BMC
Medical Research Methodology

Joint modelling rationale for chained equations

Rachael A Hughes[1*], Ian R White[2], Shaun R Seaman[2], James R Carpenter[3,4], Kate Tilling[1]
and Jonathan AC Sterne[1]

Abstract

Background: Chained equations imputation is widely used in medical research. It uses a set of conditional models, so is more flexible than joint modelling imputation for the imputation of different types of variables (e.g. binary, ordinal or unordered categorical). However, chained equations imputation does not correspond to drawing from a joint distribution when the conditional models are incompatible. Concurrently with our work, other authors have shown the equivalence of the two imputation methods in finite samples.

Methods: Taking a different approach, we prove, in finite samples, sufficient conditions for chained equations and joint modelling to yield imputations from the same predictive distribution. Further, we apply this proof in four specific cases and conduct a simulation study which explores the consequences when the conditional models are compatible but the conditions otherwise are not satisfied.

Results: We provide an additional "non-informative margins" condition which, together with compatibility, is sufficient. We show that the non-informative margins condition is not satisfied, despite compatible conditional models, in a situation as simple as two continuous variables and one binary variable. Our simulation study demonstrates that as a consequence of this violation order effects can occur; that is, systematic differences depending upon the ordering of the variables in the chained equations algorithm. However, the order effects appear to be small, especially when associations between variables are weak.

Conclusions: Since chained equations is typically used in medical research for datasets with different types of variables, researchers must be aware that order effects are likely to be ubiquitous, but our results suggest they may be small enough to be negligible.

Keywords: Chained equations imputation, Gibbs sampling, Joint modelling imputation, Multiple imputation, Multivariate missing data

Background

Multiple imputation [1] has become a popular approach for the analysis of incomplete data, with several mainstream statistical packages now incorporating multiple imputation tools. It involves making several draws of the missing data from their posterior predictive distribution given the observed data and an imputation model. For multivariate, non-monotone missing data there are two main approaches for constructing an imputation model: joint modelling and chained equations. Joint modelling imputation requires the specification of a parametric joint model for the complete data: current implementations impute under the multivariate normal model, the log linear model and the general location model [2]. However for datasets containing different types of variables the current classes of joint models [3-5] may not be appropriate for the joint distribution of the data. The alternative method, chained equations imputation [4,6], is more flexible as it specifies a separate imputation model, typically a univariate regression model, for each incomplete variable and updates the missing data for each variable in turn.

Chained equations imputation has been proposed under several different names including: fully conditional specification, stochastic relaxation, variable-by-variable

*Correspondence: rachael.hughes@bristol.ac.uk
[1]School of Social and Community Medicine, University of Bristol, Bristol, UK
Full list of author information is available at the end of the article

imputation, regression switching, sequential regressions, ordered pseudo-Gibbs sampler, partially incompatible MCMC and iterated univariate imputation [3]. In addition to handling variables of varying types, the chained equations approach has other flexible features such as incorporating restrictions, logistical and consistency bounds (for example, to handle imputation of gender specific variables or impute only questions that were not intentionally skipped in a questionnaire [4]). van Buuren and Groothuis-Oudshoorn [7] discuss the wide range of medical fields that have used chained equations imputation (e.g. addiction [8], epidemiology [9], infectious diseases [10], genetics [11], cancer [12], obesity and physical activity [13]), and a brief review of available software that have implemented chained equations imputation is given by [14]. Given the popularity of chained equations, among users of varying degrees of expertise, there is now guidance in its practical use (e.g. [14] and [15]).

Despite its widespread use, a known theoretical weakness of the chained equations method is that the implicit joint distribution underlying the separate models may not always exist: that is, the conditional models may be incompatible [4,5,16-18]. In such situations, the results after chained equations imputation may systematically differ according to the order in which the missing variables are updated in the chained equations algorithm. We shall refer to this phenomenon as an "order effect".

Previous authors [3-5] have stated that chained equations imputation under a set of normal linear regression models, with all other variables as covariates and no interactions, is equivalent to a Gibbs sampler that draws from a multivariate normal distribution. van Buuren [3] also states for a dataset of three partially observed binary variables that chained equations under a set of logistic regression models, with all other variables included as main effects only, is equivalent to a joint modelling imputation under a log linear model with the three-way factor term set to zero. However, none of these papers provides a proof beyond stating that the set of conditional models is compatible [16] and are all derived from the specified joint distribution.

Independently and concurrently with our work, Liu et al. [19] have given sufficient conditions (which include compatibility of the conditionals) under which, as the sample size tends to infinity, the stationary distribution of the Markov chain generated by the chained equations algorithm (assuming that this stationary distribution exists and that the chain converges to it) converges to the posterior predictive distribution of the missing data implied by a joint Bayesian model. That is, under these sufficient conditions, the total variation of the distance between the chained equations stationary distribution and the posterior predictive distribution tends to zero as the sample size tends to infinity. As a corollary, Liu et al show the equivalence of the two imputation methods in finite samples under a condition we have independently identified and named the "non-informative margins" condition.

Our work is complementary to that of Liu et al. Firstly, we have taken a different approach to prove the equivalence of the two imputation methods in finite samples. Additionally, in specific examples, we prove whether the non-informative margins condition is satisfied or not, and in a simulation study we demonstrate the consequences when the conditional models are compatible but do not satisfy the non-informative margins condition.

In this paper, we provide a "non-informative margins" condition that, together with compatibility of the conditionals (and assuming that the Markov chain generated by the chained equations converges to a stationary distribution), guarantees that the imputed values obtained using chained equations (at convergence) are drawn from the posterior predictive distribution of the missing data implied by a Bayesian joint model. We give examples of chained equations algorithms that satisfy the non-informative margins condition when the joint model is the multivariate normal model and the saturated multinomial model, and examples where this condition is not satisfied when the joint model is an unsaturated multinomial model and the general location model. A simulation study considers a simple chained equations algorithm in which the conditional models are compatible but do not satisfy the non-informative margins condition, and shows that it is not equivalent to any joint model procedure.

Methods
Notation
Suppose K random variables $X = (X_1, \ldots, X_K)^\intercal$ are intended to be observed on N subjects. We use subscripts i and j to index subjects and variables respectively ($i = 1, \ldots, N$; $j = 1, \ldots, K$). Let $x = (x_{ij})$ denote an $(N \times K)$ matrix, whose i, j element is x_{ij}. Column j of matrix x is denoted by $x_j = (x_{1j}, \ldots, x_{Nj})^\intercal$. It is assumed that the rows of matrix x are independent and identically distributed draws from a probability distribution with probability distribution function $p(X \mid \theta)$, where θ is an unknown parameter.

In practice some subjects have missing observations on up to $K - 1$ variables and we write $x_j = (x_j^{obs}, x_j^{mis})$ for any j, $x = (x^{obs}, x^{mis})$ and $p(x \mid \theta) = p(x^{obs}, x^{mis} \mid \theta)$, with superscript *obs* and *mis* denoting the observed and missing data respectively. In keeping with the assumptions of joint modelling imputation and chained equations imputation, the missing data mechanism is assumed to be ignorable for Bayesian inference [20] p. 120, so that

inferences about θ can be based on the marginal observed data posterior $p(\theta \mid x^{obs})$.

Joint modelling imputation

Joint modelling imputation requires the specification of a parametric joint model $p(x^{obs}, x^{mis} \mid \theta)$ for the complete data and a prior distribution $p(\theta)$ for parameter θ. Imputations are independent draws from the posterior predictive distribution of the missing data given the observed data $p(x^{mis} \mid x^{obs})$ [2] p. 105, which under the ignorability assumption is

$$p\left(x^{mis} \mid x^{obs}\right) = \int p\left(x^{mis} \mid x^{obs}, \theta\right) p\left(\theta \mid x^{obs}\right) d\theta.$$

Therefore, to draw from this posterior predictive distribution, first draw $\theta^* \sim p(\theta \mid x^{obs})$ followed by $x^{mis*} \sim p(x^{mis} \mid x^{obs}, \theta^*)$ [2] p. 105. When it is difficult to draw from the observed data posterior $p(\theta \mid x^{obs})$, Markov chain Monte Carlo methods can be used. For example, the data augmentation algorithm of Tanner and Wong [21] draws missing values from the posterior predictive distribution $x^{mis*} \sim p(x^{mis} \mid x^{obs}, \theta^*)$ and then draws θ from the complete data posterior $\theta^* \sim p(\theta \mid x^{obs}, x^{mis*})$, where $*$ denotes the last drawn values of θ or x^{mis}. Upon convergence this produces a draw from the joint posterior distribution $p(\theta, x^{mis} \mid x^{obs})$.

Chained equations imputation

For every incomplete variable the chained equations algorithm requires an imputation model, typically a univariate regression model, and an accompanying prior distribution for the model's parameter. Let $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_K)^{\mathrm{T}}$ denote the vector of random variables excluding variable X_j and $x_{-j} = (x_{-j}^{obs}, x_{-j}^{mis})$ the submatrix of x corresponding to variables X_{-j}. We write $p(x_j \mid x_{-j}, \psi_j)$ for the probability distribution function of the imputation model for variable X_j and $p(\psi_j)$ for the prior distribution of the unknown parameter ψ_j.

Chained equations draws the imputations using an iterative algorithm, typically with 10 to 20 iterations [15]. To start off, the missing values of each incomplete variable are replaced by its mean or a random sample of its observed values. Suppose, without loss of generality, that variables $X_1, \ldots, X_R$ ($R \leq K$) are incomplete and variables $X_{R+1}, \ldots, X_K$ are fully observed. Given the imputations from the last iteration $(x_1^{(t-1)}, \ldots, x_R^{(t-1)})$, iteration t of the chained equations algorithm consists of the following draws [18]

$$\psi_1^{(t)} \sim p(\psi_1) p\left(x_1^{obs} \mid x_2^{(t-1)}, x_3^{(t-1)}, \ldots, x_R^{(t-1)}, x_{R+1}, \ldots, x_K, \psi_1\right)$$

$$x_1^{mis(t)} \sim p\left(x_1^{mis} \mid x_2^{(t-1)}, x_3^{(t-1)}, \ldots, x_R^{(t-1)}, x_{R+1}, \ldots, x_K, \psi_1^{(t)}\right)$$

$$\psi_2^{(t)} \sim p(\psi_2) p\left(x_2^{obs} \mid x_1^{(t)}, x_3^{(t-1)}, \ldots, x_R^{(t-1)}, x_{R+1}, \ldots, x_K, \psi_2\right)$$

$$x_2^{mis(t)} \sim p\left(x_2^{mis} \mid x_1^{(t)}, x_3^{(t-1)}, \ldots, x_R^{(t-1)}, x_{R+1}, \ldots, x_K, \psi_2^{(t)}\right)$$

$$\vdots$$

$$\psi_R^{(t)} \sim p(\psi_R) p\left(x_R^{obs} \mid x_1^{(t)}, x_2^{(t)}, \ldots, x_{R-1}^{(t)}, x_{R+1}, \ldots, x_K, \psi_R\right)$$

$$x_R^{mis(t)} \sim p\left(x_R^{mis} \mid x_1^{(t)}, x_2^{(t)}, \ldots, x_{R-1}^{(t)}, x_{R+1}, \ldots, x_K, \psi_R^{(t)}\right).$$

During each iteration the following two steps are applied to each incomplete variable X_j in turn: ψ_j^* is drawn from the posterior distribution proportional to $p(\psi_j) p(x_j^{obs} \mid x_{-j}^*, \psi_j)$ and missing values x_j^{mis*} are drawn from the predictive posterior $p(x_j^{mis} \mid x_{-j}^*, \psi_j^*)$. The imputations from the last iteration form the imputed dataset. The whole iterative algorithm is repeated to obtain further imputed datasets.

Equivalence of joint modelling and chained equations imputation

We investigated, in finite samples, sufficient conditions under which a chained equations algorithm with compatible conditional models imputes missing data from the predictive distribution of the missing data implied by the joint model and its accompanying prior. We provide examples of chained equations algorithms (with compatible conditional models) where our identified condition is satisfied and examples where it is not satisfied.

Simulation study

We conducted a simulation study to explore the consequences for chained equations imputation when the conditional models were compatible with the same joint model but the non-informative margins condition of Proposition 1 was not satisfied. In particular, we looked for evidence of "order effects", where the distribution from which the final imputed values of the variables were drawn differed according to the order in which the variables were updated in the chained equations sampler. If the chained equations algorithm imputes all variables from the predictive distribution of the missing data implied by a specific joint model, then order effects cannot occur [22]. Thus, the existence of order effects implies that the chained equations algorithm is not equivalent to imputing from any joint model.

The simulation study was based on a general location model, discussed in the Theoretical results section below, with one incomplete binary variable Y and two continuous variables W_1 and W_2, where W_1 was also incompletely observed. We compared joint modelling imputation under the general location model, considered as a gold standard,

with the chained equations algorithm that imputes the binary variable Y under a logistic regression model and the continuous variable W_1 under a normal linear regression model.

We generated 500 datasets, each with a sample size of 100. For each dataset, the rows were independent, identically distributed realizations of the general location model $Y \sim \text{Bernoulli}(3/10)$, $W_1 \mid Y \sim N(10 + \beta Y, 9)$ and $W_2 \mid W_1, Y \sim N(9 + 8/9 + 1/9 W_1 + \beta Y, 8 + 8/9)$. The data model was a simplified version of data that can occur in the medical literature [23]. The simulation study was repeated when β, the regression coefficient for covariate Y, was set to 1 and 3. The analysis of interest was the normal linear regression of W_2 on W_1 and Y. To ensure that any observed order effects could only be due to the failure of the non-informative margins condition we considered the simplest setting, that of data missing completely at random [20] p. 16, and set the values of Y and W_1 to be missing for the first 50 individuals in the dataset. Below we describe the joint modelling imputation procedure and the chained equations algorithm that were separately applied to the same 500 datasets.

We used the data augmentation algorithm (as described under the heading "Joint modelling imputation") to perform joint modelling imputation under the general location model and the joint prior given in the general location example (see example 4 of the Results), setting hyperparameters $\tau = \nu = 1/2$ and $\kappa = 3/2$. The number of imputed datasets generated, the burn-in period and the number of iterations between imputed datasets was 100. The analysis model was applied to each dataset separately and the mean of the multiple estimates of β, the coefficient for Y, was calculated.

In the (standard) chained equations algorithm, a logistic regression model for Y given W_1 and W_2 was first fitted to those rows of the dataset in which Y was observed. Let $\hat{\psi}_Y$ denote the maximum likelihood estimate of the parameters of this model and $\hat{V}$ denote its associated estimated variance-covariance matrix. A draw ψ_Y^* was then made from the multivariate normal approximation $N(\hat{\psi}_Y, \hat{V})$ and used to impute the missing Y values. The continuous variable W_1 was imputed using the linear regression model $W_1 \mid Y, W_2 \sim N(\lambda + \xi Y + \phi W_2, \omega)$ and prior distribution $p(\lambda, \xi, \phi, \omega) \propto \omega^{-3/2}$.

To start off the chained equations algorithm the missing values of Y and W_1 were replaced with a random sample of their observed values. We augmented the chained equations algorithm such that, within each iteration we fitted the analysis model immediately after updating the binary variable Y and also immediately after updating the continuous variable W_1. The simulation study focused on systematic differences between the two resulting estimates of β. Given the imputations from the last iteration $(y^{(t-1)}, w_1^{(t-1)})$, iteration t of the augmented chained equations algorithm consisted of the following steps:

1. Generate $y^{(t)}$ by imputing values for the missing binary observations, conditioning on $w_1^{(t-1)}$ and w_2.
2. Linearly regress w_2 on $w_1^{(t-1)}$ and $y^{(t)}$ and store the estimate for the coefficient of Y, denoted by $\hat{\beta}^b$.
3. Generate $w_1^{(t)}$ by imputing values for the missing continuous observations, conditioning on $y^{(t)}$ and w_2.
4. Linearly regress w_2 on $w_1^{(t)}$ and $y^{(t)}$ and store the estimate for the coefficient of Y, denoted by $\hat{\beta}^c$.

The chained equations algorithm was implemented with 10010 iterations. The first 10 iterations were regarded as burn-in and the estimates from these iterations discarded. The remaining 10000 estimates of $\hat{\beta}^b$ were averaged, and likewise for $\hat{\beta}^c$. We denote these means as $\bar{\beta}^b$ and $\bar{\beta}^c$ and their difference by $\bar{\beta}^b - \bar{\beta}^c$. The quantity $\bar{\beta}^b - \bar{\beta}^c$ can be interpreted as an estimate of the order effect for imputation in one dataset. We estimated the (Monte Carlo) standard error of $\bar{\beta}^b - \bar{\beta}^c$ using the batch-means method, a method for computing standard errors for correlated output [24] p. 124, and calculated a 95% confidence interval from this.

Linear discriminant analysis is an alternative way to estimate a logistic regression [25,26]. A modified chained equations algorithm using linear discriminant analysis on all individuals with observed Y has been proposed as an alternative way to impute the binary variable Y [27]. Because the linear discriminant likelihood is the joint distribution of Y, W_1 and W_2, this model has the advantage of recovering some information about ψ_Y in the W margin. We repeated the simulation study using this modified chained equations algorithm.

We assessed the sensitivity of our results by repeating the simulation study using different specifications. For joint modelling imputation we increased the number of imputed datasets generated, the burn-in period and the number of iterations between imputed datasets to 250. For the standard and modified chained equations procedures we (1) increased the burn-in period of the chained equations sampler to 1000 iterations and (2) sampled every 50th iteration thereby reducing serial correlation (with a burn-in period of 10 iterations). To check that our results were not dependent upon our choice of prior distributions we repeated the simulation study with improper imputation procedures; that is, using maximum likelihood estimates of ψ_j instead of Bayesian draws ψ_j from its posterior distribution $p(\psi_j)p(x_j^{obs}|x_{-j}^*, \psi_j)$. Lastly, we also repeated the simulation study with a sample size of 1000 observations.

Results

Theoretical results

Equivalence of joint modelling and chained equations imputation

In this section, we give our key result Proposition 1, which shows that, in finite samples, compatibility of the conditionals and our proposed non-informative margins condition are sufficient for chained equations and joint modelling to yield imputations from the same predictive distribution.

Consider a joint model $p(x \mid \theta)$ and prior $p(\theta)$. From here onwards we shall use $p(. \mid .)$ to refer to any probability distribution derived from this joint model. In particular, for each $j = 1, \dots, R$, $p(x_j \mid x_{-j}, \theta)$ is the conditional distribution of x_j given x_{-j} and θ implied by the joint model, and $p(x_{-j} \mid \theta)$ is the conditional distribution of x_{-j} given θ. The distribution of x given θ factorizes as

$$p(x \mid \theta) = p(x_j \mid x_{-j}, \theta)p(x_{-j} \mid \theta).$$

Let ψ_j and $\tilde{\psi}_j$ be functions of θ such that $p(x_j \mid x_{-j}, \theta) = p(x_j \mid x_{-j}, \psi_j)$ and $p(x_{-j} \mid \theta) = p(x_{-j} \mid \tilde{\psi}_j)$. That is, the distribution of x_j given x_{-j} and the distribution of x_{-j} depend on θ only through the functions ψ_j and $\tilde{\psi}_j$, respectively, of θ. Let $p(\psi_j, \tilde{\psi}_j)$ denote the joint prior for ψ_j and $\tilde{\psi}_j$ implied by $p(\theta)$, and let $p(\psi_j)$ and $p(\tilde{\psi}_j)$ denote the corresponding marginal priors.

The chained equations algorithm applies the following two steps for each x_j in turn:

Step CE1 Draw ψ_j^* from the distribution proportional to $p(\psi_j)p(x_j^{obs} \mid x_{-j}^*, \psi_j)$.
Step CE2 Draw x_j^{mis*} from $p(x_j^{mis} \mid x_{-j}^*, \psi_j^*)$.

The choice of the parameterizations ψ_j and $\tilde{\psi}_j$ does not affect the output of step CE2, but a parsimonious choice will help to make the condition of Proposition 1 hold.

Proposition 1. Upon convergence, the chained equations algorithm defined by CE1 and CE2 and the joint model $p(x \mid \theta)$ and prior $p(\theta)$ draw from the same predictive distribution of x^{mis} if, for each incomplete variable x_j, $p(\psi_j, \tilde{\psi}_j) = p(\psi_j)p(\tilde{\psi}_j)$, i.e. if the joint prior distribution for ψ_j and $\tilde{\psi}_j$ factorizes into independent priors for ψ_j and $\tilde{\psi}_j$.

Proof. Using the condition of Proposition 1,

$$p\left(\psi_j, \tilde{\psi}_j \mid x_j^{obs}, x_{-j}^*\right) \propto p\left(\psi_j, \tilde{\psi}_j\right) p\left(x_j^{obs}, x_{-j}^* \mid \psi_j, \tilde{\psi}_j\right)$$
$$= p(\psi_j)p(\tilde{\psi}_j)p\left(x_j^{obs} \mid x_{-j}^*, \psi_j\right) p\left(x_{-j}^* \mid \tilde{\psi}_j\right)$$

Now, integrating out $\tilde{\psi}_j$,

$$p\left(\psi_j \mid x_j^{obs}, x_{-j}^*\right) \propto p(\psi_j)p\left(x_j^{obs} \mid x_{-j}^*, \psi_j\right)$$

Therefore, step CE1 yields a draw from $p(\psi_j \mid x_j^{obs}, x_{-j}^*)$.

Next, x_j^{mis} and x_j^{obs} are conditionally independent given x_{-j}^* and ψ_j^* so $p(x_j^{mis} \mid x_{-j}^*, \psi_j^*) = p(x_j^{mis} \mid x_j^{obs}, x_{-j}^*, \psi_j^*)$ and step CE2 yields a draw from $p(x_j^{mis} \mid x_j^{obs}, x_{-j}^*, \psi_j^*)$.

Hence steps CE1 and CE2 together yield a draw from $p(\psi_j, x_j^{mis} \mid x_j^{obs}, x_{-j}^*)$, and in particular they draw x_j^{mis*} from $p(x_j^{mis} \mid x_j^{obs}, x_{-j}^*)$. The latter is a full-conditional distribution corresponding to the joint density $p(x)$ implied by the joint model. Once x_j^{mis*} has been sampled, ψ_j^*, the sampled value of ψ_j, is not used again in the chained equations algorithm. So, the application of steps CE1 and CE2 to each j in turn and then iterating is a Gibbs sampler which, at convergence, yields a draw from $p(x^{mis} \mid x^{obs})$, the predictive distribution implied by the joint model and its accompanying prior.

Comment 1. The condition of Proposition 1 does not hold if the conditional and marginal parameters ψ_j and $\tilde{\psi}_j$ are not distinct (i.e. if their joint parameter space is not the product of their separate parameter spaces), and in particular if the combined dimension of ψ_j and $\tilde{\psi}_j$ is greater than that of θ. Distinctness of parameter spaces is a property of the model $p(x \mid \theta)$ and not of the prior $p(\theta)$. It will be used in the examples of the unsaturated multinomial distribution and the general location model to identify joint models where the condition of Proposition 1 does not hold.

Comment 2. Heuristically, the condition of Proposition 1 says that there is no information about ψ_j in the marginal distribution of x_{-j}, so we call it the "non-informative margins" condition. When such information does exist, it is used by the joint modelling sampler but not by the chained equations algorithm, so they may draw from different distributions. Our simulation study will show that this occurs in an example by demonstrating order effects.

Comment 3. As the non-informative margins condition has only been shown to be sufficient, then potentially if this condition is not satisfied the chained equations algorithm defined by CE1 and CE2 and the joint model $p(x \mid \theta)$ and prior $p(\theta)$ could still draw from the same predictive distribution of x^{mis}.

Comment 4. This proof holds for improper prior distributions provided the posterior distributions are proper.

Comment 5. When the non-informative margins condition holds true the chained equations algorithm is a Gibbs sampler, and so order effects cannot occur [22].

Example 1: Multivariate normal

Consider the multivariate normal joint model $X \sim N(\mu, \Sigma)$ with parameter $\theta = (\mu, \Sigma)$ and prior distribution $p(\theta) \propto |\Sigma|^{-\kappa}$ ($\kappa \in \mathbb{Q}$). We show that the corresponding chained equations algorithm, which imputes under a set of normal linear regression models, satisfies the non-informative margins condition of Proposition 1 (and hence draws from the same joint model as joint modelling imputation).

For each j we partition the mean vector μ as $(\mu_j, \tilde{\mu}_j)^{\mathrm{T}}$ and the covariance matrix Σ as

$$\begin{bmatrix} \sigma_j & \varsigma_j^{\mathrm{T}} \\ \varsigma_j & \tilde{\Sigma}_j \end{bmatrix},$$

such that $X_j \sim N(\mu_j, \sigma_j)$ and $X_{-j} \sim N(\tilde{\mu}_j, \tilde{\Sigma}_j)$. The conditional distribution of X_j given X_{-j} is the normal linear regression model $X_j \mid X_{-j} \sim N(\alpha_j + \beta_j^{\mathrm{T}} X_{-j}, \omega_j)$ where $\beta_j^{\mathrm{T}} = \varsigma_j^{\mathrm{T}} \tilde{\Sigma}_j^{-1}$, $\alpha_j = \mu_j - \beta_j^{\mathrm{T}} \tilde{\mu}_j$ and $\omega_j = \sigma_j - \varsigma_j^{\mathrm{T}} \tilde{\Sigma}_j^{-1} \varsigma_j$ [2] p. 157. Using our notation for chained equations imputation (see under the subsection "Chained equations imputation" of the Methods section) $\psi_j = (\alpha_j, \beta_j, \omega_j)$ and $\tilde{\psi}_j = (\tilde{\mu}_j, \tilde{\Sigma}_j)$.

The joint prior for ψ_j and $\tilde{\psi}_j$ derived from $p(\theta)$ is $p(\psi_j, \tilde{\psi}_j) = |\Sigma|^{-\kappa} \times |\tilde{\Sigma}_j|$ [2] p. 158-159. So, using a standard result from matrix algebra for the determinant of a partitioned matrix,

$$p(\psi_j, \tilde{\psi}_j) = \left(\sigma_j - \varsigma_j^{\mathrm{T}} \tilde{\Sigma}_j^{-1} \varsigma_j \right)^{-\kappa} \times |\tilde{\Sigma}_j|^{-\kappa} \times |\tilde{\Sigma}_j|$$
$$= \omega_j^{-\kappa} \times |\tilde{\Sigma}_j|^{-(\kappa-1)}$$
$$= p(\psi_j) \times p(\tilde{\psi}_{(j)}).$$

Therefore, the non-informative margins condition of Proposition 1 is satisfied.

Example 2: Saturated multinomial distribution

We now show for joint modelling imputation under a saturated multinomial model and a Dirichlet prior for θ, that the corresponding chained equations algorithm satisfies the non-informative margins condition of Proposition 1.

Consider K categorical variables $X = (X_1, \ldots, X_K)$, where each X_j takes possible values $1, \ldots, I_j$ ($j = 1, \ldots, K$). Variables X define a K-way contingency table. Let $c = (c_1, \ldots, c_K)$ denote a generic cell of the contingency table, θ_c denote the cell probability $pr(X = c)$ and ∂ denote the set of all cells of the contingency table. The joint distribution of X is a multinomial distribution with parameter $\theta = (\theta_c : c \in \partial)$ and index equal to 1.

Summing the table counts over variable X_j produces a collapsed contingency table defined by variables X_{-j}. Let $c_{-j} = (c_1, \ldots, c_{j-1}, c_{j+1}, \ldots, c_K)$ denote a generic cell of the collapsed table, $\tilde{\psi}_{c_{-j}}$ denote the cell probability $pr(X_{-j} = c_{-j})$ and $\tilde{\partial}_j$ denote the set of all cells of the

collapsed table. The marginal distribution of X_{-j} is multinomial with parameter $\tilde{\psi}_j = \{\tilde{\psi}_{c_{-j}} : c_{-j} \in \tilde{\partial}_j\}$, where $\tilde{\psi}_{c_{-j}} = \sum_{c_j=1}^{I_j} \theta_c$. The conditional distribution of X_j given $X_{-j} = c_{-j}$ is the multinomial distribution with parameters $\psi_{c_{-j}} = \{pr(X_j = c_j \mid X_{-j} = c_{-j}) : c_j = 1, \ldots, I_j\}$. So, the full set of parameters for the conditional distribution of X_j given X_{-j} is $\psi_j = \{\psi_{c_{-j}} : c_{-j} \in \tilde{\partial}_j\}$.

If the prior distribution for θ is Dirichlet with hyperparameter $\alpha = \{\alpha_c : c \in \partial\}$, then the implied prior distributions for $\tilde{\psi}_j$ and ψ_j are independent: the prior for $\tilde{\psi}_j$ is Dirichlet with hyperparameter $\tilde{\alpha}_j = \{\tilde{\alpha}_{c_{-j}} : c_{-j} \in \tilde{\partial}_j\}$, where $\tilde{\alpha}_{c_{-j}} = \sum_{c_j=1}^{I_j} \alpha_c$, and the prior distribution for ψ_j is the product of the set of independent Dirichlet distributions $\{\psi_{c_{-j}} \sim D(\alpha_{c_{-j}}) : c_{-j} \in \tilde{\partial}_j\}$, where $\alpha_{c_{-j}} = \{\alpha_c : c_j = 1, \ldots, I_j\}$ is a subset of α [2] p. 256. Since the prior for $(\psi_j, \tilde{\psi}_j)$ can be factored into independent distributions for ψ_j and $\tilde{\psi}_j$, the non-informative margins condition of Proposition 1 is satisfied.

Example 3: Unsaturated multinomial distribution

When the joint model is an unsaturated multinomial model, we give an example where the conditional and marginal parameters ψ_j and $\tilde{\psi}_j$ are not distinct (see comment 1 of Proposition 1). Consequently the non-informative margins condition of Proposition 1 is not satisfied.

Consider K categorical variables $X = (X_1, \ldots, X_K)$ as described in the saturated multinomial example. Assume that all cell probabilities are positive, $\theta(c) > 0$ for all c, to ensure that every multinomial distribution considered can be written as a log linear model and that all possible conditional distributions exist [28] p. 202. Let the joint model be the hierarchical log linear model that contains all two-way factors between the K variables and no higher order factors. We shall refer to this as the all two-way factor hierarchical model. Under this model, for any j the conditional distribution of X_j given X_{-j} follows a multinomial logistic regression model (or a logistic regression model when X_j is binary) where the regression model includes variables X_{-j} as main effects only (i.e. no interaction terms).

In generating class notation [29] the all two-way factor hierarchical model is written as $[1 \ 2] \ldots [1 \ K] \ldots [(K-1) \ K]$. Any hierarchical log linear model for X can be represented by an undirected graph in which the graph vertices are the variables $X_1, \ldots, X_K$ and two vertices X_g and X_h are connected by an edge if and only if the log linear model contains two-factor or higher order terms involving variables X_g and X_h. Any model containing all two-factor terms is therefore represented by the complete graph. Asmussen and Edwards [29] state that two vertices in a graph are adjacent if there is an edge between them. For any subset a of the vertices, Asmussen and Edwards define the boundary of a to be the set of vertices that are

not in a but that are adjacent to one or more vertices in a. So, for the complete graph, the boundary of X_j is X_{-j}. Theorem 2.3 of [29] states that a hierarchical log linear model is collapsible onto X_{-j} if and only if the boundary of X_j is contained in a generator of the hierarchical log linear model. Further, Theorem 4.1 of [29] states that if a hierarchical log linear model is not collapsible onto X_{-j}, then parameters ψ_j and $\tilde{\psi}_j$ are not distinct. For any j, X_{-j}, the boundary of X_j in the complete graph, contains all of the remaining $K - 1$ variables. When $K \geq 4$, X_{-j} is not contained in any of the generators, $[1 \ 2] \dots [1 \ K] \dots [(K - 1) \ K]$, of the all two-way factor hierarchical model, and hence the log linear model is not collapsible onto X_{-j}. Therefore, parameters ψ_j and $\tilde{\psi}_j$ are not distinct, and so the non-informative margins condition of Proposition 1 is not satisfied when $K \geq 4$.

Example 4: General location model

We give an example of a chained equations algorithm derived from joint modelling under a general location model that does not satisfy the non-informative margins condition of Proposition 1. Our simulation study is based on this example.

Suppose that the data on each individual consist of one incomplete binary variable Y and $K - 1$ continuous variables $W = (W_1, \dots, W_{K-1})^{\mathrm{T}}$, where one or more of the continuous variables are also incomplete. Let the joint distribution of $X = (Y, W)^{\mathrm{T}}$ be the general location model $Y \sim Bernoulli(\gamma)$ and $W \mid Y \sim N(\mu_0 + \mu_1 Y, \Sigma)$, for unknown parameters $\theta = (\gamma, \mu_0, \mu_1, \Sigma)$ [30]. Let the joint prior for θ be $p(\theta) = \gamma^{\tau - 1}(1 - \gamma)^{\nu - 1} \mid \Sigma \mid^{-\kappa}$ with hyperparameters $\tau, \nu > 0$ and $\kappa \in \mathbb{Q}$.

From the multivariate normal example above it is straightforward to show that the non-informative margins condition of Proposition 1 holds for imputing any W_j.

The conditional distribution of Y given W, $p(Y \mid W, \psi_Y)$, is the logistic regression model with covariates W [25]. The marginal distribution of W, $p(W \mid \tilde{\psi}_Y)$, can be written as a mixture of normal distributions

$$
\begin{aligned}
p(W \mid \tilde{\psi}_Y) = {} & \gamma p(W \mid Y = 1, \mu_0, \mu_1, \Sigma) \\
& + (1 - \gamma) p(W \mid Y = 0, \mu_0, \Sigma).
\end{aligned}
$$

This cannot be parameterized more parsimoniously than $\tilde{\psi}_Y = (\gamma, \mu_0, \mu_1, \Sigma) = \theta$. As ψ_Y is a function of θ, it is determined by $\theta = \tilde{\psi}_Y$.

Consequently, given $\tilde{\psi}_Y$ the parameters of the logistic regression model ψ_Y are fully determined, and so $\tilde{\psi}_Y$ and ψ_Y are not distinct. Therefore, as discussed in comment 1 above, the non-informative margins condition of Proposition 1 does not hold.

Using the same argument, the non-informative margins condition of Proposition 1 does not hold for a chained equations algorithm derived from joint modelling under the restricted general location model, with cell probabilities restricted by the all two-factor hierarchical model (discussed in the unsaturated multinomial example) and cell means restricted to be a linear function of the categorical variables.

Simulation study results

This section reports the results of the simulation study, where the chained equations conditional models were compatible with the same joint model but the non-informative margins condition of Proposition 1 was not satisfied.

Figures 1a and 1b show, for the first 30 of the 500 datasets, the value of $\bar{\beta}^b - \bar{\beta}^c$ (estimate of the order effect in one dataset) along with its 95% confidence interval, for the (standard) chained equations procedure (i.e., binary variable Y imputed under the logistic regression model). In a number of the datasets the 95% confidence intervals did not cross zero. Thus, there was clear evidence of order effects, with the magnitude of such effects varying between datasets. Such statistically significant evidence of an order effect occurred in 164 and 386 of the 500 datasets for $\beta = 1$ and $\beta = 3$ respectively. The range of absolute values of $\bar{\beta}^b - \bar{\beta}^c$ was larger for $\beta = 3$ than for $\beta = 1$. These results confirm that the chained equations procedure was not equivalent to any joint model procedure.

The average magnitudes of the order effects $\bar{\beta}^b$, $\bar{\beta}^c$ and $|\bar{\beta}^b - \bar{\beta}^c|$ over the 500 datasets are shown in Table 1. Consider the results for the chained equations algorithm (labelled LR). In keeping with Figure 1, the average magnitude of the order effect was larger for $\beta = 3$ than $\beta = 1$. The means of $\bar{\beta}^b$ and $\bar{\beta}^c$ did not differ systematically, consistent with the direction of the order effect being arbitrary and dataset dependent. Estimates $\bar{\beta}^b$ and $\bar{\beta}^c$ appeared to be equally good estimates of β.

The forest plots of $\bar{\beta}^b - \bar{\beta}^c$ with 95% confidence intervals corresponding to the modified chained equations algorithm (i.e Y imputed under the linear discriminant model) are shown in Figures 1c and 1d. Order effects were smaller than in Figures 1a and b, but were still present because the modified algorithm did not use information about W_1 and W_2 when Y was missing. Out of the 500 datasets the number of datasets that showed statistically significant evidence of an order effect was 113 for $\beta = 1$ and 330 for $\beta = 3$.

From Table 1, the complete case estimates of β were unbiased. When $\beta = 3$ the linear discriminant analysis values $\bar{\beta}^c$ and $\bar{\beta}^b$, and the joint modelling imputation estimate $\hat{\beta}_{JM}$ were slightly biased towards the null. This bias was due to the prior (which was the same for imputation under the linear discriminant analysis model and joint modelling imputation under the general location model) and it disappeared when the sample size was increased to 1000 observations and when the simulation study was
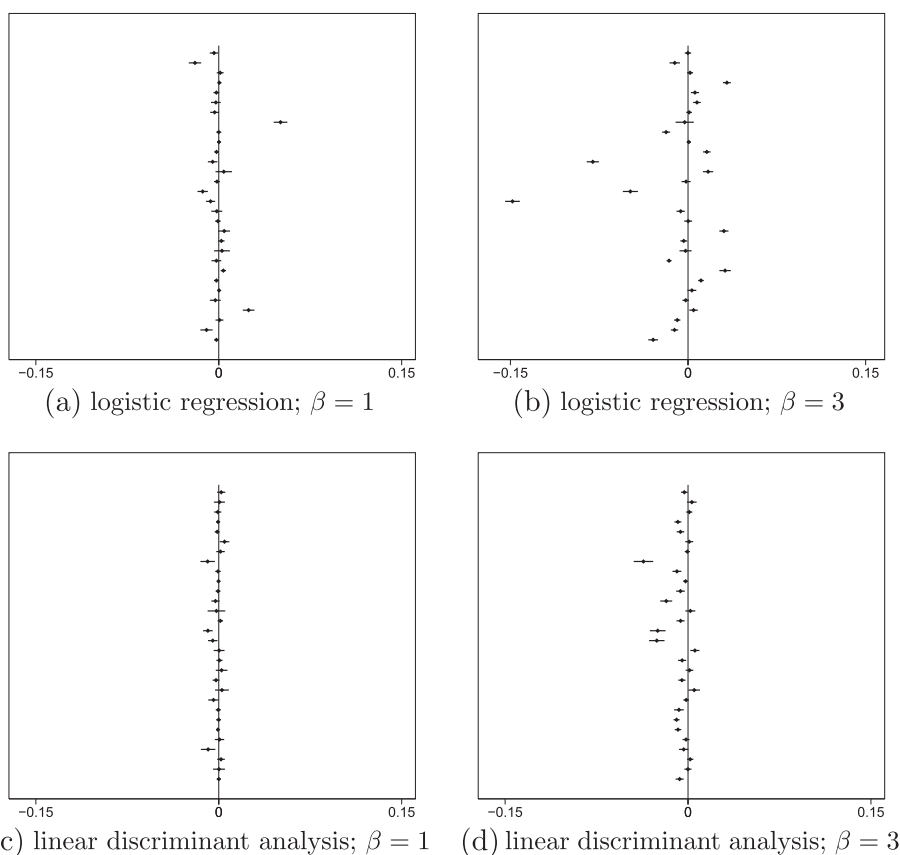
(a) logistic regression; $\beta = 1$

(b) logistic regression; $\beta = 3$

(c) linear discriminant analysis; $\beta = 1$

(d) linear discriminant analysis; $\beta = 3$

Figure 1 Four forest plots of the posterior mean differences $\bar{\beta}^b - \bar{\beta}^c$. Each panel is a forest plot of $\bar{\beta}^b - \bar{\beta}^c$ for the first 30 datasets, 95% confidence intervals calculated using the Monte Carlo standard error. Panels **(a)** and **(b)** correspond to binary variable Y imputed under the logistic regression model, with $\beta = 1$ and $\beta = 3$ respectively. Panels **(c)** and **(d)** correspond to Y imputed under the linear discriminant model, with $\beta = 1$ and $\beta = 3$ respectively.

Table 1 Over 500 datasets, average of the complete case estimates, joint modelling imputation estimates and values of $\bar{\beta}^b$, $\bar{\beta}^c$ and $|\bar{\beta}^b - \bar{\beta}^c|$ for the chained equations algorithm and the modified chained equations algorithm, with confidence intervals [mean $\pm$ 1.96 $\times$ (standard deviation $\div \sqrt{500}$)]

β	$\hat{\beta}_{CCA}{}^{\flat}$	$\hat{\beta}_{JM}{}^{\sharp}$		$\bar{\beta}^b$	$\bar{\beta}^c$	$|\bar{\beta}^b - \bar{\beta}^c|$
1	$1 \cdot 01$	$0 \cdot 97$	LR$^{\nabla}$	$0 \cdot 98$	$0 \cdot 98$	$0 \cdot 0040$
	$[0 \cdot 93, 1 \cdot 09]$	$[0 \cdot 89, 1 \cdot 05]$		$[0 \cdot 90, 1 \cdot 06]$	$[0 \cdot 90, 1 \cdot 06]$	$[0 \cdot 0034, 0 \cdot 0046]$
			LDA¶	$0 \cdot 97$	$0 \cdot 97$	$0 \cdot 0022$
				$[0 \cdot 89, 1 \cdot 05]$	$[0 \cdot 89, 1 \cdot 05]$	$[0 \cdot 0020, 0 \cdot 0024]$
3	$3 \cdot 00$	$2 \cdot 90$	LR	$2 \cdot 93$	$2 \cdot 93$	$0 \cdot 0166$
	$[2 \cdot 91, 3 \cdot 09]$	$[2 \cdot 82, 2 \cdot 98]$		$[2 \cdot 85, 3 \cdot 02]$	$[2 \cdot 85, 3 \cdot 02]$	$[0 \cdot 0145, 0 \cdot 0188]$
			LDA	$2 \cdot 89$	$2 \cdot 89$	$0 \cdot 0067$
				$[2 \cdot 81, 2 \cdot 98]$	$[2 \cdot 81, 2 \cdot 97]$	$[0 \cdot 0062, 0 \cdot 0073]$

$\flat$ complete case analysis estimate of β from the observed data only;
$\sharp$ estimate of β from joint modelling imputation;
∇ logistic regression;
$\P$ linear discriminant analysis.

conducted using improper imputation; i.e using maximum likelihood estimates instead of Bayesian draws of parameters. For joint modelling imputation, and the standard and modified chained equations procedures changing their specifications (e.g., larger number of iterations, burn-in period) gave the same pattern of results as above (results not shown but available on request from the authors).

In preliminary simulation studies, when the non-informative margins condition was satisfied the results were consistent with a zero order effect (results not shown but available on request from the authors).

Discussion

We have defined a non-informative margins condition which, together with compatibility of the conditional models, we have proved is sufficient for a chained equations algorithm to impute missing data from the predictive distribution of the missing data implied by the joint model and its prior distribution. Also, we have shown that compatibility of the conditional models is not alone a sufficient condition. In a scenario where the conditionals models were compatible but the non-informative margins condition failed, our simulation study showed that the distribution from which the final imputed values of the variables were drawn differed, in a dataset-dependent manner, according to the order in which the variables were updated in the chained equations sampler.

In work that is complementary to the finite-sample results presented in this paper, Liu et al. [19] identified sufficient conditions for chained equations imputation and imputation under a fully Bayesian model to be asymptotically equivalent; that is, for the supremum of the difference between the two imputation distributions to converge to zero as the sample size tends to infinity. This implies that when the non-informative margins condition is not satisfied but the conditional models are compatible with the same joint model, the order effects identified in our simulation study will disappear as the sample size tends to infinity.

In our simulation study the average magnitude of the order effects was small and did not induce bias. Given that chained equations imputation is a widely used approach to imputation, these results are somewhat reassuring. However, the scope of these simulations was limited and it remains possible that chained equations imputation could lead to more substantial bias in different situations; for example, when there are many partially observed variables.

When the non-informative margins condition does not hold we expect some loss of efficiency in general (because some information is discarded). However, in our simulation study we did not detect any sizable loss of efficiency.

The issue of variance estimation for chained equations imputation is beyond the scope of this paper.

The advantage of chained equations imputation is that we do not need to specify the joint distribution of the variables. In cases where it is not known that there is a joint distribution, several methods for checking compatibility have been proposed (e.g., [16,31-35]). In practice, these methods are either limited to discrete distributions or are difficult to apply for multivariate distributions of more than 2 or 3 dimensions. This means that it may not be possible to check that the conditionals are compatible with the same joint model or that our non-informative margins condition holds true. van Buuren and other authors [3,18], in the examples they considered, concluded that chained equations is a robust approach even when the set of conditionals are not compatible with the same joint model. The findings of our simulation study support this body of work. Other studies [3,4,36,37] have compared chained equations and joint modelling, when missingness is multivariate, nonmonotone and ignorable, in settings which reflect real data (e.g. mixture of different types of variables, non-linear relationships and interactions between variables, semi-continuous variables). None of these studies has reported substantial differences in the performances of joint modelling imputation and chained equations imputation. Nonetheless many authors emphasize the need for further understanding of the theoretical underpinnings of the chained equations approach and the establishment of the robustness of the chained equations method (e.g. [3,7,14,38]).

Conclusions

In finite samples, compatibility of the conditionals and our non-informative margins condition are sufficient for chained equations and joint modelling to yield imputations from the same predictive distribution. Furthermore, our simulation study demonstrated that, even in a simple setting, a chained equations procedure that does not satisfy the non-informative margins condition is not necessarily equivalent to a joint model procedure, even though when its conditional models are compatible.

When conditionals are incompatible or the non-informative margins condition is not satisfied, the distribution from which the imputed values are drawn can differ according to the order in which the variables are updated in the chained equations sampler, thereby introducing order effects.

Given the widespread use of chained equations imputation in medical research for datasets with different types of variables, researchers must be aware that order effects are likely to be ubiquitous. As noted by van Buuren [3], further work is needed to verify the robustness of chained equations to incompatibility of the conditional models in more general and realistic settings. Equally, future work

could evaluate the robustness of chained equations imputation when the sample size is small-to-moderate, the conditionals are compatible and the non-informative margins condition is not satisfied.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

IRW and JRC proposed the project. All authors contributed to the examples and/or the simulation study. RAH carried out the programming for the simulation study. RAH drafted the manuscript, and RAH, IRW and SRS edited the manuscript. All authors read and approved the final manuscript.

Author details

[1]School of Social and Community Medicine, University of Bristol, Bristol, UK. [2]MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK. [3]London School of Hygiene and Tropical Medicine, London, UK. [4]MRC Clinical Trials Unit, London, UK.

References

1. Rubin DB: *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc; 1987.
2. Schafer JL: *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall; 1997.
3. van Buuren S: **Multiple imputation of discrete and continuous data by fully conditional specification.** *Stat Methods Med Res* 2007, **16**:219–242.
4. Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P: **A multivariate technique for multiply imputing missing values using a sequence of regression models.** *Surv Methodol* 2001, **27**:85–95.
5. Gelman A, Raghunathan TE: **[Conditionally specified distributions: An introduction]: comment.** *Stat Sci* 2001, **16**:268–269.
6. van Buuren S, Boshuizen HC, Knook DL: **Multiple imputation of missing blood pressure covariates in survival analysis.** *Statist Med* 1999, **18**:681–694.
7. van Buuren S, Groothuis-Oudshoorn K: **mice: Multivariate Imputation by Chained Equations in R.** *J Stat Softw* 2011, **45**:1–67.
8. Morgenstern M, Wiborg G, Isensee B, Hanewinkel R: **School-based alcohol education: Results of a cluster-randomized controlled trial.** *Addiction* 2009, **104**:402–412.
9. Mueller B, Cummings P, Rivara F, Brooks M, Terasaki R: **Injuries of the head, face, and neck in relation to ski helmet use.** *Epidemiology* 2008, **19**:270–276.
10. Nash D, Katyal M, Brinkhof M, Keiser O, May M, Hughes R, Dabis F, Wood R, Sprinz E, Schechter M, Egger M: **Long-term immunologic response to antiretroviral therapy in low-income countries: A collaborative analysis of prospective studies.** *AIDS* 2008, **22**:2291–2302.
11. Souverein O, Zwinderman A, Tanck T: **Multiple imputation of missing genotype data for unrelated individuals.** *Ann Hum Genet* 2006, **70**:372–381.
12. Huo D, Adebamowo C, Ogundiran T, Akang E, Campbell O, Adenipekun A, Cummings S, Fackenthal J, Ademuyiwa F, Ahsan H, Olopade O: **Parity and breastfeeding are protective against breast cancer in nigerian women.** *Br J Cancer* 2008, **98**:992–996.
13. Wiles N, Jones G, Haase A, Lawlor D, Macfarlane G, Lewis G: **Physical activity and emotional problems amongst adolescents.** *Soc Psychiatry Psychiatric Epidemiol* 2008, **43**:765–772.
14. Azur M, Stuart E, Frangakis C, Leaf P: **Multiple imputation by chained equations: what is it and how does it work?** *Int J Methods Psychiatric Res* 2011, **20**:40–49.
15. White IR, Royston P, Wood A: **Multiple imputation using chained equations: Issues and guidance for practice.** *Stat Med* 2011, **30**:377–399.
16. Arnold BC, Press SJ: **Compatible conditional distributions.** *J Am Statist Assoc* 1989, **84**:152–156.
17. Heckerman D, Chickering DM, Meek C, Rounthwaite R, Kadie C: **Dependency networks for inference, collaborative filtering, and data visualization.** *J Mach Learn Res* 2000, **1**:49–75.
18. van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB: **Fully conditional specification in multivariate imputation.** *J Stat Comput Simulat* 2006, **76**:1049–1064.
19. Liu J, Gelman A, Hill J, Su Y, Kropko J: **On the stationary distribution of iterative imputations.** *Biometrika* 2013. doi: 10.1093/biomet/ast044.
20. Little RJA, Rubin DB: *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc; 2002.
21. Tanner MA, Wong WH: **The calculation of posterior distributions by data augmentation.** *J Am Statist Assoc* 1987, **82**:528–540.
22. Arnold B, Castillo E, Sarabia J: **Conditionally specified distributions an introduction.** *Stat Sci* 2001, **16**:249–265.
23. Kirkwood B, Sterne J: *Essential Medical Statistics*. Hoboken, New Jersey, US: Wiley-Blackwell; 2003.
24. Albert J: *Bayesian computation with R*. Dordrecht, Heidelberg, London, New York: Springer; 2009.
25. Efron B: **The efficiency of logistic regression compared to normal discriminant analysis.** *J Am Statist Assoc* 1975, **70**:892–898.
26. Cox D, Snell E: *Analysis of Binary Data*. second edition. London, UK: Chapman and Hall; 1989.
27. van Buuren S, Oudshoorn C: *Multivariate Imputation by Chained Equations (Mice V1.0 User's Manual)* ; 2000. http://www.stefvanbuuren.nl/publications/MICE%20V1.0%20Manual%20TNO00038%202000.pdf.
28. Whittaker J: *Graphical models in applied multivariate statistics*. New York: John Wiley & Sons, Inc; 1990.
29. Asmussen S, Edwards D: **Collapsibility and response variables in contingency tables.** *Biometrika* 1983, **70**:567–578.
30. Olkin I, Tate RF: **Multivariate correlation models with mixed discrete and continuous variables.** *Ann Math Stat* 1961, **32**:448–465.
31. Arnold B, Castillo E, Sarabia J: **Compatibility of partial or complete conditional probability specifications.** *J Stat Plann Inference* 2004, **123**:133–159.
32. Ip E, Wang Y: **Canonical representation of conditionally specified multivariate discrete distributions.** *J Multivariate Anal* 2009, **100**:1282–1290.
33. Tian G, Tan M, Ng K, Tang M: **A unified method for checking compatibility and uniqueness for discrete conditional distributions.** *Commun Stat: Theory Methods* 2009, **38**:115–129.
34. Chen H: **Compatibility of conditionally specified models.** *Stat Probability Lett* 2010, **80**:670–677.
35. Kuo K, Wang Y: **A simple algorithm for checking compatibility among discrete distributions.** *Comput Stat Data Anal* 2011, **55**:2457–2462.
36. Horton N, Kleinman K: **Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models.** *Am Stat* 2007, **61**:79–90.
37. Yu LM, Burton A, Rivero-Arias O: **Evaluation of software for multiple imputation of semi-continuous data.** *Stat Methods Med Res* 2007, **16**:243–258.
38. Kenward M, Carpenter J: **Multiple imputation: current perspectives.** *Stat Methods Med Res* 2007, **16**:199–218.