Tutorial in Biostatistics

Statistics in Medicine

Received 3 September 2009, Accepted 14 July 2010 Published online 30 November 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4067

Multiple imputation using chained equations: Issues and guidance for practice

Ian R. White,[a*†] Patrick Royston[b] and Angela M. Wood[c]

Multiple imputation by chained equations is a flexible and practical approach to handling missing data. We describe the principles of the method and show how to impute categorical and quantitative variables, including skewed variables. We give guidance on how to specify the imputation model and how many imputations are needed. We describe the practical analysis of multiply imputed data, including model building and model checking. We stress the limitations of the method and discuss the possible pitfalls. We illustrate the ideas using a data set in mental health, giving Stata code fragments. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: missing data; multiple imputation; fully conditional specification

1. Introduction

1.1. Missing data

Missing data occur in almost all medical and epidemiological research. Inadequate handling of the missing data in a statistical analysis can lead to biased and/or inefficient estimates of parameters such as means or regression coefficients, and biased standard errors resulting in incorrect confidence intervals and significance tests. In all statistical analyses, some assumptions are made about the missing data. Little and Rubin's framework [1] is often used to classify the missing data as being (i) missing completely at random (MCAR—the probability of data being missing does not depend on the observed or unobserved data), (ii) missing at random (MAR—the probability of data being missing does not depend on the unobserved data, conditional on the observed data) or (iii) missing not at random (MNAR—the probability of data being missing does depend on the unobserved data, conditional on the observed data). For example, blood pressure data are MAR if older individuals are more likely to have their blood pressure recorded (and age is included in the analysis), but they are MNAR if individuals with high blood pressures are more likely to have their blood pressure recorded than other individuals of the same age. It is not possible to distinguish between MAR and MNAR from the observed data alone, although the MAR assumption can be made more plausible by collecting more explanatory variables and including them in the analysis.

1.2. Multiple imputation and its rationale

Multiple imputation (MI) [2] is a statistical technique for handling missing data, which has become increasingly popular because of its generality and recent software developments [3, 4]. The key concept of MI is to use the distribution of the observed data to estimate a set of plausible values for the missing data. Random components are incorporated into these estimated values to reflect their uncertainty. Multiple data sets are created and then analyzed individually but identically to obtain a set of parameter estimates. Finally, the estimates are combined to obtain the overall estimates, variances and confidence intervals. Although MI can be implemented under MNAR mechanisms [2, 5], standard implementations assume MAR, and we make this assumption throughout this paper (except in Section 10.4). When correctly implemented,

[a]*MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, U.K.*
[b]*Hub for Trials Methodology Research, MRC Clinical Trials Unit and University College London, 222 Euston Road, London NW1 2DA, U.K.*
[c]*Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Worts Causeway, Cambridge CB2 8RN, U.K.*
[*]*Correspondence to: Ian R. White, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, U.K.*
[†]*E-mail: ian.white@mrc-bsu.cam.ac.uk*

MI produces asymptotically unbiased estimates and standard errors and is asymptotically efficient. The three stages of MI are described formally below:

Stage 1: *Generating multiply imputed data sets*: The unknown missing data are replaced by m independent simulated sets of values drawn from the posterior predictive distribution of the missing data conditional on the observed data. For a single incomplete variable z, this involves constructing an imputation model which regresses z on a set of variables with complete data, say $x_1, x_2, \ldots, x_k$, among individuals with the observed z. Choices of imputation model are discussed in Section 2. Methods for multiple incomplete variables are discussed in Section 1.3. Let $\widehat{\boldsymbol{\beta}}$ and $\mathbf{V}$ be the set of estimated regression parameters and their corresponding covariance matrix from fitting the imputation model. The following two steps are repeated m times. Let $\boldsymbol{\beta}^*$ be a random draw from the posterior distribution, commonly approximated by $\boldsymbol{\beta}^* \sim \text{MVN}(\widehat{\boldsymbol{\beta}}, \mathbf{V})$ [2]. Imputations for z are drawn from the posterior predictive distribution of z using $\boldsymbol{\beta}^*$ and the appropriate probability distribution. This process is known as proper imputation because it incorporates all sources of variability and uncertainty in the imputed values, including prediction errors of the individual values and errors of estimation in the fitted coefficients of the imputation model. Alternatives to proper imputation [6] are not considered here. An alternative way to draw proper imputations, predictive mean matching, is described in Section 4.2.

Stage 2: *Analyzing multiply imputed data sets*: Once the multiple imputations have been generated, each imputed data set is analyzed separately. This is usually a simple task because complete-data methods can be used. The quantities of scientific interest (usually regression coefficients) are estimated from each imputed data set, together with their variance–covariance matrices. The results of these m analyses differ because the missing values have been replaced by different imputations.

Stage 3: *Combining estimates from multiply imputed data sets*: The m estimates are combined into an overall estimate and variance–covariance matrix using Rubin's rules [2], which are based on asymptotic theory in a Bayesian framework. The combined variance–covariance matrix incorporates both within-imputation variability (uncertainty about the results from one imputed data set) and between-imputation variability (reflecting the uncertainty due to the missing information). Suppose $\widehat{\boldsymbol{\theta}}_j$ is an estimate of a univariate or multivariate quantity of interest (e.g. a regression coefficient) obtained from the jth imputed data set j and $\mathbf{W}_j$ is the estimated variance of $\widehat{\boldsymbol{\theta}}_j$. The combined estimate $\widehat{\boldsymbol{\theta}}$ is the average of the individual estimates:

$$\widehat{\boldsymbol{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\widehat{\boldsymbol{\theta}}_j. \tag{1}$$

The total variance of $\widehat{\boldsymbol{\theta}}$ is formed from the within-imputation variance $\mathbf{W} = (1/m)\sum_{j=1}^{m}\mathbf{W}_j$ and the between-imputation variance $\mathbf{B} = (1/(m-1))\sum_{j=1}^{m}(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{\theta}})^2$:

$$\text{var}(\widehat{\boldsymbol{\theta}}) = \mathbf{W} + \left(1 + \frac{1}{m}\right)\mathbf{B}. \tag{2}$$

Single imputation is sometimes considered as an alternative to multiple imputation, but it is unable to capture the between-imputation variance $\mathbf{B}$, hence standard errors are too small.

Wald-type significance tests and confidence intervals for a univariate θ can be obtained in the usual way from a t-distribution; degrees of freedom are given in references [7, 8]. Wald tests can also be constructed for a multivariate $\boldsymbol{\theta}$ [7].

1.3. Multiple imputation by chained equations

In large data sets it is common for missing values to occur in several variables. Multiple imputation by chained equations (MICE) [9] is a practical approach to generating imputations (MI Stage 1) based on a set of imputation models, one for each variable with missing values. MICE is also known as fully conditional specification [10] and sequential regression multivariate imputation [11]. Initially, all missing values are filled in by simple random sampling with replacement from the observed values. The first variable with missing values, x_1 say, is regressed on all other variables $x_2, \ldots, x_k$, restricted to individuals with the observed x_1. Missing values in x_1 are replaced by simulated draws from the corresponding posterior predictive distribution of x_1. Then, the next variable with missing values, x_2 say, is regressed on all other variables $x_1, x_3, \ldots, x_k$, restricted to individuals with the observed x_2, and using the imputed values of x_1. Again, missing values in x_2 are replaced by draws from the posterior predictive distribution of x_2. The process is repeated for all other variables with missing values in turn: this is called a cycle. In order to stabilize the results, the procedure is usually repeated for several cycles (e.g. 10 or 20) to produce a single imputed data set, and the whole procedure is repeated m times to give m imputed data sets.

An important feature of MICE is its ability to handle different variable types (continuous, binary, unordered categorical and ordered categorical) because each variable is imputed using its own imputation model. Suitable choices of imputation models are discussed in Sections 2 and 4.

1.4. Plan of the paper

The remainder of the paper is structured as follows. Section 2 describes and illustrates how to impute missing values in Normally distributed and categorical variables. Section 3 introduces the UK700 data that we use for illustration in the later sections and describes the MICE algorithm. Section 4 shows how to impute missing values in skewed quantitative variables. Section 5 focuses on how to choose the variables in the imputation model, and Section 6 focuses on how to specify the form of the imputation model when non-linear analyses are of interest. Section 7 discusses how to choose the number of imputations. Section 8 suggests how to use multiply imputed data for extended statistical analyses, such as model building and prediction. Section 9 gives an illustrative analysis of the UK700 data. Section 10 discusses theoretical limitations and pitfalls of MICE. We conclude with a general discussion in Section 11, which includes consideration of some alternatives to MICE.

We illustrate the methods using Stata code fragments where appropriate, although knowledge of Stata is not required to understand the paper. Version 11 of Stata, released in July 2009, contains a new suite of `mi` commands [12]. These do not implement MICE, which requires the user-contributed `ice` command [13–17]. Rubin's rules are implemented by the user-contributed `mim` command [18] and by the new `mi estimate` command.

2. Imputing different variable types

We here introduce approaches for imputing missing values in continuous, binary, unordered categorical (nominal) and ordered categorical variables. In this section, and in Section 4, we assume that z is a variable whose missing values we wish to impute from other (complete) variables $\mathbf{x}=(x_1,\ldots,x_k)'$. For simplicity, we assume that $\mathbf{x}$ includes a column of ones, so that k is the number of parameters estimated, including the intercept. Let n_{obs} be the number of individuals with observed z values.

2.1. Continuous variables

A linear regression model is the most common choice of model for imputing Normally distributed continuous variables:

$$z|\mathbf{x};\boldsymbol{\beta}\sim \text{N}(\boldsymbol{\beta}\mathbf{x},\sigma^2). \tag{3}$$

Let $\widehat{\boldsymbol{\beta}}$ be the estimated parameter (a row vector of length k) from fitting this model to individuals with the observed z. Let $\mathbf{V}$ be the estimated covariance matrix of $\widehat{\boldsymbol{\beta}}$, and $\widehat{\sigma}$ be the estimated root mean-squared error. We next draw the imputation parameters σ^*, $\boldsymbol{\beta}^*$ from the exact joint posterior distribution of σ, $\boldsymbol{\beta}$ [2]. First, σ^* is drawn as

$$\sigma^*=\widehat{\sigma}\sqrt{(n_{\text{obs}}-k)/g},$$

where g is a random draw from a χ^2 distribution on $n_{\text{obs}}-k$ degrees of freedom. Second, $\boldsymbol{\beta}^*$ is drawn as

$$\boldsymbol{\beta}^*=\widehat{\boldsymbol{\beta}}+\frac{\sigma^*}{\widehat{\sigma}}\boldsymbol{u}_1\mathbf{V}^{1/2},$$

where $\boldsymbol{u}_1$ is a row vector of k independent random draws from a standard Normal distribution and $\mathbf{V}^{1/2}$ is the Cholesky decomposition of $\mathbf{V}$. Imputed values z_i^* for each missing observation z_i are then obtained as

$$z_i^*=\boldsymbol{\beta}^*\mathbf{x}_i+u_{2i}\sigma^*,$$

where u_{2i} is a random draw from a standard Normal distribution. Closely related approaches that allow for deviations from the Normal assumption for continuous variables are discussed in Section 4.

2.2. Binary variables

The model usually chosen to impute binary z from $\mathbf{x}$ is the logistic regression model,

$$\text{logit}\ \Pr(z=1|\mathbf{x};\boldsymbol{\beta})=\boldsymbol{\beta}\mathbf{x}. \tag{4}$$

Let $\widehat{\boldsymbol{\beta}}$ be the estimated parameter from fitting this model to individuals with the observed z, with estimated variance–covariance matrix $\mathbf{V}$. Let $\boldsymbol{\beta}^*$ be a draw from the posterior distribution of $\boldsymbol{\beta}$, approximated by MVN($\widehat{\boldsymbol{\beta}}$, $\mathbf{V}$) [2]. For each

missing observation z_i, let $p_i^* = [1+\exp(-\boldsymbol{\beta}^*\mathbf{x}_i)]^{-1}$, and draw an imputed value z_i^* as

$$z_i^* = \begin{cases} 1 & \text{if } u_i < p_i^*, \\ 0 & \text{otherwise,} \end{cases}$$

where u_i is a random draw from a uniform distribution on (0, 1). Such a procedure is straightforward to implement.

Problems can arise due to perfect prediction, which occurs when one or more observations has fitted probability exactly 0 or exactly 1. This causes difficulty in drawing $\boldsymbol{\beta}^*$. The same difficulty arises for unordered and ordered categorical variables, and is further discussed in Section 10.3.

2.3. *Unordered categorical variables*

Unordered categorical (nominal) variables z with $L>2$ classes may be modeled using multinomial logistic regression, in which each of the classes has a logistic regression equation comparing the class with the chosen baseline class (1, say):

$$\Pr(z=l|\mathbf{x};\boldsymbol{\beta}) = \left[\sum_{l'=1}^{L} \exp(\boldsymbol{\beta}_{l'}\mathbf{x})\right]^{-1} \exp(\boldsymbol{\beta}_l\mathbf{x}), \tag{5}$$

where $\boldsymbol{\beta}_l$ is a vector of dimension $k=\dim(\mathbf{x})$ and $\boldsymbol{\beta}_1=\mathbf{0}$. Let $\boldsymbol{\beta}^*$ be the usual random draw from a Normal approximation to the posterior distribution of $\boldsymbol{\beta}=(\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_L)$, a vector of length $k(L-1)$. For each missing observation z_i, let $p_{il}^* = \Pr(z_i=l|\mathbf{x}_i;\boldsymbol{\beta}^*)$ $(l=1,\ldots,L)$ be the drawn class membership probabilities and $c_{il}=\sum_{l'=1}^{l} p_{il'}^*$. Each imputed value z_i^* is

$$z_i^* = 1 + \sum_{l=1}^{L-1} I(u_i > c_{il}),$$

where u_i is a random draw from a uniform distribution on (0, 1) and $I(u_i>c_{il})=1$ if $u_i>c_{il}$, 0 otherwise.

2.4. *Ordered categorical variables*

Ordered categorical variables z with $L>2$ classes may be modeled either using multinomial logistic regression or using the proportional odds model, which extends the binary logistic model by constraining the probabilities of class membership for z according to a proportional odds assumption between the ordered categories. Unlike the multinomial logistic model, this model has only one linear predictor, $\boldsymbol{\beta}\mathbf{x}$. The model is

$$\text{logit } \Pr(z\leqslant l|\mathbf{x};\boldsymbol{\beta},\boldsymbol{\zeta}) = \zeta_l - \boldsymbol{\beta}\mathbf{x}, \tag{6}$$

where $\zeta_0=-\infty<\zeta_1<\cdots<\zeta_L=\infty$ and $\boldsymbol{\zeta}=(\zeta_1,\ldots,\zeta_{L-1})$. $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ are estimated by maximum likelihood and values $\boldsymbol{\beta}^*$ and $\boldsymbol{\zeta}^*$ are drawn from a Normal approximation to their posterior distribution. The estimated probability of individual i belonging to class $l=1,\ldots,L$ is given by $p_{il}^* = \Pr(z_i\leqslant l|\mathbf{x}_i;\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*) - \Pr(z_i\leqslant l-1|\mathbf{x}_i;\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*)$. Imputation then proceeds as for nominal variables.

For integer-valued variables, an alternative is to impute using an overdispersed Poisson regression model.

2.5. *Categorical variables as predictors*

The models just discussed are good approaches to imputing missing values of categorical variables. When instead a categorical variable z is a predictor of another variable with missing data, it is usually included in the imputation model by a set of dummy variables. However, when z is ordered categorical, a second possibility is to include z as linear (possibly after a monotonic transformation), since ordinality is respected. The linearity approach may provide a robust solution when some levels of z are sparsely populated. In most cases, however, the use of dummy variables appears preferable [17].

3. A simple example: the UK700 trial

The UK700 trial was a multi-centre study conducted in four inner-city areas (here termed 'centres') [19]. Participants were aged 18–65 with a diagnosed psychotic illness and two or more psychiatric hospital admissions, the most recent within the previous 2 years. In the UK, such patients are typically managed in the community by a case manager. In the trial, 708 participants were randomly allocated to a case manager either with a case load of 30–35 patients (standard case management) or with a case load of 10–15 patients (intensive case management). The main trial findings have been previously reported [19].

Table I. UK700 data: description of 500 selected observations.

Variable	Description	Missing values	Range	Mean	SD
Baseline variables					
`centreid`	Study centre	0	0–3		
`afcarib`	Ethnicity (Afro-Caribbean/other)	0	1/0		
`ocfabth`	Father's occupational class at birth	87 (17 per cent)	1–6		
`cprs94`	Psychopathology score	0	0–79	19.4	12.9
`sat94`	Satisfaction with services score*	101 (20 per cent)	9–36	19.1	4.7
`rand`	Randomized group (standard/intensive case management)	0	0/1		
Outcome variables					
`cprs96`	Psychopathology score	79 (16 per cent)	0–71	18.6	13.6
`sat96`	Satisfaction with services score*	151 (30 per cent)	9–32	17.1	4.7

*Larger values indicate less satisfaction.

Table II. UK700 data: parameters for imputing `sat94` and `sat96` in the final cycle of the MICE algorithm.

	Parameters for imputing `sat94`				Parameters for imputing `sat96`			
Imputation	β^*_{sat96}	β^*_{rand}	$\beta^*_{\text{constant}}$	σ^*	β^*_{sat94}	β^*_{rand}	$\beta^*_{\text{constant}}$	σ^*
1	0.298	0.524	13.627	4.498	0.308	−0.458	11.476	4.514
2	0.284	0.514	13.985	4.522	0.224	−0.360	13.029	4.619
3	0.262	0.433	14.272	4.547	0.297	−0.275	11.653	4.534
4	0.296	0.486	13.720	4.518	0.310	−0.367	11.423	4.512
5	0.287	0.602	13.856	4.517	0.286	−0.426	11.852	4.550

In order to illustrate MI, we use a subsample of 500 individuals selected at random and summarized in Table I; these analyses are not intended to be definitive. The main aim of our analyses is to explore the effect of the intervention (intensive case management, variable `rand`) on satisfaction with services after two years of follow-up (`sat96`), allowing for baseline variables including levels of satisfaction at baseline (`sat94`). Both `sat96` and `sat94` are assumed to be Normally distributed.

Ignoring all other variables in the data set, the simplest imputation model for `sat96` is a linear regression on `sat94` and `rand`. Likewise, the simplest imputation model for `sat94` is a linear regression on `sat96` and `rand`. No imputation model is needed for `rand` because it is complete. (The alternative of running separate imputation procedures in different randomized groups is discussed later.) These imputation models assume that missingness in `sat94` (`sat96`) only depends on the observed values of `sat96` (`sat94`) and `rand`. The Stata command for running the combined imputation model with 5 imputations and loading the imputed data into memory is

```
. ice sat94 sat96 rand, m(5) clear
```

The command initializes the MICE procedure by replacing all missing values with randomly selected observed values. One cycle comprises imputing `sat94` from the linear imputation model that regresses `sat94` on `sat96` and `rand`, and then imputing `sat96` from the linear imputation model that regresses `sat96` on `sat94` and `rand`. The process is iterated through 10 cycles to produce a single imputed data set. In each cycle, variables are imputed in an increasing order of number of missing values. The procedure is repeated $m = 5$ times to produce 5 imputed data sets. The imputation parameters for the last cycle of each of the five imputations are shown in Table II, observed and imputed `sat94` and `sat96` values for a single imputation are shown in Figure 1, and the format of the imputed Stata data set is shown in Table III.

4. Handling skewed continuous variables

Non-Normally distributed continuous variables often occur. The drawback of imputing such variables by assuming Normality is that the distribution of imputed values does not resemble that of the observed values (lack of face validity). If, for instance, z is intrinsically positive and non-linearly related to the outcome but imputed by a linear regression on $\mathbf{x}$, the presence of non-positive imputed z values means that the correct model (linear regression of the outcome on $\ln z$) cannot be fitted to the MI data.

We discuss two main ways of dealing with non-Normality: transformation towards Normality and predictive mean matching. A different view, that non-Normality may not matter, is considered in Section 6.

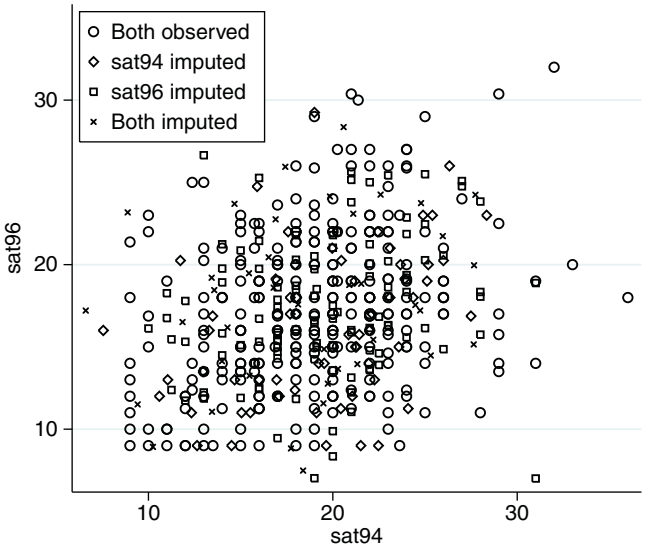

Figure 1. UK700 data: observed and imputed values for `sat94` and `sat96`.

Table III. UK700 data: data format before and after imputation, for 5 selected individuals and 2 imputed data sets. The imputed data set includes two added identifiers, `_mi` for individuals and `_mj` for imputed data sets, and includes the original data as `_mj==0`.

Before imputation			After imputation				
sat94	rand	sat96	_mi	_mj	sat94	rand	sat96
20	0	.	1	0	20	0	.
18	1	22	2	0	18	1	22
17	0	16	12	0	17	0	16
.	1	.	45	0	.	1	.
.	0	14	61	0	.	0	14
			1	1	20	0	8.35
			2	1	18	1	22
			12	1	17	0	16
			45	1	20.28	1	13.66
			61	1	19.53	0	14
			1	2	20	0	11.10
			2	2	18	1	22
			12	2	17	0	16
			45	2	24.91	1	16.81
			61	2	22.76	0	14

4.1. Transformation

When z is non-Normal, a simple monotonic transformation $f(\cdot)$ can often be found such that the marginal distribution of $f(z)$ is approximately Normal [15]. Note that ideally the *conditional* distribution of $f(z)$ given $\mathbf{x}$ would be Normal, rather than the marginal distribution of $f(z)$, but in practice this distinction may not matter much. Well-known one-parameter candidates for $f(\cdot)$ include the Box–Cox transformation $f(z)=(z^{\lambda}-1)/\lambda$ (with $f(z)\to\ln(z)$ as $\lambda\to 0$), and the shifted-log transformation $f(z)=\ln(\pm z-a)$, with a chosen such that $(\pm z-a)>0$; the sign of z is taken as positive when the distribution of z is positively (right) skewed and negative when z is negatively (left) skewed. The ancillary parameters λ and a may be estimated by maximum likelihood and confidence intervals found. As exact values are not critical, rounded values of λ or a may be chosen within their confidence intervals. In particular, if the confidence interval for λ or a includes 0, λ or a may be taken as 0 and a simple log transformation used.

The two transformations always remove skewness, but non-Normal tails (kurtosis) or other non-Normalities may remain. If so, two-parameter transformations to remove both skewness and kurtosis may be tried instead. These include the Johnson S_U family [20] and the modulus-exponential-Normal (MEN) and modulus-power-Normal (MPN) transformations [21].

To get imputed values of z, the imputed values of $f(z)$ must be back-transformed to the original scale, thus $f(\cdot)$ must be both monotonic and invertible.

Transformation to Normality is impossible with 'semi-continuous' variables for which a substantial proportion of values are equal (often zero). An example is weekly alcohol consumption. Such variables are discussed in Section 4.4.

4.2. *Predictive mean matching*

Predictive mean matching (PMM) is an *ad hoc* method of imputing missing values of z with the property that imputed values are sampled only from the observed values of z [22]. The consequence is that the distribution of imputed z often closely matches that of the observed z. Such a property is desirable, for example, when z is continuous and the Normality assumption is untenable, or when the relationship between z and $\mathbf{x}$ is non-linear. It is undesirable when imputation appropriately involves extrapolation beyond the range of the observed values of z, or possibly when the sample size is small (since then only a small range of imputed values is available).

To use PMM, the standard method described in Section 2.1 is used to give a perturbed parameter vector $\boldsymbol{\beta}^*$. For each missing value z_i with covariates $\mathbf{x}_i$, the standard procedure would next sample from a Normal distribution with mean $\boldsymbol{\beta}^*\mathbf{x}_i$. Instead, PMM identifies the q individuals with the smallest values of $|\widehat{\boldsymbol{\beta}}\mathbf{x}_h - \boldsymbol{\beta}^*\mathbf{x}_i|$ $(h=1,\ldots,n_{\text{obs}})$. Any ties are broken at random. One of these q closest individuals, say i', is chosen at random, and the imputed value of z_i is $z_{i'}$. Thus the imputed value is an observed value of z whose prediction with the observed data closely matches the perturbed prediction. We use $q=3$, which performed well in a simulation study, although a more complex adaptive method performed better [23].

4.2.1. Example. We consider PMM in a small simulated example in which $\mathbf{x}$ is univariate. Data pairs (z_i, x_i) for $i=1,\ldots,30$ were simulated by drawing z from a uniform distribution on $[0.05, 1.05]$ and x from the non-linear model $x=\ln z+\text{N}(0,\sigma^2)$ where $\sigma=0.16$; the population coefficient of determination is $R^2=0.95$. Ten values of z were deleted completely at random and were imputed using the standard approach and using PMM. The imputation model for $z|x$ was assumed linear in x, which is mis-specified, since the true relationship is approximately exponential in x.

The upper left panel of Figure 2 shows the result of one imputation with the standard approach, assuming that $z|x$ is linear in x and Normally distributed. The imputed values are clearly biased at low x where the relationship is least linear. However, PMM imputes missing z values appropriately (upper right panel of Figure 2).

Repeating the experiment with PMM and $m=10$ imputations, and then estimating β_1 in the model $E(x)=\beta_0+\beta_1\ln z$ by Rubin's rules, we find that $\widehat{\beta}_1=0.98$ (SE 0.05), which is consistent with the true value of $\beta_1=1$. With the standard approach, the correct model cannot be estimated because of the negative imputed values of z.

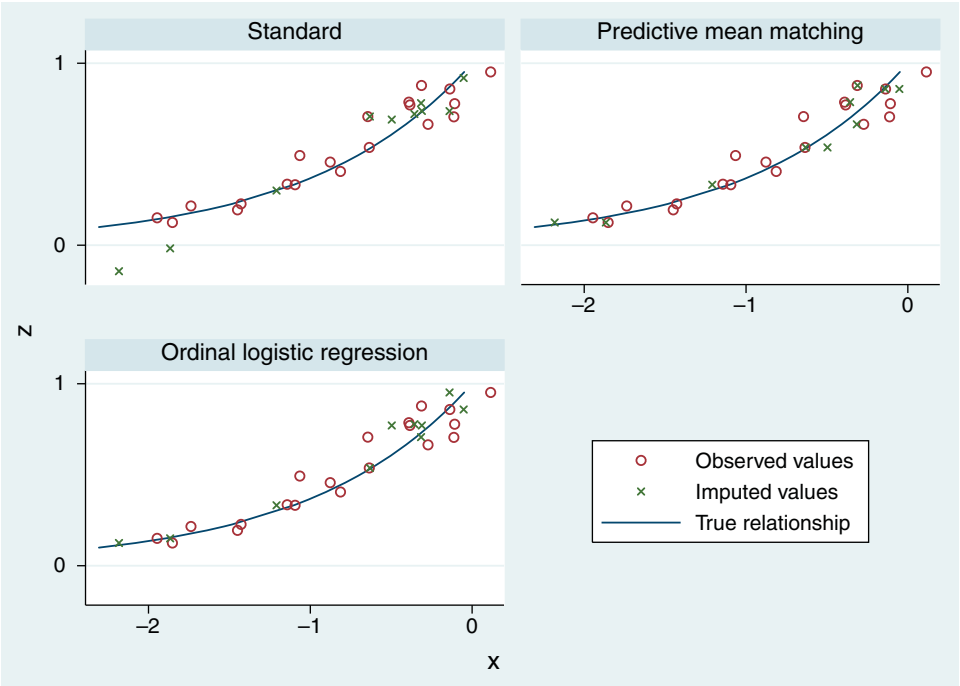

Figure 2. Artificial example of PMM with a non-linear relationship between z and x. Ten values of z missing completely at random are imputed from 20 observed values of x.

4.3. Ordinal logistic regression

An alternative approach is to treat the continuous variable z to be imputed as ordinal, and model it using the proportional odds assumption of Section 2.4. As with PMM, the imputed values of z come only from the observed distribution of z. A limitation with some software is that the maximum number of levels allowed may be limited. In Stata version 10, for example, the `ologit` command permits up to 50 levels. However, rounding the values of z appropriately is not a serious restriction.

4.3.1. Example. Continuing the example of Section 4.2.1, we apply the ordinal logistic regression approach. The results in one imputation are shown in the lower panel of Figure 2. The fit of the imputed values appears good. The estimate $\widehat{\beta}_1$ is 0.97 (SE 0.06), similar to the result using PMM.

4.4. Semi-continuous data

A variable is semi-continuous if a substantial fraction of values are equal (usually zero). Transformation does not yield a Normally distributed variable, but PMM or ordinal logistic regression may be used. An alternative approach involves separately imputing a binary variable indicating whether a value is zero or positive, and then a continuous variable for the value if positive. This approach has been applied to cost data [24]. Care is required to transform the continuous part of the distribution into Normality.

5. Specifying the imputation model: variable selection

The previous sections have focussed on drawing imputations that correctly represent what the unobserved data might have been. However, the ultimate aim in MI is to draw valid and efficient inferences by fitting models ('analysis models') to multiply imputed data. We now discuss two key requirements to avoid bias and gain precision when selecting variables for the imputation model. Section 6 discusses further requirements on the form of the imputation model.

5.1. Include the covariates and outcome from the analysis model

To avoid bias in the analysis model, the imputation model must include all variables that are in the analysis model [7]. In particular, when imputing missing values of analysis model covariates, the analysis model outcome must be in the imputation model [25]. If the imputed data are to be used to fit several different analysis models, then the imputation model should contain every variable included in any of the analysis models.

Consider, for example, the analysis of the UK700 data in Section 3, where the analysis model is a linear regression of `sat96` on `sat94` and `rand`. Suppose that the imputation model for `sat94` is chosen as a linear regression on `rand` only, wrongly omitting `sat96`. This makes `sat94` and `sat96` uncorrelated in individuals with imputed `sat94` and observed `sat96`, biasing the coefficient of `sat94` in the analysis model towards zero. The solution is simple: add `sat96` to the imputation model for `sat94`. In the UK700 data, the coefficient of `sat94` is 0.248 (standard error 0.057) when `sat96` is excluded from the imputation model for `sat94`, rather smaller than the value of 0.292 (standard error 0.055) when it is included.

When the analysis is based on a survival model, the outcome comprises time t and the censoring indicator d. Possible approaches to including the outcome when imputing covariates are including t, $\log t$ and d [26], d and $\log t$ [27] or d and t [28] in the imputation model. We have recently explored this issue in the case where the analysis model is a proportional hazards model [29]. We found that using d and $\log(t)$ can bias associations toward the null. The correct imputation model in the case of a single binary covariate involves d and $H_0(t)$, where $H_0(.)$ is the cumulative baseline hazard function; this model is approximately correct in more complex settings. In general, $H_0(.)$ is not known, but it may be adequately approximated by $H(.)$, the standard (Nelson–Aalen) estimate of the cumulative hazard function. This procedure may be implemented in Stata using `sts generate NA = na` and then including the variables `NA` and the censoring indicator `_d` in the `ice` call.

5.2. Include predictors of the incomplete variable

Including predictors of the incomplete variable in the imputation model is beneficial for two reasons.

First, it may make the MAR assumption more plausible and hence reduce bias. The MAR condition for valid inferences is that the probability of data being missing does not depend on the unobserved data, conditional on *the observed data that are included in the imputation model*. Thus the imputation model should include every variable that *both* predicts the incomplete variable *and* predicts whether the incomplete variable is missing. For example, in the UK700 data,

centreid predicts both the outcome sat96 and whether sat96 is missing, hence centreid should be included in the imputation model for sat96.

Second, adding predictors of the incomplete variable to the imputation model can improve the imputations and hence reduces the standard errors of the estimates for the analysis model.

For example, consider estimating the intervention effect on outcome sat96 in the UK700 trial. It is well known that post-randomization variables in clinical trials (treatment adherence variables, potential mediating variables and other outcome variables) should not be entered in the analysis model when estimating treatment effects. However, such 'auxiliary' variables are potentially useful in the imputation model. Some individuals with missing values of sat96 have observed values of other trial outcomes such as cprs96. We may therefore get better imputations for sat96 if we add cprs96 to the imputation model. This may also make the MAR assumption more plausible.

In practice, one could include in the imputation model all variables that *significantly* predict the incomplete variable, or whose association with the incomplete variable exceeds some threshold. One might also include any other variables which significantly predict whether the incomplete variable is missing, on the grounds that bias may be avoided if the included variables have a true association with the incomplete variable that fails to reach statistical significance, whereas little loss of precision is likely if the included variables do not predict the incomplete variable. Variable selection could be based on univariate or multivariate associations in complete cases [26]; an alternative might be to use a provisional imputation scheme with a small number of imputations (even just one) and apply a model selection procedure.

Difficulties of MI when the imputation model contains more variables than the analysis model have been much debated [30–32]. The focus in this debate has been on theoretical deficiencies of the Rubin's rules variance; there is no suggestion that bias is incurred. A pragmatic conclusion is that such deficiencies are not of practical importance [7, Section 4.5.4].

6. Specifying the imputation model: model form

As well as including all variables in the analysis model, the imputation model must also include them in an appropriate way: in the correct functional form and with any interactions that are required. We explore this problem by considering whether the association between baseline and 2-year satisfaction in the UK700 data is linear or curved. We use a quadratic analysis model of sat96 on sat94 and its square, sat94sq (although a more appropriate choice might be a fractional polynomial model [33]). We focus on individuals with the observed sat96, for reasons explained in Section 8.1, and we consider how to impute the missing values of sat94 and sat94sq.

6.1. The 'passive' approach

A first approach imputes missing values of sat94 using a linear regression on sat96, and then imputes missing values of sat94sq 'passively' [9], that is, by squaring the imputed values of sat94. In the comparison below, we call this a 'linear passive' model.

Unfortunately, if the analysis model truly involves a non-zero quadratic term, then the model for sat94 on sat96 is not a simple linear regression. Data imputed using linear regression imputation would show insufficient curvature, thus the coefficient of the quadratic term in the analysis model would be biased towards 0.

It is not easy to solve this problem, because the true imputation model is of no standard form. One possibility is to use PMM to reduce the impact of model mis-specification. In the comparison below, we call this an 'improved passive' model.

6.2. Congenial imputation model

Instead of aiming to find the true imputation model, an alternative approach relies on finding an imputation model that is 'congenial' to the analysis model but not necessarily correctly specified. Congeniality means that the imputation and analysis models are both compatible with some larger model for the data. Congeniality between the imputation procedure and the analysis model ensures (for large m) that inference on MI data approximates a maximum likelihood procedure [31].

A popular choice of the larger model is the multivariate Normal distribution. When the analysis model is a linear regression, a congenial imputation model is therefore a linear regression of the incomplete variable on all other variables from the analysis model. The multivariate Normal is rarely a good model for data: for example, it is mis-specified when some variables are categorical. However, Schafer presents evidence that procedures based on a multivariate Normal assumption perform well under this sort of model mis-specification [7].

The multivariate Normal model has a surprising application in fitting the quadratic model introduced in Section 6.1. We ignore the fact that sat94sq is defined as the square of sat94, and impute under a multivariate Normal model for sat94, sat94sq and sat96: that is, the imputation model for sat94 is a linear regression on sat94sq and

Table IV. UK700 data: estimated quadratic coefficients (×100) from fitting a quadratic regression of `sat96` on `sat94` using different imputation procedures.

Imputation method	Coefficient	Standard error
Linear passive	−1.05	0.77
Improved passive	−1.36	0.77
Just another variable	−1.34	0.80

`sat96`, and the imputation model for `sat94sq` is a linear regression on `sat94` and `sat96`. We call this approach 'just another variable' (JAV), because `sat94sq` is regarded as just another variable whose deterministic relationship with `sat94` is ignored. The JAV approach makes a very bad approximation to the joint density of `sat94` and `sat94sq`, yet it can yield valid inferences for the analysis model [34].

6.3. Method comparison

We now compare the three approaches: 'linear passive' which is mis-specified and far from congenial, 'improved passive' which may be better specified and closer to being congenial, and JAV that is highly mis-specified but congenial. We use $m = 100$ imputations to get sufficient accuracy to compare methods (see Section 7). Stata code is given in Appendix A1.

The results (Table IV) show that the linear passive approach seems to underestimate the quadratic coefficient, but the improved passive approach using PMM gives results comparable with those from JAV. (Monte Carlo errors, expressing uncertainty due to the stochastic nature of the analysis, are about 0.03 for the coefficients and 0.01 for the standard errors). In this example, the substantive interpretation is the same with all methods: there is no evidence of curvature. In other examples, the linear passive method could fail to detect curvature that is significant by the less biased methods.

6.4. Interactions

Imputing covariates whose interactions appear in the analysis model also requires care. We explore an interaction between randomized group `rand` and a dummy variable, `manual`, for the father being of manual social class at the patient's birth. If we had complete data then the analysis would involve a linear regression of `sat96` on `sat94`, `rand`, `manual` and the product `rand*manual`.

A linear passive approach imputes the three incomplete variables in the usual way: linear regression of `sat96` on `sat94`, `rand` and `manual`; linear regression of `sat94` on `sat96`, `rand` and `manual`; logistic regression of `manual` on `sat94`, `sat96` and `rand`. It then imputes the `rand*manual` interaction passively.

We can improve this passive imputation model. The interaction between `rand` and `manual` in the analysis model means that the association between `sat96` and `manual` may differ between randomized groups. Congeniality requires this interaction to be allowed for in the imputation model, so when we impute `manual`, we must allow for an interaction between `rand` and `sat96`: this gives an improved passive model. The linear passive approach, which ignores this interaction in the imputation model, is likely to underestimate the interaction in the analysis model.

A simpler congenial imputation approach involves separate imputation within each randomized group. This is convenient because randomized group is complete.

The JAV approach is based on a multivariate normal model for the four variables jointly with the product `rand*manual`, which requires that each variable is imputed using a linear regression. In particular, `rand*manual` is imputed using a linear regression on `sat96`, `sat94`, `rand` and `manual`, and `manual` is imputed using a linear regression on `sat96`, `sat94`, `rand` and `rand*manual`.

For analyses, we used $m = 500$ imputations because method comparisons with $m = 100$ were inconclusive. Stata code is given in Appendix A2. We take imputation separately by randomized group as the gold standard. The results (Table V) show that the linear passive approach suffers bias towards zero, and this bias is reduced but not completely removed by the improved passive approach. (Monte Carlo errors are about 0.02 for the coefficients and 0.01 for the standard errors.) Differences between JAV and by-group imputation are compatible with Monte Carlo error.

6.5. Recommendations

We have illustrated the passive and JAV approaches to imputation in non-linear models. Neither is without problems.

The simplest approach is passive imputation using simple linear and logistic regression models and ignoring interactions (and other subtleties) that are in the analysis model. The cost is bias of relevant terms in the analysis model (typically, but not always, towards zero) and a loss of power to detect non-linearities and interactions. Improving the passive approach relies on correct specification of the imputation models, which is hard to do; as a result, some bias is possible, as in Table V. As the number of variables increases, it becomes harder to find and estimate correct passive imputation models.

Table V. UK700 data: estimated interaction coefficients from fitting a linear regression of `sat96` on `sat94`, `rand`, `manual` and the interaction of `rand` and `manual`, using different imputation procedures.

Imputation procedure	Estimate	Standard error
Separately by randomized group	−1.77	1.15
Linear passive	−1.46	1.15
Improved passive	−1.67	1.17
Just another variable	−1.80	1.14

The JAV approach, on the other hand, produces implausible imputed values for derived variables, which can undermine the credibility of the procedure. Furthermore, congenial models such as the multivariate normal are usually strongly mis-specified. Proof that the JAV procedure is unbiased [34] relies on the MCAR assumption. Our simulation studies suggest that the model mis-specification inherent in the JAV approach may lead to bias when data are MAR but not MCAR (see Appendix B).

Thus it is hard to give concrete advice about the choice of imputation model. The best choice is to find an imputation model that is both congenial and a good representation of the data. If this cannot be done, it may be worth experimenting with more than one imputation model and exploring the impact of different choices on the analysis results. If several analyses are to be run on one imputed data set, then the imputation model should be congenial with all analyses. An awkward problem is checking for omitted non-linear and interaction terms in the analysis model: we suggest a possible strategy for this in Section 8.3.3.

7. How many imputations?

So far we have not explained how we choose the number of imputations. Standard texts on multiple imputation suggest that small numbers of imputed data sets ($m=3$ or 5) are adequate. We first review this argument, then explain why we usually prefer larger values of m, and suggest a rule of thumb.

7.1. *Efficiency argument*

The standard argument is based on the statistical efficiency of the estimates [7]. The (true) variance of a parameter estimate based on m imputations is $W+(1+1/m)B$, where as before W is the average within-imputation variance and B is the between-imputations variance. Therefore the relative efficiency of infinitely many imputations compared to m imputations is $\{W+(1+1/m)B\}/(W+B)=1+\mathrm{FMI}/m$ where $\mathrm{FMI}=B/(W+B)$ is the fraction of missing information. If we can accept 5 per cent loss of efficiency, then we need $\mathrm{FMI}/m\leqslant 0.05$, hence $m=5$ is adequate if $\mathrm{FMI}\leqslant 0.25$.

Graham *et al.* argued that we should instead choose the number of imputations to limit the loss in power for testing an association of interest [35]. To limit the loss in power to no more than 1 per cent, they needed $m\geqslant 20$ when the FMI was between 0.1 and 0.3.

7.2. *Reproducibility argument*

We believe that statistical efficiency and power are not enough. Instead, as data analysts, we want to be confident that a repeat analysis of the same data would produce essentially the same results [13, 36]. This means that we must consider the Monte Carlo error of our results, defined as their standard deviation across repeated runs of the same imputation procedure with the same data. Clearly, Monte Carlo error tends to zero as m increases.

Monte Carlo errors apply not just to parameter estimates but also to all other statistics computed using multiply imputed data, including standard errors, P-values and confidence intervals. For a 1-dimensional parameter estimate, the Monte Carlo error is easily calculated as $\sqrt{B/m}$. Monte Carlo errors for other statistics maybe computed using a jacknife procedure, implemented in Stata by `mim, mcerror` [37].

As an example, we consider the UK700 baseline data, using as analysis model a linear regression of `cprs94` on father's social class, `ocfabth`, and ethnicity, `afcarib`, and focussing on the coefficient of `ocfabth`. The FMI is about 0.17, hence MI with $m=5$ loses about $\mathrm{FMI}/m=3$ per cent precision: conventionally, this would be adequate. Yet five different imputation runs with $m=5$ give substantially different results, with P-values ranging from 0.03 to 0.08 (Table VI). Although the difference between $P<0.05$ and $P>0.05$ should not be overstated, many data analysts would not be happy with this degree of variability. Monte Carlo errors reported by `mim, mcerror` range across the five runs from 0.09 to 0.13 for the coefficient, from 0.03 to 0.28 for the lower confidence limit, from 0.11 to 0.35 for the upper confidence limit, and from 0.01 to 0.05 for the P-value.

Table VI. UK700 data: estimated coefficients of `ocfabth` from fitting a linear regression of `cprs94` on `ocfabth` and `afcarib`, in five different runs with $m=5$.

Run	Coefficient	95 per cent CI	P-value
1	−1.24	−2.45 to −0.03	0.05
2	−1.26	−2.37 to −0.16	0.03
3	−0.99	−2.10 to 0.11	0.08
4	−1.29	−2.54 to −0.03	0.05
5	−1.15	−2.40 to 0.10	0.07

7.3. A rule of thumb

Bodner performed a simulation study to explore the Monte Carlo variability of three quantities: the width of the 95 per cent confidence interval, the P-value, and the estimated FMI [38]. He chose as a key criterion that the width of the 95 per cent confidence interval should be within 10 per cent of its true value in 95 per cent of imputation runs. This led to the requirement $m \geqslant 3$, 6, 12, 24, 59 for FMI = 0.05, 0.1, 0.2, 0.3, 0.5, respectively. Bodner proposed conservatively estimating the FMI as the fraction of incomplete cases. This work has been summarized by 'the number of imputations should be similar to the percentage of cases that are incomplete' [34], at least with FMI $\leqslant 0.5$.

We prefer the following simpler argument that leads to the same rule of thumb, using results in Appendix C that give simple approximations to Monte Carlo errors. We start by considering a particular parameter β. If m is chosen so that $\text{FMI}/m \approx 0.01$, where FMI is the fraction of missing information for β, we get the following properties:

(1) The Monte Carlo error of $\widehat{\beta}$ is approximately 10 per cent of its standard error.
(2) The Monte Carlo error of the test statistic $\widehat{\beta}/\text{se}(\widehat{\beta})$ is approximately 0.1.
(3) The Monte Carlo error of the P-value is approximately 0.01 when the true P-value is 0.05, and 0.02 when the true P-value is 0.1.

We suggest that these often provide an adequate level of reproducibility in practice. Thus we require about $m \geqslant 100\times$ FMI. However, the FMI is unknown and hard to estimate (estimates of the FMI themselves have large Monte Carlo error). In addition, in any data set, the FMI is different for different parameters. In order to derive a rule of thumb, we use Bodner's approximation that the FMI for any parameter is likely to be less than the fraction of incomplete cases. This therefore suggests the rule of thumb that m should be at least equal to the percentage of incomplete cases, as stated above. For the UK700 analysis in Table VI, 17 per cent of cases are incomplete, hence this rule would suggest $m=20$.

It should be clear that this rule is not universally appropriate. Some non-MCAR missing data patterns have FMI greater than the fraction of missing data. Some settings require a greater or smaller degree of reproducibility. Larger numbers of imputations may be required for method comparison studies [39]: indeed, as already stated, we used $m=100$ or $m=500$ for the analyses in Section 6. Finally, it may be convenient to impute a larger number of data sets but use only some of them in most analyses.

8. Analysis of multiply imputed data

So far we have mainly discussed approaches for generating multiply imputed data sets. In this section, we focus on analyzing the imputed data sets using standard statistical procedures.

8.1. Individuals with missing outcomes

It is argued that individuals with imputed outcomes should be excluded from the analysis, because including them only adds noise to the estimates [40, 41]. Table VII compares linear regression analyses of `sat96` on `sat94` and `rand` using (i) data from the complete cases, (ii) multiply imputed data on all individuals, and (iii) multiply imputed data restricted to individuals with observed `sat96`. Comparing analyses (ii) and (iii) with $m=100$, estimated coefficients were similar, but Monte Carlo errors were substantially smaller for analysis (iii), and estimated standard errors were slightly smaller; results with $m=5$ were rather unstable. Method (iii) was in fact used for the analyses in Sections 5 and 6.

Method (ii) might be valuable when auxiliary variables have been used in the imputation model, because there is extra information in the imputed outcomes. In practice, however, it is only worth using method (ii) if auxiliary variables are highly correlated with the outcome variable (for standard error reduction), or if they are correlated with the outcome variable and with the probability that the variable is missing (for bias reduction).

Table VII. UK700 data: estimated coefficients and standard errors from fitting a linear regression of `sat96` on `sat94` and `rand`. Monte Carlo errors are given in square brackets.

Analysis	n	$\hat{\beta}_{\text{sat94}}$	$se(\hat{\beta}_{\text{sat94}})$	$\hat{\beta}_{\text{rand}}$	$se(\hat{\beta}_{\text{rand}})$
(i) Complete cases	288	0.297	0.056	−0.201	0.539
(ii) Multiple imputation					
$m=5$	500	0.269 [0.0200]	0.065 [0.0169]	−0.320 [0.0802]	0.450 [0.0324]
$m=100$	500	0.285 [0.0158]	0.065 [0.0117]	−0.377 [0.0316]	0.497 [0.0038]
(iii) Multiple imputation restricted to individuals with observed `sat96`					
$m=5$	349	0.296 [0.0038]	0.057 [0.0018]	−0.442 [0.0283]	0.496 [0.0111]
$m=100$	349	0.295 [0.0024]	0.056 [0.0008]	−0.377 [0.0070]	0.494 [0.0007]

Table VIII. Common statistics that can and cannot be combined using Rubin's rules (equations (1) and (2)).

Statistics that can be combined without any transformation	Mean, proportion, regression coefficient, linear predictor, C-index, area under the ROC curve
Statistics that may require sensible transformation before combination	Odds ratio, hazard ratio, baseline hazard, survival probability, standard deviation, correlation, proportion of variance explained, skewness, kurtosis
Statistics that cannot be combined	P-value, likelihood ratio test statistic, model chi-squared statistic, goodness-of-fit test statistic

8.2. Combining estimates using Rubin's rules

So far we have concentrated on regression coefficients, but interest sometimes focusses on other statistics that have been computed in each imputed data set. Statistics that are estimators of some quantity can validly be combined using Rubin's rules (equations (1) and (2)), although some transformation may be required to ensure that they are approximately Normally distributed [42, p. 108]. Statistics that are not estimators, such as measures of strength of evidence, cannot be combined using Rubin's rules. Put crudely, statistics whose value changes systematically with the sample size cannot be combined using Rubin's rules. Table VIII summarizes some common statistics which can and cannot be combined using Rubin's rules.

8.3. Model building

In many studies, analysis involves techniques of model building and model criticism such as variable selection, checking residuals, and testing for interactions and non-linear relationships. Usual practice is to perform such procedures among the complete cases. We advise against this practice because of a lack of power (for example, important predictors may be undetected) and potential bias (for example, when missing data are not MCAR, unimportant variables may be selected due to biased regression estimates) [43]. We provide guidance below on how to address model building issues in multiply imputed data sets.

8.3.1. Hypothesis testing. All model building procedures require hypothesis testing. The Wald statistics for testing the univariate or multivariate null hypothesis $\boldsymbol{\theta}=\mathbf{0}$ were discussed in the introduction. Likelihood ratio test statistics can be combined using an approximation proposed by Meng and Rubin [44]: this procedure may be convenient when the number of parameters to be tested is large, but it has not been shown to be superior to the Wald test. Thus, although likelihood ratio tests are often preferable with complete data, Wald tests should usually be used for hypothesis tests with multiply imputed data. In Stata, a set of regression parameters can be tested using `mim: testparm`.

8.3.2. Variable selection. Classical variable selection is usually implemented through iterative procedures such as forward, backwards and stepwise selection, and modification is required for application to multiply imputed data. Each variable selection step involves fitting the model under consideration to all imputed data sets (MI Stage 2) and combining estimates across imputed data sets (MI Stage 3). This multi-stage iterative process has type 1 error comparable to what would be achieved if there were no missing data [43]. However, it may be impractical under some circumstances such as (i) large data sets, (ii) large m, (iii) when multiple outcomes are of interest or (iv) when numerous variables or possible interaction terms are to be assessed. A pragmatic alternative is to analyze the multiply imputed data sets as a single data set of length $m \times n$, and to perform the variable selection procedure on this one data set. When assessing inclusion or exclusion of a covariate x, each observation receives a weight of $(1-f_x)/m$ where f_x is the fraction of missing data

for x [43]. This approach has been shown to be a good approximation. It may also be helpful in selecting functional forms for continuous predictors, for example using multivariable fractional polynomials [33].

8.3.3. Selection of non-linear and interaction terms. Selection of non-linear and interaction terms presents further difficulties. For brevity, we use 'non-linear terms' to include interaction terms. As noted in Section 6, non-linear terms can only be correctly assessed when they have been allowed for in the imputation model. However, when imputing data, one does not necessarily know what non-linear terms will be required in a sequence of analyses, and allowing for all possible non-linear terms might make the imputation model impractically large. One possible strategy is

(1) Produce a provisional and relatively simple imputation model, including non-linear terms of key scientific interest, but omitting all other non-linear terms.
(2) Use the imputed data to build and check an analysis model, including investigating the need for non-linear terms. Note that these model checks are conservative when relevant non-linear terms were omitted from the imputation model.
(3) If any convincing non-linear terms are found, then recreate the imputations including the non-linear terms computed with 'JAV' or a careful passive approach.
(4) Use the revised imputed data set to estimate the parameters of the final analysis model.

An alternative [43] would be to repeat the model building (step 2) if new imputations are drawn at step 3.

8.4. Obtaining predictions

In some circumstances we need to draw individual predictions from the analysis model. If the predictions are an intermediate step in a larger procedure—for example, in model checking, as in Section 8.5 below—they should be computed within imputed data sets and not combined over data sets. For clinical predictions, however—when we want to report probabilities of future outcomes in individuals with complete or incomplete covariates—it is appropriate to combine predictions over data sets. Let $\widehat{\eta}_{ij}$ be the imputation-specific linear predictor of interest for the ith individual obtained from the jth imputed data set. As this estimates a parameter η_i, Rubin's rules can be applied, and the combined linear predictor for individual i is $\widehat{\eta}_i = (1/m)\sum_{j=1}^{m} \widehat{\eta}_{ij}$. In logistic regression, we may want to estimate the probabilities $\pi_i = \text{expit}(\eta_i)$, where expit(.) denotes the inverse of the logit function. We could apply Rubin's rules on the probability scale, giving $\widehat{\pi}_i = (1/m)\sum_{j=1}^{m} \text{expit}(\widehat{\eta}_{ij})$ or we could use $\widehat{\pi}_i = \text{expit}(\widehat{\eta}_i)$: these are usually very similar. In Stata, imputation-specific predictions can be generated using `mim: predict`.

8.5. Model checking

A full data analysis should include model checking procedures. Such procedures could be performed on each imputed data set. This is illustrated in Figure 3 by the set of residual plots for the UK700 analysis presented in Table VII. Such plots can help to identify problems either with the imputation model or with the analysis model. For example, extreme outliers occurring in a small number of imputed data sets suggest a problem with the imputation model, whereas problems consistently occurring across all imputed data sets suggest a problem with the analysis model. Figure 3 shows slightly patterned residuals because of the bounded nature of the outcome. However, the pattern is similar across imputed data sets, not indicating any problem with the imputation model.

9. Illustrative analysis of the UK700 data

In this section we illustrate a possible analysis of the UK700 data in Stata. Our aim is to estimate the intervention effect on satisfaction with services, with adjustment for some baseline variables, and using other trial outcomes as auxiliary variables. Some output is omitted without comment.

If the data were complete then our analysis model would be

```
. regress sat96 rand sat94 Icentre* cprs94
```

where `Icentre*` are dummy variables for centres 1–3 (the reference category being centre 0). However, first we must impute the missing values. We use `cprs96` as an auxiliary variable to improve our imputations, because it is observed for 73 of the 151 individuals with missing `sat96`. Recall from Section 3 that `sat94`, `sat96` and `cprs96` are incomplete while `rand`, `Icentre*` and `cprs94` are complete.

In the previous sections we have imputed `sat94` in the same way as other variables, ignoring the fact that it is a baseline variable in a randomized trial. However, general statistical methods for missing data are not always appropriate for baseline variables in randomized trials, because they may not respect the independence of baseline variables from randomization. It is best to impute missing baselines deterministically using only other baseline data [45]. We do this

Figure 3. UK700 data: plot of residuals against fitted values for analysis (ii) of Table VII.

by imputing centre-specific means:

```
. egen sat94mean = mean(sat94), by(centreid)
. replace sat94 = sat94mean if missing(sat94)
```

We next impute the missing outcomes. The fraction of incomplete cases is 152/500≈30 per cent, hence we use $m=30$ (see Section 7). Because interactions between randomized group and baseline covariates are often of interest in randomized trials, we impute separately in each randomized group using `ice`'s `by(rand)` option. We set the seed so that our results are reproducible. We use the `genmiss(M)` option to generate indicators of missingness: for example, `Msat96` is 0 (1) for observations with observed (imputed) `sat96`.

```
. ice Icentre* sat94 sat96 cprs94 cprs96, by(rand) m(30) seed(1) clear genmiss(M)

   #missing |
      values|      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         348       69.60       69.60
          1 |          74       14.80       84.40
          2 |          78       15.60      100.00
------------+-----------------------------------
      Total |         500      100.00

   Variable | Command| Prediction equation
------------+--------+-------------------------------------------------------------
   Icentre1 |        | [No missing data in estimation sample]
   Icentre2 |        | [No missing data in estimation sample]
   Icentre3 |        | [No missing data in estimation sample]
     sat94  |        | [No missing data in estimation sample]
     sat96  |regress | Icentre1 Icentre2 Icentre3 sat94 cprs94 cprs96
    cprs94  |        | [No missing data in estimation sample]
    cprs96  |regress | Icentre1 Icentre2 Icentre3 sat94 sat96 cprs94
------------------------------------------------------------------------------

Imputing .........1.........2....[output omitted]....30
[note: imputed dataset now loaded in memory]
```

This output first shows the numbers of missing values, and then the models used to impute the two incomplete variables. The imputed data set that is now loaded in memory contains both the original data (identified by `_mj==0`) and all the imputed data sets (identified by `_mj==1`, `_mj==2`, etc).

We compare the observed and imputed data (Figure 4) using box plots over the imputations:

```
. graph box sat96 if _mj==0 | Msat96==1, over(_mj)
```

Gross discrepancies between the distributions of observed and imputed data would suggest errors in the imputation procedure, although some differences are to be expected if the data are not MCAR [46]. Figure 4 shows that the distribution of the imputed data is broadly similar to that of the observed data, but a few imputed values of `sat96` lie below the permitted range of 9–36. We could avoid this by using a transformation such as $\log\{\texttt{sat96}/(45-\texttt{sat96})\}$. A graph like Figure 1 might also be useful here, and similar checks should be done for imputed values of `cprs96`.

We fit the analysis model to the imputations displayed in Figure 4, including observations with missing outcome to benefit from the inclusion of `cprs96` in the imputation model (see Section 8.1).

```
. mim: regress sat96 rand sat94 Icentre* cprs94

Multiple-imputation estimates (regress)          Imputations =     30
Linear regression                                Minimum obs =    500
                                                 Minimum dof =  121.2

------------------------------------------------------------------------------
       sat96 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Int.]         FMI
-------------+----------------------------------------------------------------
        rand |   -.38696    .496623    -0.78   0.437    -1.36852   .594601     0.339
       sat94 |   .250918    .056591     4.43   0.000      .139408  .362428     0.229
    Icentre1 |  -1.02357    .721603    -1.42   0.158    -2.44852   .401368     0.310
    Icentre2 |   .421178    .687126     0.61   0.541    -.937743    1.7801     0.357
    Icentre3 |   .487475    .752643     0.65   0.518     -1.00255    1.977     0.386
      cprs94 |   .043834      .0199     2.20   0.029      .004528   .08314     0.318
       _cons |   11.8665    1.13555    10.45   0.000      9.62614  14.1068     0.279
------------------------------------------------------------------------------
```

The intervention improves user satisfaction with services by 0.4 units (95 per cent CI from −0.6 to +1.4 units). As the outcome standard deviation is 4.7 (Table IV), this would appear to exclude any clinically important effect.

Finally, we check that the Monte Carlo error is acceptable.

```
. mim, mcerror

Multiple-imputation estimates (regress)          Imputations =     30
Linear regression                                Minimum obs =    500
                                                 Minimum dof =  121.2

[Values displayed beneath estimates are Monte Carlo jackknife standard errors]
------------------------------------------------------------------------------
       sat96 |      Coef.   Std. Err      t    P>|t|     [95% Conf. Int.]         FMI
-------------+----------------------------------------------------------------
        rand |   -.38696    .496623    -0.78   0.437    -1.36852   .594601     0.339
             |    .05154    .020382     0.11    .0662     .062639   .07194     0.056
[output omitted]
------------------------------------------------------------------------------
```

The Monte Carlo errors are reasonably small: those for the coefficient and the standard error are less than 10 per cent of the estimated standard error, as proposed in Section 7. In particular, it is clear that the result is almost sure to be non-significant in any repeat MI run.

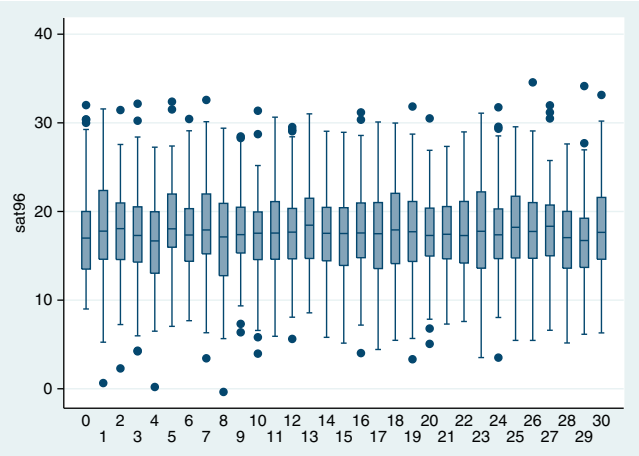

Figure 4. UK700 data: comparison of observed (0) and imputed (1–30) data for `sat96`.

10. Limitations and pitfalls in MICE

In this section we consider the main methodological limitation of the MICE procedure, that it lacks a clear theoretical rationale, and in particular that the conditional regression models may be incompatible. We then discuss several pitfalls in the practical use of MICE.

10.1. Lack of theoretical basis

Recall that MICE repeatedly samples from conditional distributions: for example, with three variables X, Y, Z, it draws X from the conditional posterior $[X|Y, Z]$, then Y from the conditional posterior $[Y|X, Z]$, and so on. This is somewhat similar to a Monte Carlo Markov Chain procedure [47], but its properties are not proven in general (e.g. [48]). Thus, justification of the MICE procedure has rested on empirical studies [10, 49] rather than theoretical arguments.

A particular concern arises when the specified conditional distributions are incompatible. A simple example arises in the UK700 data with the ordered categorical variable `ocfabth` and the continuous variable `sat94`, which are both incomplete. We might choose to impute `sat94` from a linear regression on `ocfabth` (with a linear trend across levels of `ocfabth`), and `ocfabth` from an ordered logistic regression on `sat94`. These models are incompatible: that is, there is no joint distribution for the two variables which yields both these conditional distributions. (This is a different issue from the uncongeniality discussed in Section 6, which referred to a mismatch between the imputation model and the analysis model: here the mismatch is between the imputation models for different variables.) One consequence of incompatible conditional regressions is that the distribution of imputed values, and hence the results of an analysis, may depend on the order of imputations and on which variable was imputed last. There is little evidence that this matters in practice [49]. It is probably more important to focus on the modeling issues raised in Section 6.

10.2. An example of a mis-specified imputation model

We now turn to some pitfalls of the MICE procedure, as discussed by Sterne *et al.* [46]. Possibly the easiest mistake to make in MICE is to omit the outcome from the imputation model. We saw in Section 5.1 that this biases the coefficients of incomplete covariates in regression models towards zero. Residual confounding may then lead other coefficients to be biased towards or away from zero.

For example, the authors of the QRISK study developed a new model to predict risk of cardiovascular disease using general practice data on 1.28 million individuals [50]. Here we focus on the results for women. One of the predictor variables was the ratio of total serum cholesterol to high-density lipoprotein cholesterol levels, with mean 4.0 and standard deviation 1.3. The two cholesterol variables were missing in some 60 and 70 per cent of individuals, respectively, and multiple imputation was used to impute these and other missing values. The published analysis showed a hazard ratio of 1.001 (95 per cent confidence interval 0.999 to 1.002) for a 1-unit change in the cholesterol ratio. This was most implausible, given that cholesterol is a known strong cardiovascular risk factor. Shortly afterwards, the authors reported a changed model that gave a much more plausible hazard ratio of 1.17 (95 per cent confidence interval 1.14–1.20) [51].

The difference between the two sets of results is clarified in a working document [52]. It appears that the authors included the outcome in the imputation model only via the log of the survival time, omitting the event indicator

(Section 5.1). Because most of their data were censored, the event indicator conveys most of the outcome information. This is likely to have biased the cholesterol coefficient towards zero by about 70 per cent (the fraction of missing data). The authors' revised analysis correctly included the censoring indicator as well as the log of the survival time in the imputation model. A second possible problem with the QRISK analysis was that the cholesterol ratio was imputed by imputing the two components separately. Some imputed values of the denominator may have been close to zero, leading to influentially large imputed values of the ratio. This would cause the log hazard ratio to be further biased towards zero, with an associated reduction in the standard error. By excluding extreme values of the imputed ratios, the revised analysis avoided this problem.

10.3. Perfect prediction

Another potential problem arises when any imputation model suffers from perfect prediction. Perfect prediction is a potential problem in regression models for categorical outcomes, including logistic, ordered logistic and multinomial logistic regression models. In logistic regression, perfect prediction occurs if there is a category of any predictor variable in which outcome is always 0 (or always 1): in other words, if the 2-way table of a predictor variable by the outcome variable contains a zero cell. Perfect prediction leads to infinite parameter estimates, which are not in themselves a problem; but it also leads to difficulties in estimating the variance–covariance matrix of the parameter estimates. We have discussed this problem in detail [53]. Briefly, standard software adopts one of two approaches. It may drop terms from the imputation model to avoid perfect prediction: in this case, standard imputation procedures may end up imputing using the wrong model. Alternatively, it may retain terms and estimate a singular variance–covariance matrix, leading either to very large standard errors or an unsuccessful attempt to correct the standard errors: in these cases, the Normal approximation to the log-likelihood fails, leading to very poor draws of $\boldsymbol{\beta}^*$. Various solutions exist [53]: the one implemented in `ice` involves 'augmenting' the data by adding a few extra observations to the data set (with small weight) so that no prediction is perfect. This makes it possible to use the standard Normal approximation to the likelihood in the proper imputation step, hence getting successful 'draws'. However, at the time of writing, perfect prediction still causes problems in other software.

10.4. Departures from MAR

MI, like all missing data procedures that are based around an MAR assumption, is sensitive to departures from MAR, especially with larger fractions of missing data. One way to deal with this is to include many variables in the imputation model in an effort to make MAR more plausible. This approach is well suited to MI since a complex imputation model can be built without affecting the choice of analysis model: for example, post-randomization variables can be included when imputing missing outcomes in a clinical trial, but not included in the analysis model (Section 5.2) [54]. It is possible to achieve this in a likelihood-based analysis, but this can be harder to implement [55].

If data are thought to be MNAR then there are a few analysis options. First, data can be imputed under MAR and the imputed data sets can be weighted to reflect their plausibility under specified MNAR mechanisms [5]. Second, data can be imputed under MNAR [2, Chapter 6]. A very simple example of the latter idea would be to impute clinical trial outcomes under MAR and then subtract a constant quantity δ_1 from all values in the intervention arm, while adding a second quantity δ_0 to all values in the control arm. This becomes much more complex where more than one variable is believed to be MNAR.

10.5. Non-convergence

As MICE is an iterative procedure, it is important that convergence is achieved. This may be checked by computing, at each cycle, the means of imputed values and/or the values of regression coefficients, and seeing if they are stable. For the example in Section 9, these values appeared stable from the very first cycle. We have never found 10 cycles inadequate, but larger numbers of cycles might in principle be required when incomplete variables are very strongly associated.

10.6. Checking the imputation models

Where the accuracy of the imputation model is important, it can be checked using standard model checking techniques within the MICE algorithm. For example, a plot of residuals against fitted values might be drawn for each imputed variable in the final cycle of a MICE run. Alternatively, after the MICE run, each imputation model could be re-fitted to the imputed data and the techniques of Section 8.5 used.

10.7. Too many variables

A practical difficulty with MI that afflicts many reasonable attempts to impute missing values occurs when the data set contains many (e.g. dozens of) variables; this is not necessarily an issue specific to MICE. Often, it is unknown in

advance what the analysis model should look like—how many and which variables are needed, what functional form is required for continuous variables, even what type of model is appropriate. The advice in Sections 5 and 6 may lead to very large and complex imputation models. In principle, a rich imputation structure is desirable, but in practice, fitting such a complex set of imputation models may defeat the software or lead to model instability. For example, we have found it particularly challenging to work with structures including several nominal categorical variables imputed by multinomial logistic regression; convergence of such large models is an issue, tending to make the imputation process unacceptably slow. It is hard to propose universal solutions, but careful exploration of the data may suggest smaller imputation models that are unlikely to lead to substantial bias. In general, one should try to *simplify* the imputation structure without damaging it; for example, omit variables that seem on exploratory investigation unlikely to be required in 'reasonable' analysis models, but avoid omitting variables that are in the analysis model or variables that clearly contribute towards satisfying the MAR assumption. Further practical experience and research is needed to develop useful rules of thumb.

11. Discussion

Multiple imputation is an increasingly popular method of analysis. However, like any powerful statistical technique, it must be used with caution. We have highlighted several key areas, including choice of imputation models, handling categorical and skewed variables, identifying what quantities can and cannot be combined using Rubin's rules and avoiding pitfalls. We hope to have encouraged readers to use MI widely, but with understanding and care.

We are often asked what fraction of data can be imputed. The theoretical answer is that almost any fraction of data can be validly imputed, *provided* that the imputation is done correctly and the MAR assumption is correct, but that any imperfections in the imputation procedure and any departures from MAR will have a proportionately larger impact when larger fractions of data are imputed. In the QRISK study, the large fraction of missing data (70 per cent) amplified the consequences of imperfections in the imputation procedure. It would seem wise to take special care if more than 30–50 per cent missing data are to be imputed.

It is important to report MI analyses in a way that allows readers to assess the adequacy of the methods used. Sterne *et al.* suggest reporting guidelines that include careful comparison of MI results with the results of complete-case analysis [46]: although well-implemented MI results should be superior to complete-case results, it should be possible to understand the differences between the two analyses in terms of their different assumptions about the missing data and the usually greater precision of MI.

We now consider various alternatives to the methods we have described.

MI is not the only way to handle missing data. Likelihood-based methods [55] and inverse probability weighting [56] are often good alternatives. Even complete-case analysis may be appropriate in some settings [57, 58], for example in a clinical trial with missing data only in the outcome, although sensitivity analyses would also be required to explore the impact of departures from MAR.

MICE is not the only way to perform MI. The NORM algorithm [59] is based on MCMC sampling and so is more theoretically founded, but assumes that the data are multivariate Normal. Results from NORM agree asymptotically with those from MICE when all imputation models are linear [11], but may perform poorly with non-Normal or categorical data [60]. A simpler procedure is available for monotone missing data: simply impute variables in increasing order of missingness, using appropriate regressions on more complete variables only [2, Chapter 5.4].

Finally, Stata is not the only software package that implements MI. In R, there are user-contributed libraries for MICE [9] and NORM [59]. In SAS, PROC MI implements the NORM algorithm [61], and IVEware is a user-contributed implementation of the MICE algorithm [11, 62]. In SPSS, the missing values module implements the MICE algorithm [63].

We have focussed on unstructured data sets. Longitudinal data can be straightforwardly imputed by regarding the different time points as different variables. Imputing clustered (multi-level) data is harder, since the imputations should respect the multi-level structure. If clusters are large, then it may be reasonable to treat cluster as a fixed effect in the imputation model. If clusters are very small—say, all of size 1 or 2—then it may be possible to format the data with one record per cluster and different variables for the first and second cluster members, and then to use conditional imputation so that the second variable is only imputed when the cluster contains more than one individual. If clustering is not the main focus of analysis, then it may be adequate in practice to ignore the clustering in the imputation model and only allow for it in the analysis model: this is a topic for future research. Finally, a set of MLwiN macros for imputing multi-level data are available from www.missingdata.org.uk.

Other topics for future research are exploring the implications of incompatible conditional models in MICE, evaluating the performance of congenial but mis-specified imputation models, exploring how large the imputation model(s) can safely be, developing methods valid under specified MNAR mechanisms, exploring the performance of PMM, and exploring multivariable model building with fractional polynomials or splines.

From a practical perspective, it is worth considering at what stage of the data analysis process missing data should be imputed. Early work assumed that imputation would be done just once on a data set and that the imputed data would be released to all users. This arrangement is appealing for large surveys, when the imputer may have access to confidential information that is helpful for imputing but which cannot be publicly released. However, it requires the imputer to allow for any association that might be of interest to future analysts, a virtually impossible task. We prefer the arrangement of imputing data once for each project to be conducted on a data set, since it should be possible to produce a moderate list of variables and interactions that might be considered and include them all in the imputation model. It may then be desirable to re-impute the data in a way that is congenial with key analysis models. A third alternative is imputing data once for each individual analysis: this makes it easier to ensure that the imputation model is appropriate, but typically involves more imputation effort than most analysts are willing to tolerate.

Appendix A: Stata implementation for passive and JAV approaches

A.1. Non-linear models (*Section 6.3*)

The JAV approach is implemented by

```
. generate sat94sq = sat94^2
. ice sat96 sat94 sat94sq, m(100) clear
```

The linear passive approach changes the `ice` call to

```
. ice sat96 sat94 sat94sq, m(100) clear passive(sat94sq:sat94^2)
```

The improved passive approach changes the `ice` call to

```
. ice sat96 sat94 sat94sq, m(100) clear passive(sat94sq:sat94^2) match
```

After each form of imputation, the analysis model is fitted by

```
. mim: regress sat96 sat94 sat94sq
```

A.2. Interactions (*Section 6.4*)

A linear passive approach uses

```
. generate randmanual = rand*manual
. ice sat96 sat94 rand manual randmanual, m(500) clear passive(randmanual:rand*manual)
```

The improved passive approach uses

```
. generate randmanual = rand*manual
. generate sat96rand = sat96*rand
. ice sat96 sat94 rand manual randmanual sat96rand, m(500) clear
    passive(randmanual:rand*manual \ sat96rand:sat96*rand)
```

Separate imputation uses

```
. ice manual sat96 sat94, m(500) clear by(rand)
. generate randmanual = rand*manual
```

The JAV approach uses

```
. generate randmanual = rand*manual
. ice sat96 sat94 rand manual randmanual, m(500) clear cmd(regress)
```

where `cmd(regress)` overrides the default logistic imputation models for the binary variables `manual` and `randmanual`. After each form of imputation, the analysis model is fitted by

```
. mim: regress sat96 sat94 rand manual randmanual
```

Appendix B: Bias of the JAV method under MAR

One data set of size 5000 was generated from the model $x \sim N(0, 1)$, $y \sim N(x+x^2, 1)$. Values of x were deleted either with probability 0.3 (an MCAR mechanism) or if $y>3$ (an extreme MAR mechanism). Missing values of x were imputed

using the JAV approach: that is, x2 was computed as x^2, and MICE was performed with $m=10$, imputing x using linear regression on y and x2 and imputing x2 using a linear regression on y and x. Estimated values of β_2 from the analysis model $y \sim N(\beta_0+\beta_1 x+\beta_2 x^2, \sigma^2)$ are given in the table.

Missing data mechanism	Coefficient β_2 (true value = 1)	
	Estimate (MC error)	Standard error (MC error)
MCAR	1.002 [0.002]	0.011 [0.001]
MAR	1.440 [0.004]	0.024 [0.002]

The JAV procedure appears unbiased under MCAR but is biased under this form of MAR.

Appendix C: Derivation of Monte Carlo error

Recall that the Monte Carlo error of the point estimate $\widehat{\beta}$ is $\sqrt{B/m}$ and that the standard error of $\widehat{\beta}$ is $\sqrt{W+(1+1/m)B}$. We ignore Monte Carlo error in the standard error because it is typically much smaller (proportionately) than Monte Carlo error in the point estimate. The Monte Carlo error of the test statistic $Z=\widehat{\beta}/\mathrm{se}(\widehat{\beta})$ is therefore approximately $\sqrt{B/m}/(\sqrt{W+(1+1/m)B})$ which can be shown to equal $1/\sqrt{1+m/\mathrm{FMI}}$ where as before $\mathrm{FMI}=B/(B+W)$. The two-tailed P-value is given by $P=2(1-\Phi(|Z|))$, where $\Phi(.)$ is the standard Normal distribution function. A Taylor series approximation gives the Monte Carlo error of the P-value as approximately equal to the Monte Carlo error of Z multiplied by $|\mathrm{d}P/\mathrm{d}Z|=2\phi(Z)/\sqrt{(1+m/\mathrm{FMI})}$, where $\phi(.)$ is the standard Normal density function. The results given in Section 7 follow on setting $\mathrm{FMI}/m=0.01$.

Acknowledgements

The authors thank the UK700 investigators for permission to use a subset of their data; Shaun Seaman for very helpful comments and discussions; and all the audiences on our courses for asking good questions.

Ian White and Patrick Royston were supported by UK Medical Research Council grants U.1052.00.006 and U.1228.06.01.00002.01.

References

1. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: Hoboken, NJ, 2002.
2. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
3. Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. *Statistics in Medicine* 2007; **26**:3057–3077.
4. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* 2007; **61**:79–90.
5. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research* 2007; **16**(3):259–275.
6. Robins J, Wang N. Inference for imputation estimators. *Biometrika* 2000; **87**(1):113–124.
7. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall: London, 1997.
8. Barnard J, Rubin DB. Small-sample degrees of freedom with multiple imputation. *Biometrika* 1999; **86**:948–955.
9. van Buuren S, Oudshoorn CGM. Multivariate Imputation by Chained Equations: MICE V1.0 User's manual. *TNO Report PG/VGZ/00.038*. TNO Preventie en Gezondheid: Leiden, 2000. Available from: http://www.multiple-imputation.com/.
10. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 2007; **16**(3):219–242.
11. Raghunathan TE, Lepkowski JM, Hoewyk JV, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 2001; **27**:85–95.
12. StataCorp, *Stata Statistical Software*: *Release 11*. Stata Press: College Station, TX, 2009.
13. Royston P. Multiple imputation of missing values. *Stata Journal* 2004; **4**:227–241.
14. Royston P. Multiple imputation of missing values: update. *Stata Journal* 2005; **5**:188–201.
15. Royston P. Multiple imputation of missing values: update. *Stata Journal* 2005; **5**:527–536.
16. Royston P. Multiple imputation of missing values: Further update of ice, with an emphasis on interval censoring. *Stata Journal* 2007; **7**:445–464.
17. Royston P. Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables. *Stata Journal* 2009; **9**:466–477.
18. Carlin JB, Galati JC, Royston P. A new framework for managing and analysing multiply imputed data sets in Stata. *Stata Journal* 2008; **8**:49–67.

19. Burns T, Creed F, Fahy T, Thompson S, Tyrer P, White I, for the UK700 trial group. Intensive versus standard case management for severe psychotic illness: a randomised trial. *Lancet* 1999; **353**:2185–2189.
20. Johnson NL. Systems of frequency curves generated by methods of translation. *Biometrika* 1949; **36**:149–176.
21. Wright EM, Royston P. Age-specific reference intervals ('normal ranges'). *Stata Technical Bulletin* 1996; **34**:24–34.
22. Little RJA. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics* 1988; **6**:287–296.
23. Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis* 1996; **22**:425–446.
24. Burton A, Billingham LJ, Bryan S. Cost-effectiveness in clinical trials: using multiple imputation to deal with incomplete cost data. *Clinical Trials* 2007; **4**:154–161.
25. Moons KG, Donders RA, Stijnen T, Harrell FE. Jr. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology* 2006; **59**:1092–1101.
26. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**:681–694.
27. Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *Journal of Clinical Epidemiology* 2003; **56**:28–37.
28. Barzi F, Woodward M. Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology* 2004; **160**:34–45.
29. White IR, Royston P. Imputing missing covariate values for the Cox model. *Statistics in Medicine* 2009; **28**:1982–1998.
30. Fay RE. When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Sections*. American Statistical Association, Alexandria, VA, 1992; 227–232.
31. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**:538–558.
32. Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**:473–489.
33. Royston P, Sauerbrei W. *Multivariable Model-building*: *A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables*. Wiley: Chichester, 2008.
34. Von Hippel PT. How to impute squares, interactions, and other transformed variables. *Sociological Methodology* 2009; **39**:265–291.
35. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* 2007; **8**:206–213.
36. Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *American Statistician* 2001; **55**:244–254.
37. Royston P, Carlin JB, White IR. Multiple imputation of missing values: new features for mim. *Stata Journal* 2009; **9**:252–264.
38. Bodner TE. What improves with increased missing data imputations? *Structural Equation Modeling*: *A Multidisciplinary Journal* 2008; **15**:651–675.
39. Wood A, White I, 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**:89–99.
40. Little RJA. Regression with missing X's: a review. *Journal of the American Statistical Association* 1992; **87**:1227–1237.
41. Von Hippel PT. Regression with missing Ys: an improved strategy for analyzing multiply imputed data. *Sociological Methodology* 2007; **37**(1):83–117.
42. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. Wiley: Chichester, 2007.
43. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data. *Statistics in Medicine* 2008; **27**:3227–3246.
44. Meng XL, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* 1992; **79**:103–111.
45. White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomised trials. *Statistics in Medicine* 2005; **24**:993–1007.
46. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal* 2009; **338**:b2393.
47. Gilks W, Richardson S, Spiegelhalter D. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC: London, Boca Raton, FL, 1996.
48. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Statistical Methods in Medical Research* 2007; **16**(3):199–218.
49. van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 2006; **76**:1049–1064.
50. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *British Medical Journal* 2007; **335**:136.
51. Hippisley-Cox J, Vinogradova Y, Robson J, May M, Brindle P. QRISK: authors' response. *British Medical Journal* 2007. Available from: http://www.bmj.com/cgi/eletters/bmj.39261.471806.55v1#174181.
52. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. QRISK cardiovascular disease risk prediction algorithm—comparison of the revised and the original analyses technical supplement, 2007. Available from: http://www.qresearch.org/Public_Documents/QRISK1%20Technical%20Supplement.pdf.
53. White IR, Daniel R, Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics and Data Analysis* 2010; **54**:2267–2275.
54. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001; **6**:330–351.
55. Carpenter JR, Kenward MG. *Missing Data in Clinical Trials—A Practical Guide*. National Institute for Health Research, Publication RM03/JH17/MK: Birmingham, 2008. Available from: http://www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml.
56. Hogan JW, Lancaster T. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research* 2004; **13**:17–48.
57. Allison PD. Multiple imputation for missing data: a cautionary tale. *Sociological Methods and Research* 2000; **28**:301–309.
58. White IR, Carlin JB. Is multiple imputation always better than complete-case analysis for handling missing covariate values? *Statistics in Medicine*, DOI: 10.1002/sim.3944.

59. Schafer JL. Software for multiple imputation, 2008. Available from: http://www.stat.psu.edu/~jls/misoftwa.html.
60. Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine* 2007; **26**:1368–1382.
61. SAS Institute Inc., *SAS/STAT 9.1 User's Guide*. SAS Institute Inc.: Cary, NC, 2004, chapter 46.
62. Raghunathan TE, Solenberger PW, Hoewyk JV. *IVEware Imputation and Variance Estimation Software*, 2007. Available from: http://www.isr.umich.edu/src/smp/ive.
63. SPSS inc. *Build Better Models When You Fill in the Blanks*. Available from: http://www.spss.com/media/collateral/statistics/missing-values.pdf (20 April 2010).