of overfitting of imputation models. This is due to the large number of incomplete variables and the large number of predictors in the imputation models. For Bayesian MI, a hierarchical multinomial logistic model is defined in addition to the other univariate imputation models. The results of the simulations for the AC analysis, model ICE(FE-FS) and Bayesian MI are presented in Table 4.

Bayesian MI using a univariate hierarchical multinomial imputation model produced the least bias for the coefficients of trtphase. The coverages of 95% CIs for these coefficients were close to the nominal value for all the methods. Owing to the poor performance of model ICE(FE-FS) for the estimates of continuous variables, Bayesian MI is preferred when there are incomplete continuous and categorical variables in the data set, provided that there are fully observed variables in the data set that can be used as predictors in the imputation model.

5.4. Simulation scenario 4: artificial data for the repeated measurements of the incomplete normally distributed explanatory variables with different correlation structures

To investigate whether the relatively poor performance of the ICE approaches was due to the presence of non-normally distributed continuous variables, a simulation study was carried out by generating five explanatory variables for $N = 199$ subjects from a multivariate normal distribution. Four of them were generated to correspond to the time varying explanatory variables phys, psyc, hltsys and satis, measured at five time points $j = 1, \ldots, 5$. The remaining variable

Table 5. Choice of MI method depending on the correlation structure between the repeated measurements

Correlation between repeated measurements of incomplete variables	*Choice of method for the following incomplete variable types:*			
	Univariate normal	*Multivariate normal*	*Non-normal continuous*	*Mixture of normal binary or categorical†*
Unstructured	Multivariate normal imputation ICE(RE) Bayesian MI‡	Multivariate normal imputation Bayesian MI‡	Multivariate normal imputation Bayesian MI‡	Bayesian MI‡
Exchangeable	Multivariate normal imputation ICE(RE) Bayesian MI‡	Multivariate normal imputation Bayesian MI‡	Multivariate normal imputation Bayesian MI‡	Bayesian MI‡
AR(1)	Multivariate normal imputation ICE(FE-MTW) Bayesian MI‡	Multivariate normal imputation ICE(FE-MTW) Bayesian MI‡	Multivariate normal imputation Bayesian MI‡	ICE(FE-MTW) Bayesian MI‡

†Based on the results from simulation scenario 3.
‡Assuming the true correlation structure between repeated measurement of the incomplete variables.

represents the fully observed categorical explanatory variable which is constant over time, site. After generating normally distributed continuous values for site, the subjects were assigned to any of the three cancer site groups on the basis of the percentiles of the latent normally distributed variable, by keeping the prevalences of each group the same as in the continuity-of-care data.

The random vector of explanatory variables was drawn from the 21-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}_{21\times1} = (0, 0, \ldots, 0)^{\mathrm{T}}$ and the covariance matrix $\boldsymbol{\Omega}_{21\times21}$. The dimension of the distribution was specified as 21 to match the number of explanatory variables and time points, i.e. four time varying variables were measured at five time points and a baseline variable was constant over time ($4 \times 5 + 1 \times 1$). Additionally in this section the performances of the MI methods, when the true correlation between the repeated measurements of incomplete variables is not captured by the imputation models, was also investigated. Multivariate normal data sets were generated assuming different true correlation structures. The desired covariance matrices were obtained by specifying the correlations between repeatedly measured variables to be as follows:

(a) unstructured, the correlation structure between the repeated measurements and the explanatory variables was kept the same as in the continuity-of-care data,
(b) exchangeable, with correlation 0.4 between the repeated measurements as well as between any of the two variables, and
(c) first-order auto-regressive (AR(1)), with correlation 0.6 between two consecutive time points and 0.4 between the variables measured at the same time point.

The outcome data were simulated 500 times following the random-intercepts model given in equation (1). 50% item missingness for each variable was then defined by using equations (6)–(8) given in Section 5.1. The correlations between variables averaged over 500 simulated data sets are given in Table 6 and the results are presented in Figs 2 and 3, and Table 7 in the on-line appendix.

Multivariate normal imputation still outperformed all the ICE approaches under true unstructured and exchangeable correlation structures, which could be because of its correct spec-

ification of the true correlation structure. Bayesian MI with multivariate normal imputation models which assume the true correlation structure resulted in small bias. Using univariate Bayesian normal imputation models resulted in slightly more bias than using multivariate models. This may be due to the fact that Bayesian univariate imputation models do not account for the correlations between the incomplete variables. When the true correlation is unstructured, the fixed effects ICE methods resulted in more bias than does AC analysis. The performance of model ICE(FE-MTW) improved considerably for the AR(1) process and was comparable with that of multivariate normal imputation in terms of bias and efficiency.

A simulation scenario was also carried out by reducing the number of incomplete normally distributed variables to 1. Only three of the explanatory variables, namely phys, psyc and period, were kept in the data sets. 40% item missingness was imposed only on psyc, keeping the outcome concare and the other explanatory variables complete. All the methods estimated the regression coefficient of psyc with small bias, except for model ICE(FE-FS) (see Fig. 4 in the on-line appendix). Model ICE(RE) was observed to be the best method in terms of precision (see Table 8 in the on-line appendix).

6. Discussion

The performances of the MI methods and AC analysis were compared for data missing at random in repeated measurements observational studies in this paper and the strengths and weaknesses of each MI method were highlighted. For item missingness, where missingness occurred in the explanatory variables with the outcomes fully observed, MI has the potential to produce more precise estimates relative to those obtained from using the AC analysis. The preferred MI methods for imputing different types of incomplete variables are summarized in Table 5. When the incomplete variables are continuous, the multivariate normal imputation method produced the least bias. An assumption of multivariate normality did not appear to cause bias when imputing non-normal continuous variables as also observed by Demirtas *et al.* (2008). Additionally, multivariate normal imputation is theoretically well justified and was least sensitive to the different correlation assumptions, as it assumes the most flexible correlation structure for the repeated measurements of the incomplete variables, i.e. unstructured correlations. Therefore it can be recommended for the imputation of continuous variables in repeated measurement studies. However, Bayesian MI may be preferable when there is a mixture of incomplete continuous and categorical variables in the data, particularly for categorical variables. The Bayesian method also offers flexibility regarding the choice of distributions for the continuous variables, which may be useful when handling non-normal continuous variables with missing values. However, it may be important to consider the appropriateness of the chosen distributions carefully, as an incorrect choice may lead to bias. The distribution of the incomplete variables should be inspected in detail, e.g. by using Q–Q-plots and by comparing the distributions of the imputed and observed values. Another advantage of Bayesian MI is that the uncertainty arising from imputing missing values is directly incorporated through the number of iterations used by the MCMC algorithm (Lunn *et al.*, 2009). Bayesian MI using univariate imputation models could perform poorly when there are no fully observed variables in the data. Although in our simulation scenarios it performed well in comparison with the ICE methods, in practice caution is needed when using univariate Bayesian MI when there are no fully observed variables.

The ICE approaches were sensitive to the correlation structures for the repeated measurements of the incomplete variables. The performance of these methods for imputing continuous incomplete variables also depended on whether these variables satisfied the assumptions of normality. Model ICE(FE-MTW) could be used for normally distributed incomplete continuous

variables with auto-regressive correlation between the repeated measurements. The fixed effects ICE approach could be used under the same conditions only if there are sufficient data to fit the imputation model as all time points are included. The random-coefficients imputation models used by model ICE(RE) allow a complex correlation structure between the repeated measurements of the incomplete variables but may have convergence problems due to the large number of random parameters being estimated.

As a viable alternative to standard and Bayesian imputation methods for imputing categorical variables, Vermunt *et al*. (2008) proposed a latent class model for imputing several categorical variables by using a mixture model of independent multinomial distributions. van der Palm *et al*. (2012) showed that, when the number of missing data patterns for the categorical variables is large, the latent class model approach is superior to ICE. In addition to using a different imputation model, the latent class approach differs from multivariate normal imputation and ICE in that the imputations are based on a fully frequentist framework that uses the non-parametric bootstrap. Random-effects latent class models can be fitted by using Latent Gold 4.5 software (Vermunt and Magidson, 2008); however, in this paper we have focused on the MI methods which are available in free software (e.g. MLwiN, WinBUGS and R) or software which is commonly used in practice (e.g. Stata).

AC analysis showed greater bias in regression coefficient estimates of complete time varying explanatory variables that were related to missingness. When the estimates of incomplete variables are considered, MI methods did not always necessarily produce less bias compared with the AC analysis but nevertheless provided greater precision in general. Therefore, MI methods may be preferable to AC analysis in the presence of missing data in the explanatory variables. Aucejo *et al*. (2013) have also focused on the case when covariates are missing and proposed a new strategy which conducts sets and imputes bounds for missing data.

A complete missing data analysis requires the investigation of departures from missingness at random to a missingness not at random mechanism through a sensitivity analysis of parameter estimates (Sterne *et al*., 2009). However, when there are several incomplete variables with complicated missingness patterns, the modelling approaches which are designed to conduct a sensitivity analysis are not straightforward and are beyond the scope of this paper.

In this paper, we highlighted the issues that can arise in applying MI methods in the presence of various features of repeated measurements data. Although we assessed the methods under different settings of missing data with several simulation scenarios, caution is needed to refrain from overgeneralizing the results which are based on a particular data set. A detailed investigation of the missingness pattern, degree of missingness, distribution of the incomplete variables, prevalence of the categorical variables and the correlation between the repeated measurements of the incomplete variables should be carried out to choose the optimal imputation method.

Acknowledgements

The authors thank Ian White for his valuable comments, which substantially improved an earlier draft. Oya Kalaycioglu was supported by the Ministry of National Education of Turkey within the graduate studies scholarship scheme.

References

Aucejo, E. M., Bugni, F. A. and Hotz, V. J. (2013) Identification and inference on regressions with missing covariate data. London School of Economics and Political Science, London. (Available from `http://public.econ.duke.edu/~vjh3/working_papers/missing_x.pdf`.)

van Buuren, S. (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Statist. Meth. Med. Res.*, **16**, 219–242.
van Buuren, S. and Oudshoorn, K. G. (2000) Multivariate imputation by chained equations: MICE V1.0 user's manual. *Report PG/VGZ/00.038*. Toegepast Natuurwetenschappelijk Onderzoek, Leiden. (Available from `http://web.inter.nl.net/users/S.van.Buuren/mi/docs/Manual.pdf`.)
van Buuren, S. and Oudshoorn, K. G. (2011) MICE: Multivariate imputation by chained equations in R. *J. Statist. Softwr.*, **45**, no. 3.
Carpenter, J. and Goldstein, H. (2005) Multiple imputation using MLwiN. *Multilev. Modllng Newslett.*, **16**, 9–18.
Carpenter, J. R., Goldstein, H. and Kenward, M. G. (2011) REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *J. Statist. Softwr*, **45**, no. 4.
Carpenter, J. and Kenward, M. (2005) Example analyses using WinBUGS 1.4. Department of Statistics, London School of Hygiene and Tropical Medicine, London. (Available from `http://www.missingdata.org.uk`.)
Carrigan, G., Barnett, A. G., Dobson, A. J. and Mishra, G. D. (2007) Compensating for missing data from longitudinal studies using WinBugs. *J. Statist. Softwr.*, **19**, no. 7.
Demirtas, H., Freels, S. A. and Yucel, R. M. (2008) Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *J. Statist. Computn Simuln*, **78**, 69–84.
Fitzmaurice, G. M. (2003) Methods for handling dropouts in longitudinal clinical trials. *Statist. Neerland.*, **57**, 75–99.
Goldstein, H., Carpenter, J., Kenward, M. G. and Levin, K. A. (2009) Multilevel models with multivariate mixed response types. *Statist. Modllng*, **9**, 173–197.
Harrell, F. E. (2001) *Regression Modelling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.
von Hippel, P. T. (2007) Regression with missing Y's: an improved strategy for analyzing multiply imputed data. *Sociol. Methodol.*, **37**, 83–117.
von Hippel, P. T. (2009) How to impute interactions, squares and other transformed variables. *Sociol. Methodol.*, **39**, 265–291.
King, M., Jones, L., Richardson, A., Murad, S., Irving, A., Aslett, H., Ramsay, A., Coelho, H., Andreou, P., Tookman, A., Mason, C. and Nazareth, I. (2008) The relationship between patients' experiences of continuity of cancer care and health outcomes: a mixed methods study. *Br. J. Cancer*, **98**, 529–536.
Lee, K. J. and Carlin, J. B. (2010) Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *Am. J. Epidem.*, **171**, 624–632.
Lunn, D., Best, N., Spiegelhalter, D., Graham, G. and Neuenschwander, B. (2009) Combining MCMC with 'sequential' PKPD modelling. *J. Pharmokinet. Pharmodyn.*, **36**, 19–38.
Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000) WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statist. Comput.*, **10**, 325–337.
Molenberghs, G. and Kenward, M. G. (2007) *Missing Data in Clinical Studies*. Chichester: Wiley.
Nevalainen, J., Kenward, M. G. and Virtanen, S. M. (2009) Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. *Statist. Med.*, **28**, 3657–3669.
van der Palm, D. W., van der Ark, L. A. and Vermunt, J. K. (2012) A comparison of incomplete-data methods for categorical data. *Statist. Meth. Med. Res.*, to be published, doi 10.1177/0962280212465502.
Paskett, E. D., Naughton, M. J., McCoy, T. P., Case, L. D. and Abbott, J. M. (2007) The epidemiology of arm and hand swelling in premenopausal breast cancer survivors. *Cancer Epidem. Biomark. Prevn*, **16**, 775–782.
Peyre, H., Leplege, A. and Coste, J. (2011) Missing data methods for dealing with missing items in quality of life questionnaires: a comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Qual. Life Res.*, **28**, 287–300.
Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. and Solenberger, P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.*, **27**, 85–95.
R Development Core Team (2008) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Royston, P. (2005) Multiple imputation of missing values: update. *Stata J.*, **5**, 188–201.
Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
Rubin, D. B. (2008) Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 20–28.
Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
StataCorp (2011) *Stata Statistical Software: Release 12*. College Station: StataCorp.
Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. and Carpenter, J. R. (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Br. Med. J.*, **338**, article b2393.
Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A. and Sijtsma, K. (2008) Multiple imputation of incomplete categorical data using latent class analysis. *Sociol. Methodol.*, **38**, 369–397.

Vermunt, J. K. and Magidson, J. (2008) *LG-syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module*. Belmont: Statistical Innovations.

White, I. R. and Carlin, J. B. (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statist. Med.*, **29**, 2920–2931.

White, I. R., Royston, P. and Wood, A. M. (2011) Multiple imputation using chained equations: issues and guidance for practice. *Statist. Med.*, **30**, 377–399.

Yu, L. M., Burton, A. and Rivero-Arias, O. (2007) Evaluation of software for multiple imputation of semi-continuous data. *Statist. Meth. Med. Res.*, **16**, 243–258.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Online appendix to A comparison of multiple imputation methods for handling missing data in repeated measurements observational studies'.