Statistical Methods in Medical Research 2007; **16**: 219–242

Multiple imputation of discrete and continuous data by fully conditional specification

Stef van Buuren TNO Quality of Life, Leiden, The Netherlands and University of Utrecht, The Netherlands

The goal of multiple imputation is to provide valid inferences for statistical estimates from incomplete data. To achieve that goal, imputed values should preserve the structure in the data, as well as the uncertainty about this structure, and include any knowledge about the process that generated the missing data. Two approaches for imputing multivariate data exist: joint modeling (JM) and fully conditional specification (FCS). JM is based on parametric statistical theory, and leads to imputation procedures whose statistical properties are known. JM is theoretically sound, but the joint model may lack flexibility needed to represent typical data features, potentially leading to bias. FCS is a semi-parametric and flexible alternative that specifies the multivariate model by a series of conditional models, one for each incomplete variable. FCS provides tremendous flexibility and is easy to apply, but its statistical properties are difficult to establish. Simulation work shows that FCS behaves very well in the cases studied. The present paper reviews and compares the approaches. JM and FCS were applied to pubertal development data of 3801 Dutch girls that had missing data on menarche (two categories), breast development (five categories) and pubic hair development (six stages). Imputations for these data were created under two models: a multivariate normal model with rounding and a conditionally specified discrete model. The JM approach introduced biases in the reference curves, whereas FCS did not. The paper concludes that FCS is a useful and easily applied flexible alternative to JM when no convenient and realistic joint distribution can be specified.

1 Introduction

Multiple imputation (MI) is a general statistical method for the analysis of incomplete data sets.[1,2] A statistical analysis using multiple imputation typically comprises three major steps. The first step involves specifying and generating plausible synthetic data values, called imputations, for the missing values in the data. This step results in a number of complete data sets (m) in which the missing data are replaced by random draws from a distribution of plausible values. The number of imputations, m, typically varies between 3 and 10. The second step consists of analyzing each imputed data set by a statistical method that will estimate the quantities of scientific interest. This step results in m analyses (instead of one), which will differ only because the imputations differ. The third step pools the m estimates into one estimate, thereby combining the variation within and across the m imputed data sets. Under fairly liberal conditions, this step results in statistically valid estimates that translate the uncertainty caused by the missing data into the width of the confidence interval.

Address for correspondence: Stef van Buuren, TNO Quality of Life, Leiden, The Netherlands. E-mail: stef.vanbuuren@tno.nl

10.1177/0962280206074463

MI is a highly modular statistical method in the sense that the steps can be executed separately, and with relatively limited interaction between the steps. The major rule that connects steps 1 and 2 is that every relation to be studied in the step 2 should, in some way, be included into the specification of the plausible values for the missing data in step 1. Failure to do so may bias the estimates towards the null, the amount of which depends on the amount of missing data and the strength of the relationship of interest. It should be pointed out however that for such failures to occur the relations have to be quite strong and the amount of missing information has to be quite high.[3]

Rubin formulated the main principles of MI already at the end of the 70 s,[4] but the uptake of the technique has been rather slow. The number of applications of MI in health is currently growing at a fair rate. Figure 1 plots the number of citations per years of Rubin's[1] book in medical journals and in all journals (source: www.scopus.com). About half of all applications of MI occur in the medical field. There is a steady rise in the number of citations. Of course, we have to take into account that the citation database has more coverage in the recent years. For comparison, we included the number of references in medical journals to the classic EM paper by Dempster *et al.*[5] Relative to that work, the number of applications of MI is growing.

We refer to Little and Rubin[6] for a discussion of the relative merits of approaches to missing data other than MI, for example ad hoc methods, direct maximum likelihood and weighting. Schafer's[7] book is the standard work of imputation for multivariate data. Introductions into MI have been written by Schafer, Stern *et al.* and Allison.[8–10] The overview by Schafer and Graham[11] addresses many practical points relevant to the

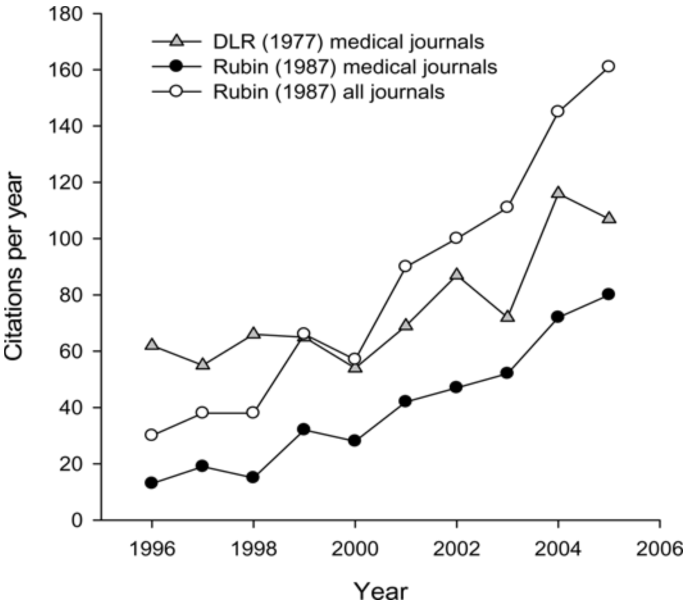


Figure 1 Number of citations per year in medical journals of the EM-algorithm[5] and multiple imputation.[1] Source: www.scopus.com, assessed May 8, 2006.

application of MI. Overviews of MI in health have been written by Rubin and Schenker and Barnard and Meng.[12,13] Evaluations of MI and comparative reviews have appeared in various medical fields: epidemiology,[14–16] psychiatric and developmental research,[17] nursing research,[18–21] public health,[22–24] cost and outcomes research,[25–27] quality of life,(28) and physical activity,[29,30] educational research,[31,32] and chemometrics.[33] More methodologically oriented comparative reviews have appeared on multilevel models,[34] structural equation modeling,[35,36] methods for longitudinal data,[37,38] attrition problems in longitudinal data,[39,40] drop out in clinical trials,[41–46] and meta-analysis.[47] Ibrahim *et al.*[48] provide a comparative review of various advanced missing data methods. Schafer[49] compares Bayesian MI methods with direct maximum likelihood methods. Taken together, these references provide abundant evidence on the value and vitality of MI in health research.

The present paper deals with the question how to create multiple imputations for multivariate data. The paper provides an overview of methods for generating multiple imputations, starting from basic methods where the missing values are confined to one variable, and continuing to more advanced methods for dealing with general patterns of missing values in multivariate data of various types, including mixes of categorical and continuous data. We distinguish between approaches based on both joint modeling (JM) and fully conditional specification (FCS). An application on pubertal data from the Fourth Dutch Growth Study illustrates the principles.

2 Method

2.1 Notation

Let Y_j be one of k incomplete random variables ($j = 1, \ldots, k$) and let $Y = (Y_1, \ldots, Y_k)$. The observed and missing parts of Y_j are denoted by Y_j^{obs} and Y_j^{mis}, respectively, so $Y^{\text{obs}} = (Y_1^{\text{obs}}, \ldots, Y_k^{\text{obs}})$ and $Y^{\text{mis}} = (Y_1^{\text{mis}}, \ldots, Y_k^{\text{mis}})$ stand for the observed and missing data in Y. Let $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_k)$ denote the collection of the $k-1$ variables in Y except Y_j. Let R_j be the response indicator of Y_j, with $R_j = 1$ if Y_j is observed and $R_j = 0$ if Y_j is missing. Let $R = (R_1, \ldots, R_k)$ and $R_{-j} = (R_1, \ldots, R_{j-1}, R_{j+1}, \ldots, R_k)$. Let $X = (X_1, \ldots, X_l)$ be a set of l complete covariates on the same subjects. In order to avoid distracting complexities, we assume that the observations in Y, X and R correspond to a simple random sample from the population of interest.

2.2 Imputation models

Rubin[1] (Ch. 5) distinguished three tasks for creating imputations under an explicit model: the *modeling* task the *imputation* task and the *estimation* task. The modeling task is to provide a specification for the hypothetical joint distribution $P(Y, X, R)$ of all data. The imputation task sets out to derive the posterior predictive distribution $P(Y^{\text{mis}}|Y^{\text{obs}}, X, R)$ of the missing values Y^{mis} given the observed data. The estimating task consists of calculating the posterior distribution of the parameters of this distribution, so that random draws can be made from it. According to Rubin's framework, the imputations follow from the specification of the joint model $P(Y, X, R)$.

In practice, it is often difficult to specify a realistic joint model $P(Y, X, R)$. Model $P(Y, X, R)$ embraces both the model for generating the imputations and the scientifically interesting model for which the data were sampled in the first place. This dual role of $P(Y, X, R)$ puts a heavy burden on its specification. Several classes of joint models have been proposed. Schafer developed joint models (JM) for imputation under the multivariate normal (MVN), the log-linear and the general location model.[7] The methods are theoretically elegant, but they often lack flexibility to account for important features of the data. For example, if the data contain derived variables (e.g., sum scores, transformations, indices) one would like the imputation procedure to ensure consistency between the constituent parts. Multivariate imputation according to a joint model could also create impossible combinations like 'pregnant fathers', which are better avoided in the imputed data. The rows or columns could have a meaningful order, for example, as in longitudinal data. Real data often consist of a mix of different scale types (e.g., binary, unordered, ordered, continuous). Also, the relation between Y_j and predictors Y_{-j} can be complex, for example, nonlinear or be subject to censoring or rounding or contain interactions that are important. Enforcing parametric joint models $P(Y, X, R)$ on the data potentially discards interesting features in the data that we may wish to investigate and may thus severely limit the class of scientific models that may be legitimately applied to the imputed data.

Fortunately, imputations of high quality can be generated without an explicit specification of $P(Y, X, R)$. An *imputation model* $P(Y^{\text{mis}}|X, Y^{obs}, R)$ describes how synthetic values for $Y^{\text{mis}} = (Y_1^{\text{mis}}, \ldots, Y_k^{\text{mis}})$ are generated. The imputation model can be an explicit probability model, or an implicit model, like hot-deck (Little and Rubin, 2002, p. 67). In principle, the imputation model can correspond to any method to augment the data, as long as it yields imputations that are *proper* in the sense of Rubin (1987, p. 119). A procedure is proper if particular conditions hold for the complete-data statistics and the within and between imputation variances in the case $m = \infty$. An important requirement for a procedure to be proper is that the variability of the parameters of the imputation model should be included into the generated imputations, a property that Schafer[7] calls 'Bayesianly proper'. It is actually difficult to demonstrate properness analytically in a given case (Schafer, 1997, p. 145). Brand *et al.*[50] for a validation strategy based on simulation to assess various aspects of properness. Note that the imputation model $P(Y^{\text{mis}}|X, Y^{\text{obs}}, R)$ need not make an explicit reference to a specification for $P(Y, X, R)$ and that it does not automatically follow from the joint distribution $P(Y, X, R)$. Imputation models bypass the need to specify $P(Y, X, R)$, though their use creates new responsibilities for substantiating its correctness for a given statistical analysis. Instead of specifying $P(Y, X, R)$, using models $P(Y^{\text{mis}}|X, Y^{\text{obs}}, R)$ is a separate modeling activity that comes with its own goals and rules.[3,49,51–53]

This paper is based on the idea that we may bypass the (joint) modeling task, and directly specify a sensible model for creating multivariate imputations $P(Y^{\text{mis}}|X, Y^{\text{obs}}, R)$. A convenient way of doing that is to generate imputations in multivariate data variable-by-variable by specifying a conditional model $P(Y_j^{\text{mis}}|X, Y_{-j}, R)$ for each $Y_j, j = 1, \ldots, k$.

2.3 Ignorability

Let us first look at the role of R within the imputation model. The imputation model for variable j, $P(Y_j^{\text{mis}}|X, Y_{-j}, R)$, exploits relations between and within Y, X and R. Let us for the moment assume that $k = 1$, so that there is only one Y with missing data. In that case, the information about Y that is present in X and R is summarized by the conditional distribution $P(Y|X, R)$. Cases with missing Y, that is, with $R = 0$, do not provide any information about $P(Y|X, R)$, and so in actual data analysis it is only possible to fit models for $P(Y|X, R = 1)$. It is, however, the distribution $P(Y|X, R = 0)$ that we need to draw imputations from, and the central problem is how to specify that distribution. The conventional procedure is to equate $P(Y|X, R = 0) = P(Y|X, R = 1)$, which corresponds to the assumption that the response mechanism is ignorable (Rubin, 1987, pp. 51–53).

The assumption of ignorability is often sensible in practice and generally provides a natural starting point. If, on the other hand, the assumption is not reasonable (e.g., when data are censored), we may use other forms for $P(Y|X, R = 0)$. The fact that $R = 0$ allows for the possibility that $P(Y|X, R = 1) \neq P(Y|X, R = 0)$ (Rubin, 1987, p. 205). By definition, the specification of $P(Y|X, R = 0)$ needs assumptions external to the data. As long as the imputations reflect the correct amount of uncertainty about the values that are missing, there is nothing in the theory of MI that prevents appropriate inferences under $P(Y|X, R = 0)$. MI will also work for nonignorable response mechanisms.

Example: Suppose that a growth study measures body weight in kg (Y) and gender (X_1: $1 =$ boy, $0 =$ girl) of 15-year old children, and that some of the body weights are missing. We can model the weight distribution for boys and girls separately for those with observed weights, that is, $P(Y|X_1 = 1, R = 1)$ and $P(Y|X_1 = 0, R = 1)$. If we assume that the response mechanism is ignorable, then imputations for a boy's weight can be drawn from $P(Y|X_1 = 1, R = 0) = P(Y|X_1 = 1, R = 1)$. The same can be done for girls. This procedure leads to correct inferences on the combined sample of boys and girls, even if boys have substantially more missing values, or if the body weights of the boys and girls are very different. The procedure is however not appropriate if, within the group of boys or the girls, the occurrence of the missing data is related to body weight. For example, some of the heavier children may not want to be weighed, resulting in more missing values for the more obese. It will be clear that assuming $P(Y|X_1, R = 0) = P(Y|X_1, R = 1)$ will then underestimate the prevalence of overweight and obesity. In this case, it may be more realistic to specify $P(Y|X_1, R = 0)$ such that imputation accounts for the excess body weights in the children that were not weighed. There are many ways to do that. In all these cases the response mechanism will be nonignorable.

The assumption of ignorability is essentially the belief on the part of the user that the available data are sufficient to correct for the effects of the missing data. The assumption cannot be tested on the data itself, but it can be checked against suitable external validation data. There are two main strategies that we may pursue if the response mechanism is not ignorable. The first is to expand the data and assume ignorability on the expanded data. In the above example, fat children may simply not want anybody to know their weight, but perhaps have no objection if their waist circumference (X_2) is measured. The ignorability assumption $P(Y|X, R = 0) = P(Y|X, R = 1)$ is less stringent for $X = (X_1, X_2)$ than for $X = (X_1)$, and hence more realistic. The second strategy is to formulate $P(Y|X, R = 0)$ different from $P(Y|X, R = 1)$, describing which body weights would have been observed if they had been measured. Candidates for such models include the pattern mixture model and the selection model, though application of such models requires untestable *a priori* assumptions beyond the data (Little and Rubin, 2002, Ch. 15; Schafer, 1997, p. 28).[6,7]

We may disregard R in the imputation model if we are prepared to make the assumption of ignorability. If this is not realistic, then we can pursue one of the two strategies outlined above. Of course, any such methods need to be explained and justified as part of the statistical analysis.

3 Univariate and monotone imputation

For both theoretical and practical reasons, it is useful to distinguish between monotone and nonmonotone missing data patterns. A pattern is monotone if the variables can be ordered such that, for each person, all earlier variables are observed if the later variable is observed. Monotone patterns often occur as a result of dropout in a longitudinal study. It is often useful to sort variables and cases to approach a monotone pattern. Figure 2 depicts various monotone and nonmonotone missing data patterns.

3.1 Univariate methods

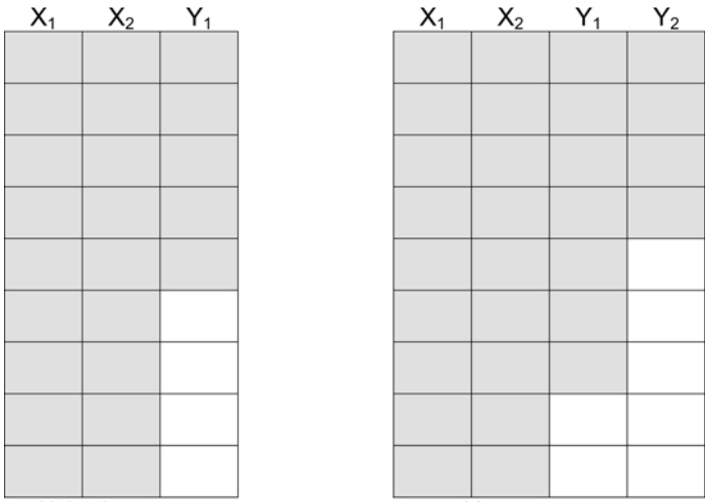
An important special case of a monotone missing data pattern occurs when $k = 1$. In that case, there is only one Y that needs to be imputed and the remaining data X are all complete. Table 1 contains an overview of various methods that have been proposed for generating multiple imputations for univariate data. Many methods are variations on the linear regression method proposed by Rubin (1987, p. 166).[1]

3.2 Monotone patterns

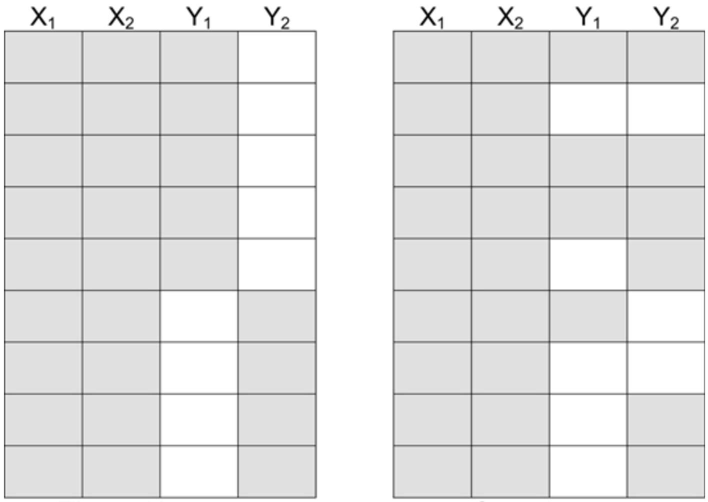
Imputations for multivariate missing data can be imputed by a sequence of univariate methods if the missing data pattern is *monotone-distinct*.[1] Suppose that variables $Y_1, \ldots, Y_k$ are ordered in a monotone pattern such that for $j = 1, \ldots, k-1$, all cases with missing data in Y_j also have missing data in $Y_{>j}$. If, in addition, the parameters $\phi_1, \ldots, \phi_k$ of the imputation models are *a priori* independent, that is, if they factor into independent marginal priors, we can draw a set of multivariate imputations using the following sequence of univariate imputation models

$$
\begin{aligned}
&P(Y_1^{\text{mis}}|X, \phi_1)\\
&P(Y_2^{\text{mis}}|X, Y_1^*, \phi_2)\\
&\ldots\\
&P(Y_k^{\text{mis}}|X, Y_1^*, \ldots, Y_{k-1}^*, \phi_k)
\end{aligned}
$$

where notation Y_j^* stands for the jth imputed variable. The sequence can be replicated m times from different starting points to obtain multiple imputations. Univariate methods such as listed in Table 1 can be used as building blocks. There is no need to iterate.



Univariate pattern Monotone pattern



File matching pattern General pattern

Figure 2 Four types of missing data patterns in multivariate data. The grey parts represent observed data, whereas the empty parts indicate missing data.

Since this procedure is convenient, it is often useful to identify whether the data can be ordered to a (nearly) monotone pattern. It is beneficial to impute to entries that destroy the monotone pattern first and then apply the above method.[7,77,78] It may however be impossible to reorder variables into a monotone pattern. In that case, we need a truly multivariate imputation method.

Table 1 Overview of imputation methods in univariate missing data problems

Type of variable	Method	References
Ignorable methods		
Continuous	Linear regression	Rubin (1987)[1] Schenker and Taylor (1986)[54]
	Linear regression + empirical residuals	Rubin (1987)[1] Schenker and Taylor (1986)[54]
	Predictive mean matching	Rubin (1986)[55] Little (1988)[56] Schenker and Taylor (1986)[54]
	Nonlinear regression	Harrell (2001)[57]
	Truncated normal model	Schafer (1997, p. 204)[7]
Binary	Logistic regression	Rubin (1987, p. 169)[1]
	Probit regression	Albert and Chib (1993)[58]
	Measurement error and reporting model	Yucel and Zaslavsky (2005)[59]
Categorical	Polytomous logistic regression	Brand *et al.* (2003)[50]
	Discriminant analysis	Brand (1999)[60]
Semi-continuous[a]	Two step: logistic + linear	Rubin (1987, p. 180)[1]
	General location model	
Counts	Poisson regression	Raghunathan *et al.* (2001)[61]
General	Approximate Bayesian bootstrap	Rubin (1987)[1] Parzen *et al.* (2005)[62]
	Hot-deck	Reilly and Pepe (1997)[63]
	Machine learning methods	Junninen (2004)[64]
	Polya tree	Paddock (2002)[65]
Nonignorable methods		
Continuous	Normal selection model	Heckman (1976)[66]
	Logit selection model	Greenlees *et al.* (1983)[67]
Censored data	Data augmentation	Wei and Tanner (1991)[68]
Clustered censored data	GEE	Pan and Connett (2001)[69]
Interval censored	Proportional hazard model	Goetghebeur and Ryan (2000)[70] Pan (2000)[71]
Limited dependent variables	DeFries–Fulker regression	Bechger *et al.* (2002)[72]
Below detection limit	Custom model	Hopke *et al.* (2001)[73] Lubin *et al.* (2004)[74]
Pedigree relations	Custom model	Fridley (2003)[75]
Bracketed responses	Custom model	Heeringa *et al.* (2002)[76]

[a] A mixture of discrete and continuous distribution.

4 Multivariate imputation methods

4.1 Joint modeling (JM)

The JM approach partitions the observations into groups of identical missing data patterns and imputes the missing entries within each pattern according to a joint model for X, Y and R that is common to all observations. The first such model was published by Rubin and Schafer.[77] Schafer developed sophisticated JM methods for generating multivariate imputations under the MVN, the log-linear and the general location model.[7] These methods start by specifying a parametric multivariate density $P(Y, X, R|\theta)$ for the data Y, X and R given the model parameters θ. Under an appropriate prior distribution for θ, it is possible to derive the appropriate submodel for each missing data pattern, from which imputations are drawn, usually under the assumption of an ignorable missing data mechanism. These methods are available as tools in S-Plus 7.0 and SAS V8.2 and are widely applied.

4.2 Fully conditional specification (FCS)

The FCS approach is to impute the data on a variable-by-variable basis by specifying an imputation model per variable. FCS is an attempt to define $P(Y, X, R|\theta)$ by specifying a conditional density $P(Y_j|X, Y_{-j}, R, \theta_j)$ for each Y_j. This density is used to impute Y_j^{mis} given X, Y_{-j} and R. Starting from simple guessed values, imputation under FCS is done by iterating over all conditionally specified imputation models. Methods listed in Table 1 may act as building blocks. One iteration consists of one cycle through all Y_j. If the joint distribution defined by the specified conditional distributions exists, then this process is a Gibbs sampler.

FCS has some practical advantages over JM. FCS allows tremendous flexibility in creating multivariate models. One can easily specify models that are outside any known standard multivariate density $P(X, Y, R|\theta)$. FCS can use specialized imputation methods that are difficult to formulate as a part of a multivariate density $P(X, Y, R|\theta)$. Imputation methods that preserve unique features in the data, for example, bounds, skip patterns, interactions, bracketed responses and so on can be incorporated. It is straightforward to maintain constraints between different variables in order to avoid logical inconsistencies in the imputed data. It would be rather difficult to formulate such constraints in terms of the multivariate density $P(X, Y, R|\theta)$. Each conditional density has to be specified separately, so some modeling effort may be required on the part of the user. Computational shortcuts like the sweep operator[6] cannot be used anymore, so the calculations could be more intensive than for JM.

Despite the lack of a satisfactory theory, FCS seems to work quite well in many applications. A number of simulation studies provide evidence that FCS generally yields estimates that are unbiased and that possess appropriate coverage, at least in the variety of cases investigated.[50,60,61,79,80]

The basic idea of FCS is already quite old and has been proposed using a variety of names: stochastic relaxation,[81] variable-by-variable imputation,[60] regression switching,[52] sequential regressions,[61] ordered pseudo-Gibbs sampler,[82] partially incompatible MCMC,[78] iterated univariate imputation,[83] chained equations[84] and FCS.[79]

4.3 Relations between FCS and JM

FCS is related to JM in some cases. If $P(X, Y)$ has an MVN model distribution, then all conditional densities are linear regressions with a constant normal error variance. So, if $P(X, Y)$ is MVN then $P(Y_j|X, Y_{-j})$ follows a linear regression model. The reverse is also true: if the imputation models $P(Y_j|X, Y_{-j})$ are all linear with constant normal error variance, then the joint distribution will be MVN. We refer to Arnold *et al.* (p. 186) for a description of the precise conditions.[85] Thus, imputation by FCS using all linear regressions is identical to imputation under the MVN model. In that case, the algorithm is a real Gibbs sampler, and convergence is guaranteed.

Another special case occurs for binary variables with only two-way interactions in the log-linear model. For example, in the case $k = 3$, suppose that $Y_1, \ldots, Y_3$ are modeled by the log-linear model that has the three-way interaction term set to zero. It is known that the corresponding conditional distribution $P(Y_1|Y_2, Y_3)$ is the logistic regression model $\log(P(Y_1)/1 - P(Y_1)) = \beta_0 + \beta_2 Y_2 + \beta_3 Y_3$.[86] Analogous definitions exist for $P(Y_2|Y_1, Y_3)$ and $P(Y_3|Y_1, Y_2)$. This means that if we use logistic regressions for Y_1, Y_2 and Y_3, we are effectively imputing under multivariate 'no three-way interaction' log-linear model. In this case, the method is also a Gibbs sampler.

5 Issues in FCS

5.1 Compatibility

It is quite easy to specify a set of conditional distributions for which no multivariate density exists. An example is the combination of $Y_2|Y_1 \sim N(\alpha_2 + \beta_1 Y_1, \sigma_1^2)$ with $Y_1|Y_2 \sim N(\alpha_1 + \beta_2 \log(Y_2), \sigma_2^2)$, but the issues involved are actually quite subtle. Incompatibility is a theoretical weakness of FCS, because it is not known to which multivariate distribution the algorithm converges. The limiting distribution to which the algorithm converges may depend on the order of the univariate imputation steps, which may or may not be desirable in a given context. Consequently, assessing convergence is a somewhat ambiguous activity. The issue is known as *incompatibility of conditionals*, and has been studied by various authors.[85,87–89] Gelman and Speed[89] showed that the joint distribution for $Y_1, \ldots, Y_3$, if it exists, is uniquely specified by the following set of three conditionals: $P(Y_1|Y_2, Y_3), P(Y_2|Y_3)$ and $P(Y_3|Y_1)$. Imputation under FCS typically specifies general forms for $P(Y_1|Y_2, Y_3)$, $P(Y_2|Y_1, Y_3)$ and $P(Y_3|Y_1, Y_2)$ and estimates the free parameters for these conditionals from the data. Typically, the number of parameters in imputation is much larger than needed to uniquely determine $P(Y_1, Y_2, Y_3)$.

Not much is known about the consequences of incompatibility on the quality of imputations. Van Buuren *et al.*[79] report some simulations under some strongly incompatible models and observe that the adverse effects on the estimates after MI were only minimal. More work is needed to verify such claims in more general and more realistic settings.

In cases where the multivariate density is of genuine scientific interest, incompatibility clearly represents a problem because the data cannot be represented by a formal model. So given the dual role of $P(Y, X, R)$ for both analysis and imputation (Section 2.2), incompatibility is clearly undesirable within a JM context. In imputation however, the objective is to augment the data and preserve the relations in the data. In that case, the joint distribution is more like a nuisance factor that has no intrinsic value. Gelman remarked: 'One may argue that having a joint distribution in the imputation is less important than incorporating information from other variables and unique features of the dataset (e.g., zero/nonzero features in income components, bounds, skip patterns, nonlinearity, interactions).[83]

FCS is highly important from a practical point of view because it adapts so well to the data. FCS is guaranteed to work if the conditionals are compatible and some evidence is available on the robustness of FCS against incompatibility.

5.2 Assessment of convergence

When m sampling streams are calculated in parallel, monitoring convergence is done by plotting the draws in each stream against time for a set of selected parameters. The pattern should be inspected for any absence of trend, and convergence can be assessed by test statistics that combine within and between variation.[90]

In practice, we have seen many cases where essentially nothing happened after the first few iterations. In those applications, we have therefore set the main number of FCS iterations quite low, usually somewhere between 5 and 20 iterations. This number is much lower than in other applications of MCMC methods, which often require thousands of iterations. There are exceptions however. In order to demonstrate this, consider a small simulation experiment with three variables: one complete covariate X and two incomplete variables Y_1 and Y_2. The data consisted of 10 000 draws from the MVN distribution with correlations $\rho(X, Y_1) = \rho(X, Y_2) = 0.9$ and $\rho(Y_1, Y_2) = 0.7$. The number of complete cases (CC) was varied as $n_{CC} = (1000, 500, 250, 100, 50, 0)$. Missing data were randomly created in two patterns (X, NA, Y_2) and (X, Y_1, NA), both of size $(10\,000 - n_{CC})/2$, where symbol 'NA' stands for the missing entry. A missing data pattern like this may result in statistical matching problems, where Y_1 and Y_2 are jointly observed only for a subset of n_{CC} cases.[55] The difficulty in this particular problem is that the correlation $\rho(Y_1, Y_2)$ under conditional independence of Y_1 and Y_2 given X is equal to $0.9 \times 0.9 = 0.81$, whereas the true value equals 0.7. We used compatible linear regressions $Y_1 = \beta_{1,0} + \beta_{1,2}Y_2 + \beta_{1,3}X + \varepsilon_1$ and $Y_2 = \beta_{2,0} + \beta_{2,1}Y_1 + \beta_{2,3}X + \varepsilon_2$ to impute Y_1 and Y_2, so the algorithm is a Gibbs sampler.

Figure 3 shows the development of $\rho(Y_1, Y_2)$ calculated on the completed data after every iteration of the Gibbs sampler. At iteration 1, $\rho(Y_1, Y_2)$ is around 0.40 (not shown in the figure), due to the random starting imputations. At iteration 2, $\rho(Y_1, Y_2)$ jumps to the value expected given X only. After iteration 2, the influence of the n_{CC} pairs with both Y_1 and Y_2 observed percolates into the imputations, so that the chains slowly move into the direction of the population value of 0.7. The speed of convergence heavily depends on the value of n_{CC}. If $n_{CC} = 1000$, that is, if 90% of the record are incomplete, the streams are essentially flat after about 15 iterations. If $n_{CC} = 0$, the correlation $\rho(Y_1, Y_2)$ is unidentified because there is no information about it in the data. The streams do not converge at all, and wander widely within the Cauchy–Schwarz bounds (0.6–1.0 here). The Cauchy–Schwarz inequality provides the upper and lower bounds for a correlation $\rho(Y_1, Y_2)$ in a positive semi-definite correlation matrix. The lesson from this simulation is that we should be quite careful about convergence in missing data patterns that results from, for example, statistical matching problems.

One final note of interest in this analysis is the following. In the case $n_{CC} = 0$ we could stop at iteration 200 and take the imputations from there. From a Bayesian perspective, this still would yield a valid inference on $\rho(Y_1, Y_2)$. The mean value of $\rho(Y_1, Y_2)$ was equal to 0.812, and its standard error after pooling was large for this sample size: 0.087.

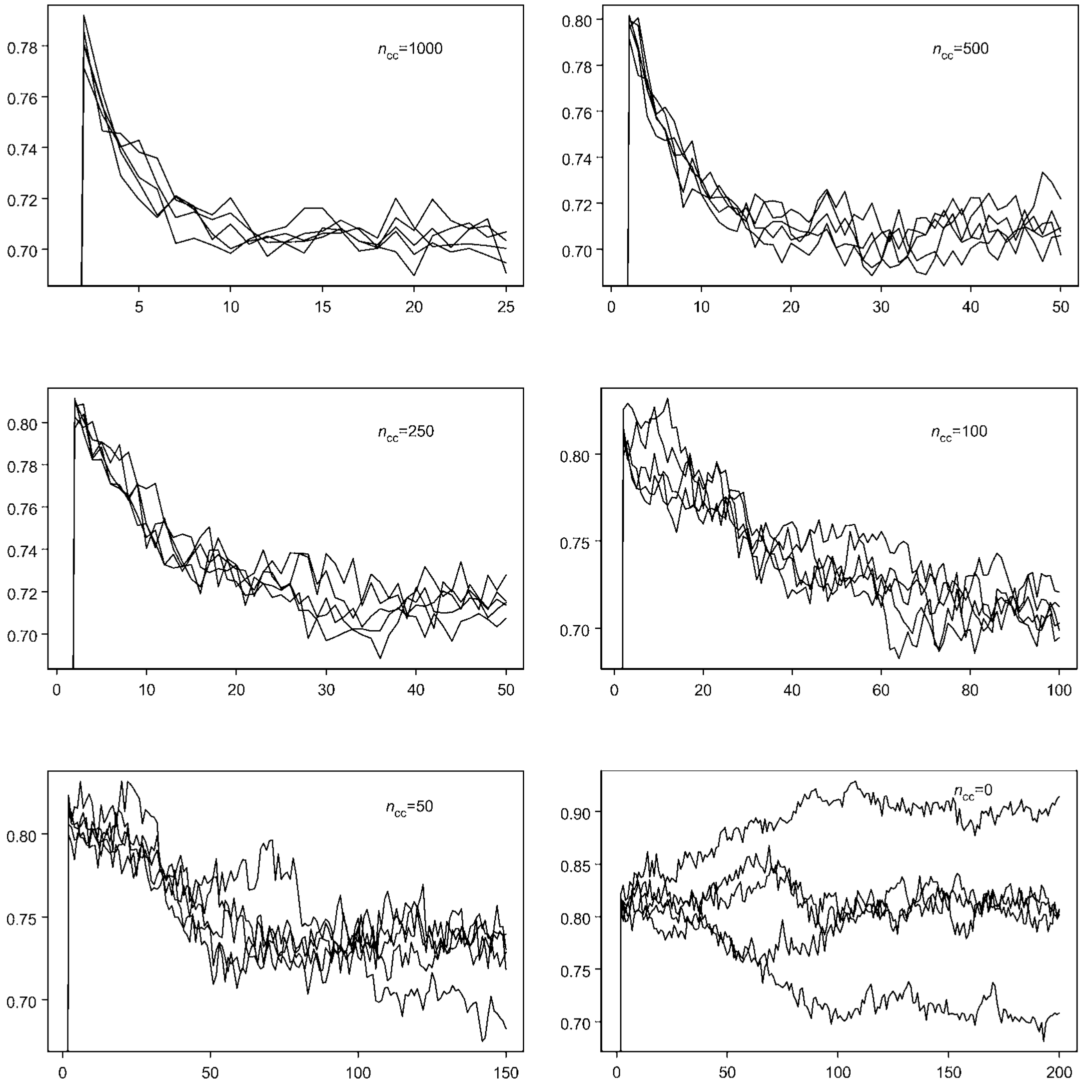


Figure 3 Correlation between Y_1 and Y_2 in the imputed data per iteration in five independent runs of the Gibbs sampler. The number n_{CC} represents the sample size for which both Y_1 and Y_2 are observed. The vertical line at iteration 1 represents the jump in correlation that occurs when the unconditional random starting imputation is replaced by the first conditional imputation.

This is a signal that $\rho(Y_1, Y_2)$ can be anywhere within interval defined by the Cauchy–Schwarz bounds. Under the assumption of a flat prior distribution of an unidentified parameter, this adequately summarizes the available evidence about $\rho(Y_1, Y_2)$. So even in this pathological case with 100% missing data, the analysis tells the appropriate story. The key factor here is that the appropriate amount of variation between streams is achieved. As long as that is the case, pooling under MI seems to acts as a safety net for estimates that are off-target.

Of course, it never hurts to do a couple of extra iterations or to start more streams, but good results can often be obtained with a small number of iterations.

5.3 Software

Systems for creating multiple imputations by FCS include FRITZ,[81] IVEWARE in SAS,[61] HERMES missing data engine,[60] MICE in S-Plus and R,[84] and ICE, an implementation of MICE to Stata.[91–92]

6 Fourth Dutch Growth Study

The Fourth Dutch Growth Study[93] collected data on 14 500 Dutch children between 0 and 21 years. The development of secondary pubertal characteristics was measured by the so-called Tanner stages, which divides the continuous process of maturation into discrete stages for the ages between 8 and 21 years.[94] Stages for girls are defined for menarche (two stages), breast development (five stages B1–B5), and pubic hair (six stages P1–P6). Collecting the data requires the examination of the child by a trained nurse. In the growth study, many children did not receive Tanner scores, usually because the nurse felt that the measurement was ‘unnecessary’, or because the child did not give permission. Table 2 provides the contingency table of the data. Table 3 lists the response patterns for the three measures in 3801 girls (out of 3804) that had complete information on age, height and weight. Strictly speaking, age, height and weight are not completely observed covariates because they had three missing values in the original sample of 3804 girls. For the matter of illustration, these three rows are ignored here, so age, height and weight are assumed to be complete covariates. About 34% of the pubertal data were missing. Figure 4 shows that older girls had more missing values in scores for breast development and pubic hair.

Mul *et al.*[95] published reference curves for these data by deleting all girls that had one or more missing scores. The analysis by Mul *et al.* consisted of a regression of an incompletely observed outcome (Tanner stage) on a completely observed covariate (age). Under the assumption of ignorability, this CC analysis will not bias the age-conditional reference intervals,[96] though it may create sparse data, especially for the older girls. In addition, deleting incomplete records in analyses where the Tanner stages have a role as predictors may yield biased estimates. In order to study the influence of these effects, we multiply imputed the missing Tanner stages.

The data consist of three complete covariates (X_1 = Age, X_2 = Height, X_3 = Weight) and three incomplete variables (Y_1 = Menarche, Y_2 = Breast stage, Y_3 = Pubic hair stage). The data are imputed five ($m = 5$) times by two multivariate methods: MVN and FCS. The MVN method draws imputations under the MVN model and rounds the imputations to the nearest integer to accommodate the categorical nature of the Tanner stages. The FCS method creates imputations for Y_1 by means of logistic regression conditional on X and Y_{-1} under the standard noninformative prior (Rubin, p. 169).[1] For Y_2 and Y_3, imputations were generated by polytomous logistic regression (Brand,

Table 2 Frequency table of pubertal development: menarche status, breast development and pubic hair of 3801 Dutch girls

Menarche status	Breast development stage	Pubic hair stage						
		P1	P2	P3	P4	P5	P6	Missing
No	B1	458	53	5	0	0	0	11
	B2	121	131	50	4	0	0	11
	B3	19	47	100	39	6	0	8
	B4	0	2	25	58	15	2	5
	B5	0	1	0	13	12	1	0
	Missing	0	0	1	0	0	0	155
Yes	B1	6	1	0	0	0	0	0
	B2	2	3	0	0	0	0	0
	B3	0	2	14	19	10	5	3
	B4	0	0	11	127	141	21	4
	B5	0	0	6	53	489	128	6
	Missing	0	0	1	0	1	0	587
Missing	B1	6	1	0	0	0	0	0
	B2	1	3	0	0	0	0	2
	B3	0	1	3	0	1	0	1
	B4	0	0	0	2	0	0	2
	B5	0	0	0	3	2	1	1
	Missing	0	0	1	0	0	0	777

Source: Fourth Dutch Growth Study.[93]

1999, Ch. 4).[60] For Y_2 the generalized logit model for polytomous categories[97]

$$\ln \frac{P(Y_2 = c)}{P(Y_2 = 1)} = [X, Y_{-2}]\beta'_c \quad \text{for } c = 2, \ldots, 5$$

was fitted by the `multinom()` function of Venables and Ripley (2002).[98] This function yields estimates $\hat{\beta} = [\hat{\beta}_2, \ldots, \hat{\beta}_5]$, and its posterior variance–covariance $\hat{V}(\hat{\beta})$ was calculated by the function `vcov()`. A random draw is made from $\beta^* \sim N(\hat{\beta}, \hat{V}(\hat{\beta}))$, which is then plugged back into the object generated by `multinom()`. For each observation with missing Y_2, the function `predict.multinom()` calculated the class probability

Table 3 Response patterns (1 = observed, 0 = missing) of pubertal characteristics of 3801 girls from the Fourth Dutch Growth Study[93]

	Menarche	Breast	Pubic hair	Frequency
	1	1	1	2200
	0	1	1	24
	1	1	0	48
	1	0	1	3
	0	1	0	6
	0	0	1	1
	1	0	0	742
	0	0	0	777
Total	808	1523	1573	3801

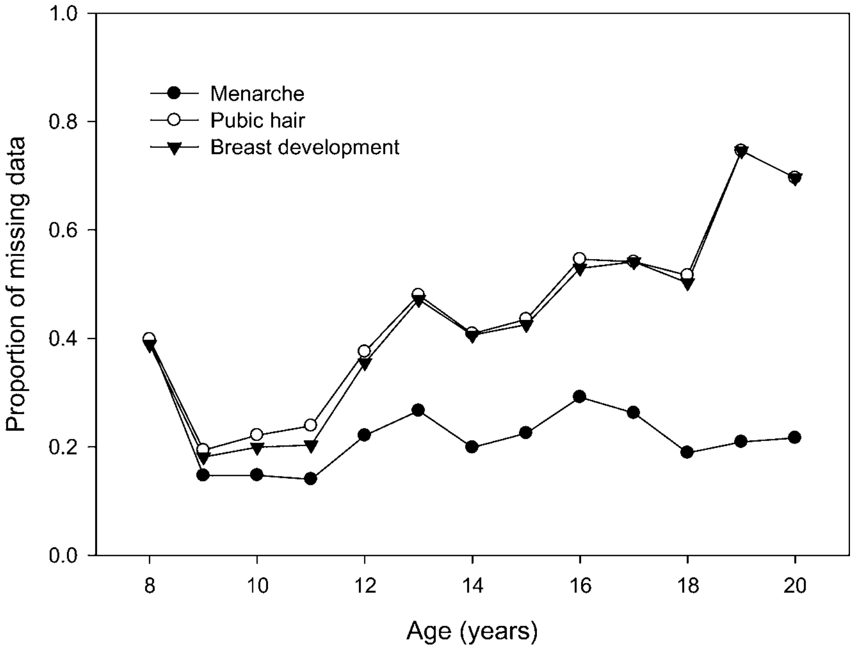


Figure 4 The probability of missingness for menarche, breast development and pubic hair development as a function of age of the girl.
Source: Fourth Dutch Growth Study.[93]

conditional on X and Y_{-2}, which were then used to draw imputations for the missing category score. Brand *et al.*[50] investigated the quality of the imputations of this method and found that it leads to minimal bias and appropriate coverage under a variety of missing data mechanisms. An analogous procedure was followed for imputing Y_3.

The above procedure for polytomous regression becomes computationally prohibitive if sample size is large, as finding $\hat{V}(\hat{\beta})$ requires calculation of the Hessian matrix. An alternative is not to draw β^* from its posterior but set it equal to the 'plug-in estimate', that is $\beta^* = \hat{\beta}$. Such a procedure is improper in terms of Rubin as it ignores the variability of $\hat{\beta}$. However, the difference between using the proper procedure and the plug-in methods is generally quite small if sample size is large. As the sample consisted of about 2200 complete records, we used the fast plug-in estimate.

After imputation, we conducted several complete-data analyses that assessed different aspects of the solution. These analyses were performed on 1) CC, 2) the imputed and rounded data under the fully normal model (MVN) and 3) the imputed data under the FCS. All calculations were performed in S-Plus using the MICE V1.12 library.[84]

We used correspondence analysis of Y_2 and Y_3 to investigate how well imputation preserves the structure between the stages of B1–B5 and P1–P6. For a three-dimensional solution, the CC analysis yielded canonical correlations of 0.940, 0.613 and 0.385. Under MVN, the canonical correlations averaged over the five imputed data sets were equal to 0.927, 0.647 and 0.402. For FCS, we obtained 0.940, 0.627 and 0.396, which is slightly closer to the CC analysis. The scale values per category were quite similar in the different solution.

Next, we modeled the distribution of body weight (X_3) for a given age (X_1), height (X_2) and stages of pubertal development ($Y_1, \dots, Y_3$). Table 4 contains the results of modeling log weight by a simple linear model with only main effects under the three missing data methods. Pubic hair (Y_3) was not a significant predictor in any model, and was therefore omitted. Due to a larger sample size, the standard errors of the estimates of MVN or FCS are smaller than of CC. The models predict equally well: all had $r^2 = 0.79$. For MVN and FCS, r^2 was calculated by taking the average r^2 of the five regressions. Though some differences occur in the individual estimates (e.g., for menarche, age) or in the fraction of missing information (e.g., for B2), the overall impression is that the models behave very similarly.

The above analysis suggests that the results of rounded MVN and FCS hardly differ, but that conclusion would not be correct. In fact, the methods may lead to substantially different estimates for the reference curves. We refitted the reference curves on the imputed data and compared the results to the curves published by Mul *et al.*[95] For each stage transition of breast development, a reference curve was fitted conditional on age by a series of four logistic additive models

$$\log \frac{P(Y_2 < c)}{P(Y_2 \geq c)} = \alpha_c + f_c(X_1), \quad c = 2, \dots, 5$$

where $f_c()$ are arbitrary univariate functions of age (X_1). These models were fitted by the S-Plus `gam()` function with a binomial distribution with a logit link. Hastie and Tibshirani (1990, Ch. 6)[99] for more details. The default number of degrees of smoothness (df = 4) generally provided a good compromise between smoothness and fit, and was used in all analyses.

Table 4 Parameters estimates of three linear regression models for predicting 100 log (body weight in kg) in Dutch girls

Parameter	CC		MVN			FCS		
	est	se	est	se	fmi	est	se	fmi
Age (yrs)	0.72	0.18	0.97	0.12	0.03	0.77	0.12	0.02
Height (cm)	1.27	0.04	1.25	0.03	0.05	1.25	0.03	0.01
Menarche	2.80	0.77	2.65	0.61	0.23	3.85	0.90	0.23
B1*	0		0			0		
B2	5.37	0.98	5.27	0.94	0.24	5.44	0.85	0.06
B3	8.98	1.18	9.30	0.98	0.09	8.50	0.98	0.08
B4	11.32	1.44	11.82	1.15	0.08	11.23	1.22	0.19
B5	18.13	1.62	16.62	1.37	0.18	17.83	1.30	0.12
Intercept	164.72	6.00	164.44	4.58	0.07	164.66	4.35	0.01
n	2200			3801			3801	
r^2	0.79			0.79			0.79	

The CC analysis uses only the complete cases ($n = 2200$), the rounded MVN and FCS methods are fitted on the full sample after multiple imputation and pooling ($n = 3801$). Stage B1 is the reference stage. fmi = fraction of missing information. Symbol r^2 denotes the proportion of variance explained by the model.

Under the assumption that the breast stage data are ignorable given $[X, Y_{-2}^{\text{obs}}]$, the reference curves emanating from the imputed data and from the CC have the same expectation. Figures 5 and 6 contain the resulting references under the MVN and FCS methods. The thick lines are the published reference curves based on the CC only. Under ignorability, the reference curves from the imputed data should on average be equal to the published curves.

Figure 5 shows that the rounded MVN method produces biased estimates at several points. For very young children, the MVN method results in probabilities for stage B2 which are too high. For example, the imputed data indicates that at an age of 8.5 years 10% of the girls have entered stage B2. According to the complete data analysis (which is valid here), that point is actually located at about 9.0 years. At the other end of the age spectrum, the method overestimates the age at which 50% of the girls have entered the final stage B5 by more than eight months (15.0 years instead of 14.3 years). These are large and clinically relevant differences. In general, the rounded MVN produces imputations that do not follow the bends and twists in the observed data. Note that approximately half of the case is imputed here, so the effects of the imputed data on the results are attenuated by the observed data. Analyzing just the cases with the imputed values would lead to even larger discrepancies.

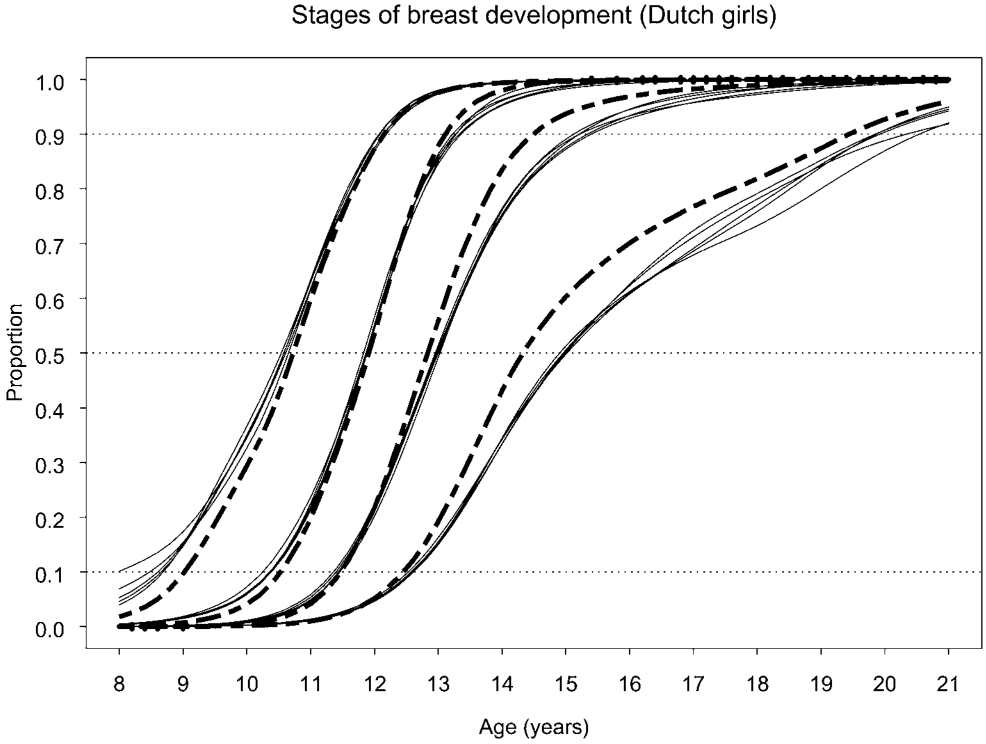


Figure 5 Reference curves per stage of breast development according to two methods: complete cases only (thick lines) and multiple imputation ($m = 5$) under a rounded multivariate normal (MVN) model imputation model. The MVN model is off target.

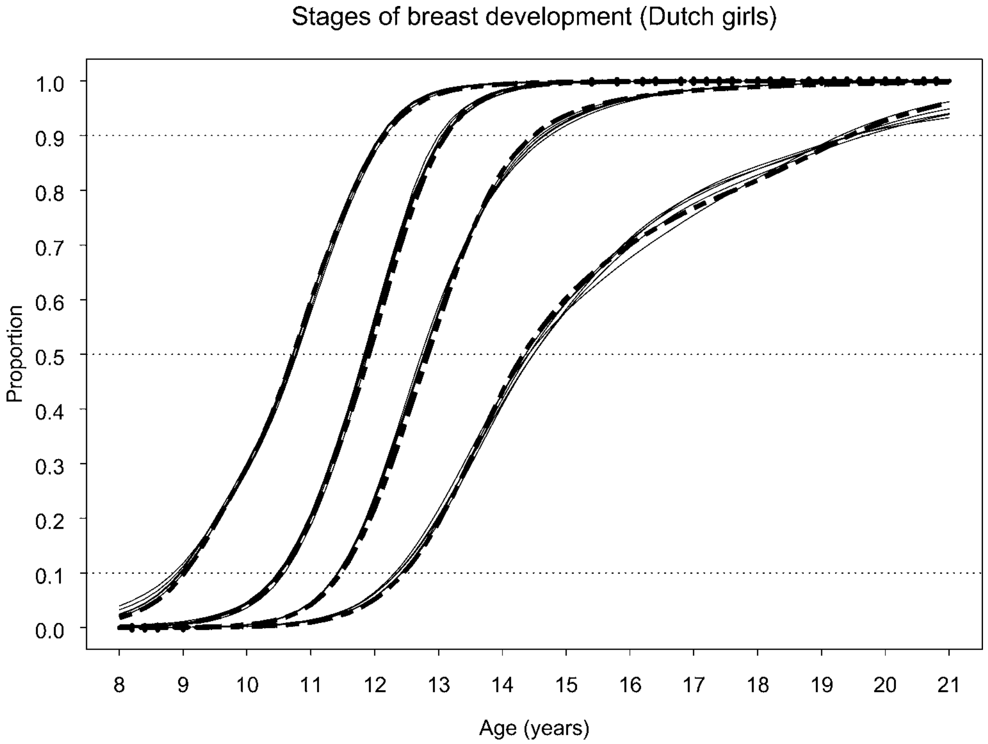


Figure 6 Reference curves per stage of breast development according to two methods: complete cases only (thick lines) and multiple imputation ($m = 5$) under a conditionally specified (FCS) imputation model. The FCS model is on target.

In contrast, the FCS imputation method in Figure 6 behaves very well. There is a tendency that imputation leads to somewhat smoother reference curves because of the higher sample size, but the effect is only slight. All in all, we conclude that the FCS method preserves the important features in the relationship between breast stage and age that are ignored in rounded MVN.

Like Horton *et al.*, Ake and Allison[100–102] we therefore do not recommend the rounded MVN method when data are categorical. Horton *et al.* expected that bias problems of rounding would taper off if variables have more categories, but our analyses suggest the MVN methods may introduce biases also for discrete data with more than two categories. The FCS method appears to be free of such problems.

7 Discussion

Creating imputations in multivariate health data is not an easy task. The ultimate goal of imputation is to yield valid inferences for the statistical estimates of interest from the imputed data. To achieve that goal, imputation should preserve the structure in the data, as well as the uncertainty about this structure, and include any knowledge about the process that generated the missing data. Two main approaches have been proposed, JM and FCS. JM stays close to the theory and leads to imputation procedures whose statistical properties are known under a correctly specified model. FCS is its semi-parametric and flexible cousin that emphasizes features in the data.

Several authors have been critical on JM in particular contexts. Schenker and Taylor[54] performed a simulation study and observed that 'the fully parametric method breaks down in several situations, whereas the partially parametric methods maintain their good performance'. Belin *et al.*[103] assessed the usefulness of the general location model for a mental health services study and conclude: 'Our investigations suggest that either the model or the companion assumption of ignorable nonresponse are not suitable in our applied context with numerous variables and a complicated pattern of missing data.' Gelman and Raghunathan[104] address the difficulty of maintaining consistencies in the imputed data and note that 'separate regressions often make more sense than joint models'. Briggs *et al.*[105] imputed cost data and wrote 'using the algorithm based on multivariate normality resulted in failure of the algorithm to converge', and were forced to dichotomize their data. In order to bypass the limitations of joint models, Gelman (p. 541) concludes: 'Thus we are suggesting the use of a new class of models – inconsistent conditional distributions – that were initially motivated by computational and analytical convenience.'[83]

Within the JM context, the data often need to be transformed before imputation (to make the observed data conform to the imputation model) and after imputation (to make the imputed values conform to the observed data) (Schafer, 1997 p. 147–148, 202–203, 214, 272, 374).[7] While such transformations enhance the performance of the JM. Various authors[100–102] observed that rounding imputed values to the closest observed value in the data can introduce a bias in the parameter estimates, whereas if the imputed data are not rounded, no bias would occur. Our analysis of the pubertal data provide evidence that rounding bias in JM may also show up in real life categorical data with more than two categories. Chen *et al.*[107] also provide some support for the idea that normal methods do not work well for ordinal data. Our analysis of the pubertal data showed that FCS appears to be less sensitive to such biases. We therefore recommend that continuous data are imputed as continuous and discrete data are imputed as discrete. Conditional specification is the most convenient way to do that. Despite its theoretical weaknesses, we conclude that FCS is a useful and flexible alternative to JM when the joint distribution of the data is not easily specified.

Missing data problems require careful consideration and thought. It will be clear by now that MI is not an automatic technical fix for the missing data. Rather, it is a general and principled strategy for attacking missing data problems. The process of specifying the imputation model is a scientific modeling activity on its own, that comes with its own model building principles. The fact that highly automated and sophisticated procedures are available does not free the imputer or the analyst from the responsibility to consider the appropriateness of the assumptions underlying the imputation model for the problem at hand. The implication is that medical researchers should include a short description of their missing data method into their scientific articles. The most natural location for that description is the section on the statistical analysis.

Acknowledgement

I thank Peter van der Heijden, Ian White and Patrick Royston for their constructive and insightful feedback on an earlier draft of this paper.

References

1 Rubin DB. *Multiple imputation for nonresponse in surveys*. Wiley, 1987.

2 Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**(434): 473–89.

3 Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001; **6**(3): 330–51.

4 Scheuren F. Multiple imputation: how it began and continues. *American Statistician* 2005; **59**(4): 315–9.

5 Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology* 1977; **39**: 1–38.

6 Little RJA, Rubin DB. *Statistical analysis with missing data*. second edition Wiley, 2002.

7 Schafer JL. *Analysis of incomplete multivariate data*. Chapman & Hall, 1997.

8 Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research* 1999; **8**(1): 3–15.

9 Stern HS, Sinharay S, Russell D. The use of multiple imputation for the analysis of missing data. *Psychological Methods* 2001; **6**(3): 317–29.

10 Allison PD. *Missing data*. Sage, 2002.

11 Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological Methods* 2002; **7**(2): 147–77.

12 Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine* 1991; **10**(4): 585–98.

13 Barnard J, Meng XL. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research* 1999; **8**(1): 17–36.

14 Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology* 1995; **142**(12): 1255–64.

15 Kmetic A, Joseph L, Berger C, Tenenhouse A. Multiple imputation to account for missing data in a survey: estimating the prevalence of osteoporosis. *Epidemiology* 2002; **13**(4): 437–44.

16 Abraham WT, Russell DW. Missing data: a review of current methods and applications in epidemiological research. *Current Opinion in Psychiatry* 2004; **17**(4): 315–21.

17 Croy CD, Novins DK. Methods for addressing missing data in psychiatric and developmental research. *Journal of the American Academy of Child and Adolescent Psychiatry* 2005; **44**(12): 1230–40.

18 Kneipp SM, McIntosh M. Handling missing data in nursing research with multiple imputation. *Nursing Research* 2001; **50**(6): 384–9.

19 Patrician PA. Multiple imputation for missing data. *Research in Nursing and Health* 2002; **25**(1): 76–84.

20 McCleary L. Using multiple imputation for analysis of incomplete data in clinical research. *Nursing Research* 2002; **51**(5): 339–43.

21 Fox-Wasylyshyn SM, El-Masri MM. Handling missing data in self-report measures. *Research in Nursing and Health* 2005; **28**(6): 488–95.

22 Molenberghs G, Burzykowski T, Michiels B, Kenward MG. Analysis of incomplete public health data. *Revue d'Epidemiologie et de Sante Publique* 1999; **47**(6): 499–514.

23 Zhou XH, Eckert GJ, Tierney WM. Multiple imputation of public health research. *Statistics in Medicine* 2001; **20**(9–10): 1541–9.

24 Raghunathan TE. What do we do with missing data? Some options for analysis of

incomplete data. *Annual Review of Public Health* 2004; 25: 99–117.

25 Crawford SL, Tennstedt SL, McKinlay JB. A comparison of analytic methods for non-random missingness of outcome data. *Journal of Clinical Epidemiology* 1995; **48**(2): 209–19.

26 Faris PD, Ghali WA, Brant R, Norris CM, Galbraith PD, Knudtson ML, APPROACH Investigators. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Journal of Clinical Epidemiology* 2002; **55**(2): 184–91.

27 Oostenbrink JB, Al MJ. The analysis of incomplete cost data due to dropout. *Health Economics* 2005; **14**(8): 763–76.

28 Chavance M. Handling missing items in quality of life studies. *Communications in Statistics – Theory and Methods* 2004; **33**(6): 1371–83.

29 Catellier DJ, Hannan PJ, Murray DM, Addy CL, Conway TL, Yang S, Rice JC. Imputation of missing data when measuring physical activity by accelerometry. Med Sci Sports Exerc 2005; **37**(11 Suppl): S555–62.

30 Wood AM, White IR, Hillsdon M, Carpenter J. Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. *International Journal of Epidemiology* 2005; **34**(1): 89–99.

31 Smits N, Mellenbergh GJ, Vorst HCM. Alternative missing data techniques to grade point average: Imputing unavailable grades. *Journal of Educational Measurement* 2002; **39**(3): 187–206.

32 Peugh JL, Enders CK. Missing data in educational research: a review of reporting practices and suggestions for improvement. *Review of Educational Research* 2004; **74**(4): 525–56.

33 Walczak B, Massart DL. Dealing with missing data: part II. *Chemometrics and Intelligent Laboratory Systems* 2001; **58**(1): 29–42.

34 Longford NT. Multilevel analysis with messy data. *Statistical Methods in Medical Research* 2001; **10**(6): 429–44.

35 Olinsky A, Chen S, Harlow L. The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research* 2003; **151**(1): 53–79.

36 Allison PD. Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology* 2003; **112**(4): 545–57.

37 Twisk J, de Vente W. Attrition in longitudinal studies: how to deal with missing data. *Journal of Clinical Epidemiology* 2002; **55**(4): 329–37.

38 Demirtas H. Modeling incomplete longitudinal data. *Journal of Modern Applied Statistical Methods* 2004; **3**(2): 305–21.

39 Streiner DL. The case of the missing data: Methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry* 2002; **47**(1): 68–75.

40 Kristman VL, Manno M. Methods to account for attrition in longitudinal data: do they work? A simulation study. *European Journal of Epidemiology* 2005; **20**(8): 657–62.

41 Little R, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics* 1996; **52**(4): 1324–33.

42 Liu G, Gould AL. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *Journal of Biopharmaceutical Statistics* 2002; **12**(2): 207–26.

43 Houck PR, Maz umdar S, Koru-Sengul T, Tang G, Mulsant BH, Pollock BG, Reynolds CF 3rd. Estimating treatment effects from longitudinal clinical trial data with missing values: Comparative analyses using different methods. *Psychiatry Research* 2004; **129**(2): 209–15.

44 Tang L, Unntzer J, Song J, Belin TR. A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine* 2005; **24**(14): 2111–28.

45 Beunckens C, Molenberghs G, Kenward MG. Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials* 2005; **2**(5): 379–86.

46 Barnes SA, Lindborg SR, Seaman J. Multiple imputation techniques in small sample clinical trials. *Statistics in Medicine* 2006; **25**(2): 233–45.

47 Pigott TD. Missing predictors in models of effect size. *Evaluation and the Health Professions* 2001; **24**(3): 277–307.

48 Ibrahim JG, Herring AH, Chen MH, Lipsitz SR. Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association* 2005; **100**(469): 332-46.

49 Schafer JL. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica* 2003; **57**(1): 19–35.

50 Brand JPL, Van Buuren S, Groothuis-Oudshoorn K, Gelsema ES. A toolkit in SAS for the evaluation of multiple imputation methods. *Statistica Neerlandica* 2003; **57**(1): 36–45.

51 Meng XL. Multiple imputation with uncongenial sources of input (with discussion). *Statistical Science* 1995; **10**: 538–73.

52 Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**(6): 681–94.

53 Abayomi K, Gelman A, Levy M. Diagnostics for multivariate imputations. Assessed from Gelman's weblog November 2005.

54 Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis* 1996; **22**(4): 425–46.

55 Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business Economics and Statistics* 1986; **4**: 87–94.

56 Little RJA. Missing data adjustments in large surveys (with discussion). *Journal of Business Economics and Statistics* 1988; **6**: 287–301.

57 Harrell F. *Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis* Springer, 2001.

58 Albert JH, Chib S. Bayesian analysis of binary and polychotomous variables. *Journal of the American Statistical Association* 1993; **88**: 669–79.

59 Yucel RM, Zaslavsky AM. Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association* 2005; **100**(472): 1123–32.

60 Brand JPL. *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. Erasmus University, 1999.

61 Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 2001; **27**: 85–95.

62 Parzen M, Lipsitz SR, Fitzmaurice GM. A note on reducing the bias of the approximate Bayesian bootstrap imputation variance estimator. *Biometrika* 2005; **92**(4): 971–4.

63 Reilly M, Pepe M. The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medicine* 1997; **16**(1–3): 5–19.

64 Junninen H, Niska H, Ruuskanen J, Kolehmainen M, Tuppurainen K. Methods for imputation of missing values in air quality data sets. *Atmospheric Environment* 2004; **38**(18): 2895–907.

65 Paddock SM. Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse. *Biometrika* 2002; **89**(3): 529–38.

66 Heckman JJ. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 1976; **5**: 475–92.

67 Greenlees WS, Reece JS, Zieschang KD. Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association* 1983; **77**: 251–61.

68 Wei GCG, Tanner MA. Applications of multiple imputation to the analysis of censored regression data. *Biometrics* 1991; **47**(4): 1297–309.

69 Pan W, Connett JE. A multiple imputation approach to linear regression with clustered censored data. *Lifetime Data Analysis* 2001; **7**(2): 111–23.

70 Goetghebeur E, Ryan L. Semiparametric regression analysis of interval-censored data. *Biometrics* 2000; **56**(4): 1139–44.

71 Pan W. A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* 2000; **56**(1): 199–203.

72 Bechger TM, Boomsma DI, Koning H. A limited dependent variable model for heritability estimation with non-random

ascertained samples. *Behavior Genetics* 2002; **32**(2): 145–51.

73 Hopke PK, Liu C, Rubin DB. Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the arctic. *Biometrics* 2001; **57**(1): 22–33.

74 Lubin JH, Colt JS, Hartge P, Camann D, Davis S, Cerhan JR, Severson RK, Bernstein L, Hartge P. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental Health Perspectives* 2004; **112**(17): 1691–6.

75 Fridley B, Rabe K, de Andrade M. Imputation methods for missing data for polygenic models. *BMC Genetics* 2003; **4**(Suppl 1): S42.

76 Heeringa SG, Little RJA, Raghunathan TE. Multivariate imputation of coarsened survey data on household wealth. In Groves RM, Dillman DA, Eltinge JL, Little RJA, eds. *Survey Nonresponse*. Wiley, 2002.

77 Rubin DB, Schafer JL. Efficiently creating multiple imputations for incomplete multivariate normal data. 1990 Proceedings of the Statistical Computing Section, American Statistical Association 1990; 83–8.

78 Rubin DB. Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* 2003; **57**(1): 3–18.

79 Van Buuren S, Brand JPL, Groothuis-Oudshoorn K, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 2006; 76(**12**): 1049–1064.

80 Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *American Statistician* 2001; **55**: 244–54.

81 Kennickell AB. Imputation of the 1989 survey of consumer finances: stochastic relaxation and multiple imputation. ASA 1991 Proceedings of the Section on Survey Research Methods 1991; 1–10.

82 Heckerman D, Chickering DM, Meek C, Rounthwaite R, Kadie C. Dependency networks for inference, collaborative filtering, and data visualisation. *Journal of Machine Learning Research* 2001; **1**: 49–75.

83 Gelman A. Parameterization and Bayesian Modeling. Journal of the American Statistical Association 2004; **99**(466): 537–45.

84 Van Buuren S, Groothuis-Oudshoorn K. *Multivariate imputation by chained equations: MICE V1.0 user's manual*. TNO Quality of Life, 2000, PG/VGZ/00.038.

85 Arnold BC, Castillo E, Sarabia JM. *Conditional specification of statistical models*. Springer, 1999.

86 Goodman LA. The multivariate analysis of qualitative data: interactions among multiple classifications. *Journal of the American Statistical Association* 1970; **65**: 226–56.

87 Besag J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 1974; **36**: 192–236.

88 Arnold BC, Press SJ. Compatible conditional distributions. *Journal of the American Statistical Association* 1989; **84**: 152–6.

89 Gelman A, Speed TP. Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 1993; **55**: 185–8.

90 Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* 1991; **7**: 457–511.

91 Royston P. Multiple imputation of missing values. *The Stata Journal* 2004; **4**: 227–41.

92 Royston P. Multiple imputation of missing values: update of ice. *Stata Journal* 2005; **5**: 527–36.

93 Fredriks MA, van Buuren S, Burgmeijer RJ, Meulmeester JF, Beuker RJ, Brugman E, Roede MJ, Verloove-Vanhorick SP, Wit JM. Continuing positive secular growth change in The Netherlands 1955–1997. *Pediatric Research* 2000; **47**: 316–23.

94 Marshall WA, Tanner JM. Variations in pattern of pubertal changes in girls. *Archives of Diseases in Childhood* 1969; **44**: 291–303.

95 Mul D, Van Buuren S, Frediks MA, Oostdijk W, Verloove-Vanhorick SP,

Wit JM. Pubertal development in The Netherlands 1965–1997. *Pediatric Research* 2001; **50**: 479–86.

96 Little RJA. Regression with missing X's: a review. *Journal of the American Statistical Association* 1992; **87**: 1227–37.

97 McCullagh P, Nelder JA eds. Generalized linear models, second edition. Chapman & Hall, 1989.

98 Venables WN, Ripley BD eds. Modern applied statistics with S, fourth edition Springer-Verlag, 2002.

99 Hastie TJ, Tibshirani RJ. Generalized additive models. Chapman & Hall, 1990.

100 Horton NJ, Lipsitz SR, Parzen M. A potential for bias when rounding in multiple imputation. *American Statistician* 2003; **57**(4): 229–32.

101 Ake CF. Rounding after multiple imputation with non-binary categorical covariates SUGI 30 Proceedings 2005, 112–30, pp. 1–11.

102 Allison PD. Imputation of categorical variables with PROC MI . SUGI 30 Proceedings 2005, 113–30, pp. 1–14.

103 Belin TR, Hu MY, Young AS, Grusky O. Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. *Statistics in Medicine* 1999; **18**: 3123–35.

104 Gelman A, Raghunathan TE. Discussion of Arnold *et al*. Conditionally specified distributions. *Statistical Science* 2001; **16**: 249–74.

105 Briggs A, Clark T, Wolstenholme J, Clarke P. Missing . . . . presumed at random: cost-analysis of incomplete data. *Health Economics* 2003; **12**: 377–92.

106 Chen L, Valois RF, Toma-Drane M, Drane JW. Multiple imputation for missing ordinal data. *Journal of Modern Applied Statistical Methods* 2005; **4**(1): 288–99.