Statistical Methods in Medical Research 2007; **16**: 199–218

Multiple imputation: current perspectives

Michael G Kenward and James Carpenter Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London, UK

This paper provides an overview of multiple imputation and current perspectives on its use in medical research. We begin with a brief review of the problem of handling missing data in general and place multiple imputation in this context, emphasizing its relevance for longitudinal clinical trials and observational studies with missing covariates. We outline how multiple imputation proceeds in practice and then sketch its rationale. We explore the problem of obtaining proper imputations in some detail and distinguish two main classes of approach, methods based on fully multivariate models, and those that iterate conditional univariate models. We show how the use of so-called uncongenial imputation models are particularly valuable for sensitivity analyses and also for certain analyses in clinical trial settings. We also touch upon other forms of sensitivity analysis that use multiple imputation. Finally, we give some open questions that the increasing use of multiple imputation has thrown up, which we believe are useful directions for future research.

1 Introduction

Since its introduction nearly 30 years ago[1] in the survey analysis setting, multiple imputation (MI) has become an important and influential approach in the statistical analysis of incomplete data. It now has a very large bibliography, including several reviews and texts.[2–8] During this period, its range of application has grown to include the analysis of observational data from public health research and clinical trials. In parallel with these developments, tools for MI have been incorporated into several mainstream statistical packages. Inevitably, its increasing breadth of use has thrown up new issues and challenges. In this paper, we outline how MI proceeds in practice and sketch its rationale. We consider its application to both observational and experimental medical data, and review recent developments and available tools. We conclude with some open questions that the increasing use of MI has thrown up, which we believe are useful directions for future research.

We begin in the next section by introducing some standard definitions and outlining the well-established framework for dealing with the problem of missing data, within which we place MI. At the same time, we differentiate between medical data from observational and experimental medical research (e.g., randomized trials), within each of which we see MI as having an important but rather different role. In Section 3, we introduce the basic steps that constitute the MI procedure. Although a full understanding

Address for correspondence: Michael G. Kenward, Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. E-mail: mike.kenward@lshtm.ac.uk

 10.1177/0962280206075304

of the formal justification of the method is not needed for the rest of this paper, the essential ideas are sketched in the Appendix. Section 4 discusses methods for generating proper imputations, reviewing recent developments and highlighting some remaining unresolved questions. There is now increasing agreement that in many settings, the analysis of partially observed data sets should be accompanied by appropriate and relevant sensitivity analyses. In Section 5 we discuss some of the ways in which MI can be used for this. Section 6 concludes with a discussion and outlines some areas for future research.

2 Analyzing incomplete datasets

There is now a very large literature on the problem of statistical analysis when data are missing. A useful elementary introduction is provided by Allison,[9] with more extensive, technical accounts by Little and Rubin[8] and Schafer.[5] Much of material the rest of this section is expanded in the latter two references.

Suppose that we have a study in which it is intended that each subject provides observations on p different variables $\boldsymbol{z} = (z_1, \ldots, z_p)^{\mathrm{T}}$, but in practice for some subjects, some of these observations may be missing. Divide the data (for one subject) into those observed and those missing, denoted, respectively, $\boldsymbol{z}_{\mathrm{o}}$ and $\boldsymbol{z}_{\mathrm{m}}$. Denote by $\boldsymbol{r}$ a vector of p indicator variables associated with $\boldsymbol{z}$, each element of which takes the value one if the corresponding value of $\boldsymbol{z}$ is observed and zero if it is missing. We term the process by which data become missing as the *missing data mechanism* and this can viewed as defining a probability distribution for $\boldsymbol{r}$ given the $\boldsymbol{z}$, denoted $P(\boldsymbol{r} \mid \boldsymbol{z})$. If the variables can be ordered in such a way that, for a subject, z_i is missing implies z_k is missing for all $k > i$ then we say that the missing data pattern is *monotone*. This is most likely to occur with longitudinal data when i indexes time or visit, and missing values arise due to attrition or dropout.

We assume that the aim of the statistical analysis is to make inferences about the population represented by all the subjects in the study, and for this we have a *substantive* model, typically (although not necessarily) some form of regression model, in which some of the z's will be outcome variables and some explanatory variables. For the moment, we do not distinguish between these, but when this becomes necessary we will use y and x for outcome and explanatory variables, respectively. Examples are a logistic regression model, as might be used in a cohort study, for which y represents the binary disease status and x the set of exposure and confounder variables; or an analysis of covariance model for a randomized trial in which y represents the outcome at the final time point and x the baseline measurements and treatment indicator. For simplicity, we omit the distinction between the outcome and explanatory variables in the following expressions.

Our starting point is the joint model for the complete data $\boldsymbol{z}$ and the missingness indicator $\boldsymbol{r}$, $f(\boldsymbol{z}, \boldsymbol{r})$. This can be written in many ways, of which two are particularly useful. The first is the *selection model* factorization

$$f(\boldsymbol{z}, \boldsymbol{r}) = f(\boldsymbol{z}) P(\boldsymbol{r} \mid \boldsymbol{z})$$

This provides a natural viewpoint for defining key assumptions as it shows the substantive model, $f(\boldsymbol{z})$, explicitly. The alternative factorization gives the *pattern-mixture model*[10]

$$f(\boldsymbol{z}, \boldsymbol{r}) = f(\boldsymbol{z} \mid \boldsymbol{r})P(\boldsymbol{r}) \tag{1}$$

which we return to at the end when considering sensitivity analysis using MI. The pattern-mixture representation allows, in principle, a different outcome model for different missing value patterns. Unfortunately, the relationship between parameters in the two factorizations, such as treatment effect or disease/exposure odds-ratio, can be very complicated.

In this framework, valid inference about the population[a] is obtained by fitting the substantive model to the *complete* data. Faced with missing data, the aim remains valid inference, but now using only the available, or observed, data. The following definitions, based on those introduced by Rubin,[11] provide a formal framework for considering how this can be done. This framework is constructed using the missing data mechanism. Data are said to be missing completely at random (MCAR) if the probability of being missing is independent of $\boldsymbol{z}$

$$P(\boldsymbol{r} \mid \boldsymbol{z}) = P(\boldsymbol{r})$$

Data are said to be missing at random (MAR) if the conditional distribution of $\boldsymbol{r}$ given the observed data is independent of the unobserved data

$$P(\boldsymbol{r} \mid \boldsymbol{z}) = P(\boldsymbol{r} \mid \boldsymbol{z}_o) \tag{2}$$

When neither MCAR nor MAR hold the missing data mechanism is said to be missing not are random (MNAR).

Under MCAR, any analysis is valid if subjects are excluded because they have some missing data, and these missing data arise completely by chance. For example, if patients dropout of a trial completely randomly, an analysis using only the subset of patients who do not dropout provides valid inferences. Data are also MCAR if the pattern of observed data arises by design – provided that the design is independent of $\boldsymbol{z}$. For example, in a follow-up study, we may choose to only follow up a random 50% of those initially enrolled. The key point is that in both examples, the observed sample remains an unbiased representation of the original population.

The MAR assumption occupies a very special position in the missing value framework, not because it is especially plausible in practice, but because it represents the most general condition under which valid inference can be obtained without reference to the missing data mechanism. However, such analyses must be likelihood based. Under both MCAR and MAR, we can ignore the model $f(\boldsymbol{r}|\boldsymbol{z})$, hence these are often grouped together and termed *ignorable*. Non-likelihood analyses, such as those based on estimating equations, are in general only valid under MCAR, and need modification (e.g., by weighting[12]) to give valid inference under MAR.

---

[a]That is, consistent estimators (which 'home-in' on the true value in the population as the sample size increases), confidence intervals which achieve their nominal coverage, etc.

When data are MNAR valid analysis requires explicit incorporation of the missing data mechanism, which in most situations will be unknown. It is important to appreciate that, apart from data with design-based incompleteness, the observed data cannot be used to distinguish between MAR and MNAR mechanisms without additional untestable assumptions. This is simply because, it is the missing data alone that can distinguish between the MAR and NMAR assumptions, although this is not transparent in some formulations. So, although the MAR assumption leads to considerable simplification when analyzing missing data, it can rarely be justified in the way that the normality assumption in regression can be justified by examining residuals.

Thus, it is an inescapable feature of analyses of partially observed data that they depend on untestable assumptions. In this sense, such analyses belong to a more general class of problem that includes unobserved confounders, measurement error, and noncompliance. As a consequence, there is a general agreement that ideally the analysis should be repeated under varied assumptions, to see if the substantive conclusions are sensitive to them. It is perhaps in such *sensitivity analyses* that MNAR models have their appropriate role.

Many methods can be used for the analysis of incomplete data. A key distinction is between what we term *principled* and *unprincipled methods*. Principled methods are based on statistical models for the data, and the methods of model fitting, analysis and inference are based on formal statistical paradigms, principally frequentist and Bayesian. Unprincipled methods are characterized by ad hoc procedures – typically manipulating the data so that the analysis originally intended for fully observed data can be run. Examples are completers only analyses, last observation carried forward for longitudinal data, and single imputation methods (without proper variance correction). Although unprincipled methods may give valid inference in certain settings, such settings are typically narrow and often both unrealistic and difficult to establish in practice. Moreover, because they lack a principled foundation, they often behave unexpectedly in different settings. An example of this in a clinical trial setting is given by Molenberghs *et al*.[13]

Among principled methods we distinguish between two broad approaches. In the first, the missing data are integrated out of the joint distribution for the observed and missing observations. This calculation yields the distribution of the *observed* data $\{\boldsymbol{z}_o, \boldsymbol{r}\}$ from the distribution of the *complete* data $\{\boldsymbol{z}_o, \boldsymbol{z}_m, \boldsymbol{r}\}$. In so doing it shows us the role played by the parameters from the original substantive model in the distribution of what we have actually observed.

To look at this in a little more detail, we denote the substantive model by $f(\boldsymbol{z} \mid \boldsymbol{\phi})$ and the missing value model/mechanism by $P(\boldsymbol{r} \mid \boldsymbol{z}; \boldsymbol{\beta})$ where the parameter vectors $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ are distinct, that is, the two models share no parameters. The joint distribution of the observed data can then be written as

$$\begin{aligned} f(\boldsymbol{z}_o)P(\boldsymbol{r} \mid \boldsymbol{z}_o) &= \int f(\boldsymbol{z}_o, \boldsymbol{z}_m)P(\boldsymbol{r} \mid \boldsymbol{z}_o, \boldsymbol{z}_m)\mathrm{d}\boldsymbol{z}_m \\ &= f(\boldsymbol{z}_o)\ E_{\boldsymbol{z}_m|\boldsymbol{z}_o}\{P(\boldsymbol{r} \mid \boldsymbol{z}_o, \boldsymbol{z}_m)\} \end{aligned} \tag{3}$$

We call the conditional distribution $\boldsymbol{z}_m$ given $\boldsymbol{z}_o$ the conditional predictive distribution for the missing data and denote it $g(\boldsymbol{z}_m \mid \boldsymbol{z}_o)$. Under MAR $P(\boldsymbol{r} \mid \boldsymbol{z}_o, \boldsymbol{z}_m) = P(\boldsymbol{r} \mid \boldsymbol{z}_o)$. Then, from equation (3), a likelihood analysis for the substantive model parameter

$\boldsymbol{\phi}$ can be based on $f(\boldsymbol{z}_o)$ alone. When the substantive model is a regression and only responses are missing, $f(\boldsymbol{z}_o)$ can be readily obtained and the analysis is comparatively straightforward.[25]

However, when covariates are missing it is not usually so simple. In a regression model, inferences are made conditionally on the observed covariate values. Thus if none is missing, their joint distribution does not enter into the analysis: it is irrelevant. However, once some are missing we need to introduce a joint distribution for all $\boldsymbol{z} = (\boldsymbol{y}, \boldsymbol{x})$ not just $(\boldsymbol{y} \mid \boldsymbol{x})$.

Under MAR (i.e., ignoring $f(\boldsymbol{r} \mid \boldsymbol{z}_o)$),

$$f(\boldsymbol{z}_o, \boldsymbol{z}_m) = f(\boldsymbol{y}_o, \boldsymbol{y}_m \mid \boldsymbol{x}_o, \boldsymbol{x}_m) f(\boldsymbol{x}_m \mid \boldsymbol{x}_o) f(\boldsymbol{x}_o)$$

so that

$$f(\boldsymbol{z}_o) = f(\boldsymbol{y}_o \mid \boldsymbol{x}_o) f(\boldsymbol{x}_o) = \int \int f(\boldsymbol{y}_o, \boldsymbol{y}_m \mid \boldsymbol{x}_o, \boldsymbol{x}_m) f(\boldsymbol{x}_m \mid \boldsymbol{x}_o) f(\boldsymbol{x}_o) \, \mathrm{d}\boldsymbol{x}_m \, \mathrm{d}\boldsymbol{y}_m$$

Although we integrate over the unobserved covariates, inferences are still made conditionally on the *observed* covariate values, and as usual, the marginal distribution of the observed covariates $f(\boldsymbol{x}_o)$ is irrelevant for inference about parameters in the substantive regression model. However, the integration required is now often analytically intractable.

A variety of methods have been developed or applied to tackle this problem, often indirectly. Examples are the expectation-maximization (EM) algorithm,[14] Monte Carlo Newton Raphson and Monte Carlo likelihood,[15] mean score methods,[16] and fully Bayesian methods based on Markov Chain Monte Carlo (MCMC).[17]

An alternative, which is often computationally attractive, is to treat the missing covariates as responses. Then, as noted following (3), $f(\boldsymbol{z}_o)$ can usually be readily obtained (especially if response and missing covariates are continuous) and the model fitted. However, we now need to calculate the regression parameters originally of interest from the fitted model (which has at least one of the intended covariates as a response). However, often this is not hard, and sometimes computational tricks can make the desired parameter estimates fall out automatically. For example, this approach can often be used in a clinical trial analysis to obtain a baseline adjusted treatment estimate when some baseline values are missing (for full details see Chapter 3 in Carpenter and Kenward[25]).

In general, though, these issues mean that handling missing covariates can be much more complicated than dealing with missing outcomes only, and consequently quite different approaches may be appropriate in the two settings.

The second broad approach uses weighting by the inverse probability of completely observing $\boldsymbol{z}$ to adjust for non-random missingness. It is in this way, for example, generalized estimating equations can be modified to provide valid inferences under MAR. The original idea comes from Horvitz and Thompson,[18] and has recently been extensively developed, especially with respect to efficiency and robustness by Robins and co-workers.[19,20] An important point to note is that such estimators do not need, at least in their simplest form, knowledge of the conditional predictive distribution $g(\boldsymbol{z}_m \mid \boldsymbol{z}_o)$. Rather they use the missing data mechanism $P(\boldsymbol{r} \mid \boldsymbol{z})$, or $P(\boldsymbol{r} \mid \boldsymbol{z}_o)$ under MAR.[21]

In the light of this, we now consider what MI has to offer in a medical context. To do this we distinguish two broad types of problem. Although an oversimplification, this helps to clarify the ideas.

1) Randomized trials with missing longitudinal outcomes.
   Randomized trials commonly involve the collection of outcome variables on a number of occasions, and some missing data, especially due to early withdrawal or dropout, is almost inevitable. In this setting, missingness mainly occurs in the outcome variable rather than the covariates. Inference is often required about the treatment effect at a pre-specified occasion, typically the last. Dropout may be associated with termination of treatment or other forms of non-compliance, and often an intention to treat (ITT) analysis will be required.
2) Missing covariates in observational studies.
   It is common in epidemiological studies for there to be some missing values on many, if not all, exposures and confounders. The underlying design may be complex, with hierarchical and/or longitudinal features.

Above, we distinguished likelihood and weighting methods for missing data. MI is a principled, likelihood-based method, that is, it uses a form of integration over the distribution of the missing data. The way this works is sketched in the Appendix. It has certain key features that make it very attractive for both setting (1) and (2)

1) It is based on two distinct models: the substantive model, the target of the analysis, and the *imputation model*, essentially defining what we called the conditional predictive distribution (3). As we describe below, these two models do not have to be wholly consistent with each other.
2) In terms of the substantive model, MI uses only quantities from *complete-data* analyses. These are typically much easier to obtain than those from direct modelling of the incomplete data.
3) The basic steps can be applied very generally – they are not problem specific.
4) MI preserves the correct conditional framework when there are incomplete covariates.
5) MI can be used to construct comparatively simple sensitivity analyses.

We have already noted there are other methods (e.g., the EM algorithm) that can be used. The key advantages of MI are flexibility and generality. Further, as a method, it works in a fairly transparent way.

As MI is principled, and likelihood based, it has the advantage of *efficiency*. In contrast to completers only analyses, MI incorporates information from subjects with incomplete sets of observations. The second advantage is *bias correction*. When the missing data mechanism is MAR, as opposed to MCAR, the method corrects for biases in completers-only analyses and other ad hoc analyses. Of course, both these properties rest on the assumptions underlying the substantive and imputation models, but as we stressed earlier, MI is not special in this; *all* analyses for missing data rest on untestable

assumptions. A third advantage of MI is that the imputation model and substantive model are kept separate. This means

1) additional covariates can be included in the imputation model, to maximize the chance MAR holds, which we do not want to adjust for/condition on in the substantive model, and;
2) it is relatively simple to perform sensitivity analysis, that is, to explore the implications of departures from assumptions. The imputation model can be changed relatively easily to reflect possible departures from assumptions such as MAR; the substantive model is then fitted to the imputed data in the usual way.

The latter is important (e.g., for ITT analyses), and we return to this in Section 5.

3 The MI procedure

To see how this separation works, consider a simple linear regression setting in which we assume that the substantive model is

$$y_i \mid x_i \sim N(\alpha + \beta x_i; \sigma^2) \tag{4}$$

for $i = 1, \ldots, n$ pairs (y_i, x_i), with some of the x_i missing. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$, with $\boldsymbol{x}$ divided into the observed and missing portions $\boldsymbol{x}_{\mathrm{o}}$ and $\boldsymbol{x}_{\mathrm{m}}$ respectively. We assume that the probability that x_i is missing may depend on y_i, noting that, if this is true, a simple regression of y_i on x_i using the complete pairs only would lead to biased estimators of the target regression parameters.

The first step in a MI analysis for this problem is the introduction of an *imputation* model. In the broadest sense, this is a regression model that relates the distribution of the missing data to the observed data, although there are some additional subtleties here to which we return later. Here we might posit the following regression model

$$x_i \mid y_i \sim N(\xi + \eta y_i; \tau^2)$$

Effectively we are assuming that (y_i, x_i) has a bivariate normal distribution. In specifying a distribution model for a covariate, we are *adding* to the usual assumptions required for a regression model. Recall that with complete data no such additional assumptions are needed.

The imputation model is fitted to complete pairs using the Bayesian paradigm. The reason behind the need to use Bayesian imputations is outlined in the Appendix. From the resulting posterior distribution, a set of random draws is made for the *missing* x_i's, given the observed y_i's, and these are used to fill in for the missing observations and so complete the dataset. This imputation process is then repeated M times, resulting in M completed datasets.

In the second step, the substantive model is fitted, in turn, to each of the completed-datasets. A great advantage here is the ability to apply complete-data methods.

This produces M sets of estimates, together with their associated measures of precision. Typically, but not necessarily, conventional maximum-likelihood procedures might be used for this. The estimates and measures of precision are then combined according to rules developed by Rubin, to produce an overall MI estimate and associated standard error. Finally, inferences are then drawn from these quantities using conventional-frequentist procedures, such as those based on t- or F-tests and confidence intervals. One interesting feature of MI is the combination of the Bayesian paradigm at the imputation step and approximate frequentist inference at the end.

In the simple regression setting this procedure would be as follows. For the jth imputation ($j = 1, \ldots, M$) a draw is made from the posterior distribution of the imputation parameters:

$$(\tilde{\xi}_j, \tilde{\eta}_j, \tilde{\tau}_j^2) \sim f(\xi, \eta, \tau \mid \boldsymbol{y}, \boldsymbol{x}_o)$$

Note that under NMAR we would also need to condition on the missing value process in defining this distribution; under MAR (2) we can omit this. Using these parameter values, the set of missing x_i's is replaced by random draws from the distribution

$$N(\tilde{\xi}_j + \tilde{\eta}_j y_i, \tilde{\tau}_j^2) \tag{5}$$

This whole process is then repeated M times to produce the M completed datasets.

The substantive model is fitted in turn to each dataset to provide M sets of parameter estimates. Focusing on the regression coefficient β, we then have M estimates $(\tilde{\beta}_1, \ldots, \tilde{\beta}_M)$ and M associated variances $\tilde{V}_1, \ldots, \tilde{V}_M$. The MI estimator of β is just the average of the $\tilde{\beta}_j$'s

$$\tilde{\beta}_{\mathrm{MI}} = \frac{1}{M} \sum_{j=1}^{M} \tilde{\beta}_j \tag{6}$$

and the estimator of the variance of this is a simple combination of within- and between-imputation variability. Define $\bar{V}$ to be the *average within-imputation variance*

$$\bar{V} = \frac{1}{M} \sum_{j=1}^{M} \tilde{V}_i$$

and B to be the *between-imputation variance* of the estimators

$$B = \frac{1}{M-1} \sum_{j=1}^{M} (\tilde{\beta}_j - \tilde{\beta}_{\mathrm{MI}})^2$$

Then the variance of $\tilde{\beta}_{\mathrm{MI}}$ is estimated by

$$\tilde{V}_{\mathrm{MI}} = \bar{V} + (1 + M^{-1})B \tag{7}$$

The term $(1 + M^{-1})$ adjusts for the fact we are effectively conditioning on the finite, M, number of imputations.

These formulae can be generalized in a obvious way for vector valued estimands. Rubin[2] shows that inferences can then be based (for $p \times 1$ estimand $\boldsymbol{\beta}$), on the approximate relationship

$$(\tilde{\boldsymbol{\beta}}_{\mathrm{MI}} - \boldsymbol{\beta})^{\mathrm{T}} \tilde{\mathbf{V}}_{\mathrm{MI}}^{-1} (\tilde{\boldsymbol{\beta}}_{\mathrm{MI}} - \boldsymbol{\beta}) \sim F_{p,\nu} \tag{8}$$

or the square root of this, when $p = 1$. For example, under the null hypothesis that a single parameter β is zero, we would use the familiar t-statistic

$$\frac{\tilde{\beta}}{\sqrt{\tilde{V}}}$$

on ν degrees of freedom (DF). The residual DF ν are calculated according to formulae given in standard references.[2,22] It is possible for other complete data quantities such as P-values and likelihood-ratio statistics to be combined across imputations,[23,24] but the method above dominates practical use.

We make one remark here concerning the generalization of Rubin's formulae (6) and (7), to vector-valued parameters. The expression (8) involves the inverse of the between-imputation variance matrix B. With few imputations, this matrix may be poorly estimated or even singular (when $M < p$, the number of parameters in the vector $\boldsymbol{\beta}$). There exist approximations to this that are more stable in such situations. However, given modern computing resources, simply increasing M is usually the most practical solution.

When the imputation model uses only variables in the substantive model and is compatible with the substantive model then, under MAR, we should expect the results of MI to be similar, if a little less efficient, than a full-likelihood analysis. In these circumstances, if the likelihood-based analysis is comparatively easy to do (whether frequentist or Bayesian), then there is little to be gained from MI. This is precisely the situation with a continuous outcome from a longitudinal clinical trial with dropout, when the primary analysis is to be made assuming MAR. Often a likelihood-based mixed model analysis will then suffice. However, when additional variables are to be included in the imputation model or particular forms of MNAR analysis are relevant then this is no longer necessarily the case, and MI has several potential advantages. We return to this point in Section 5.

In the early use of MI in a sample survey setting, it was important for practical reasons that the number of imputations, M, be kept small. An important feature of the method was its success in practice with M as small as 3 and a value of $M = 5$ was typically used. Practical experience, together with theoretical results of Rubin,[2] suggested that such values of M should be sufficient, except when large proportions of data are missing. The following remark is typical[6] "Unless rates of missing information are unusually high, there tends to be little or no practical benefit to using more than five to ten imputations." With the great increase in available computing power, it has become practicable to do

MI with many more imputations in routine problems and examine the consequences of this. We have found, in contrast to the commonly expressed view given above, that values of M can be required that are far greater than 5–10, in some instances reaching 100–200 before the results are sufficiently accurate for critical inferences.[25]

4 Creating appropriate imputations

Unsurprisingly many of the practical issues with MI concern the choice of, and Bayesian draws from, the imputation model. Schafer[5] summarizes the formal requirements for a MI to be valid (pp. 109–110). For simpler settings, like the regression model used as an illustration earlier, it is possible to check formally that these conditions hold. For more realistic problems such a justification is difficult to construct and broader guidelines are needed. Rubin[2] provides the following (pp. 126–127):

- Draw imputations following the Bayesian paradigm as repetitions from a Bayesian posterior distribution of the missing values under the chosen models for non-response and data, or an approximation to this posterior distribution that incorporates appropriate between-imputation variability.
- Choose models of non-response appropriate for the posited response mechanism.
- Choose models for the data that are appropriate for the complete-data statistics likely to be used – if the model for the data is correct, then the model is appropriate for all complete-data statistics.

We now consider these points. In general, the imputation model should both contain variables known to be predictive of missingness *and* accommodate structure, for example, interactions, in the substantive model. In particular, for imputing covariates, the outcome variable must be included as an explanatory variable in the imputation model. Failure to accommodate the structure appropriately can cause bias in the resulting analysis.[26] Thus a key practical issue is selecting the variables for the imputation model. This is clearly bound up with the substantive setting of the problem and so it is difficult to provide general guidelines. However, if many potential variables are available some selection will have to be made and this should depend on subject matter considerations as well as more formal variable selection procedures. Because of the Bayesian nature of the imputation step, over-fitting, that is including redundant predictors, may be expected to reduce the precision of the final estimates to some extent (typically not large in our experience), but should not cause other problems like bias. In contrast, omission of important predictors of missingness would be expected to produce bias. In the light of this, it may be better in the imputation model to err on the side of over- rather than under-fitting.

It has been suggested that inferences are fairly robust to the choice of imputation distribution itself. Clearly, this depends very much on the setting, in particular on the substantive model and the nature, pattern and proportion of the missing data. However, some success has been reported using a normal regression model for the case of missing binary data (Section 5.1 in ref. 5), although predictably inference can be inaccurate when probabilities are extreme.[27]

We now consider more formally schemes for generating proper imputations. Recall that in the general case, we have a study in which it is intended that p variables ($\boldsymbol{z}$) will be observed from each subject, and some of these ($\boldsymbol{z}_{\mathrm{m}}$) may be missing. Each of these may be an outcome or an explanatory variable in the substantive model, or may be in the imputation model only (as an explanatory variable). In its most general form, the imputation model provides a, typically multivariate, regression of $\boldsymbol{z}_{\mathrm{m}}$ on $\boldsymbol{z}_{\mathrm{o}}$ with conditional distribution

$$g(\boldsymbol{z}_{\mathrm{m}} \mid \boldsymbol{z}_{\mathrm{o}}, \boldsymbol{r})$$

We have seen already that under MAR, $\boldsymbol{r}$ can be ignored to give

$$g(\boldsymbol{z}_{\mathrm{m}} \mid \boldsymbol{z}_{\mathrm{o}})$$

which provides considerable simplification. In particular, for simpler modelling structures the required imputation model can be estimated from using subjects with complete data alone.

Care needs to be taken in conceptualizing MAR when the missingness pattern is non-monotone.[28,29] If different subjects have different patterns of missingness then it is difficult to conceive of plausible missing value mechanisms that maintain the ignorability assumption. Although mathematically it is sufficient for each subject to have their own MAR mechanism, which could potentially be different from every other subject, this is often rather contrived. In practice, to keep the problem manageable, this issue is usually put to one side, and $\boldsymbol{r}$ is simply omitted from the imputation model.

The main problem is then to construct and fit an appropriate imputation model $g(\boldsymbol{z}_{\mathrm{m}} \mid \boldsymbol{z}_{\mathrm{o}})$, and make posterior Bayesian draws from it. With the exception of problems where missingness is confined to the outcome variables, or only a single variable is incomplete (which is unusual) we are then faced with constructing a joint model for a combinations of types of variable: continuous, binary, ordinal, and nominal categorical. This is the situation we are most likely to meet with observational data, and much of what follows is mainly relevant to that setting. The joint modelling of disparate types of variable is well known to be an awkward problem. Two routes have established themselves. We begin with the cross-sectional setting.

In the first approach, a joint multivariate normal distribution is assumed for all variables except nominal (with more than two categories). In the second, a series of univariate conditional models are used in the *spirit* of the Gibbs sampler, although only in very special circumstances can these be shown to correspond to a genuine joint distribution. We consider each of these two routes in turn. A third approach uses non-parametric methods through Bayesian versions of hot-deck imputation, Rubin (ref. 2, Chapter 5), Herzog and Rubin,[30] and Heitjan and Little[31] are comparatively early examples. Much recent work in medical contexts has focused on explicit model-based methods and we concentrate here on these.

In many settings, simpler approximate imputation draws can be substituted for full blown Bayesian imputation via MCMC. A range of such methods for generating proper imputations is given (ref. 32, Section 10.2.3). In sufficiently large samples, we can often do acceptably well by approximating the posterior predictive distribution using a multivariate normal distribution with mean and covariance matrix given by the

maximum-likelihood estimates. In some settings, we can obtain these using data from subjects with no missing values, which may be acceptably precise when the proportion of missing data is not great. Consider the simple normal regression model used as an illustration in Section 3

$$x_i \mid y_i \sim N(\xi + \eta y_i, \tau^2)$$

Using ordinary least squares we can get consistent estimators, $\hat{\boldsymbol{\gamma}}$ and $\hat{\tau}^2$ say, of $\boldsymbol{\gamma} = (\xi, \eta)^{\mathrm{T}}$ and τ^2, from the complete pairs. Approximate draws from a Bayesian posterior for these parameters can then be made as follows

$$\tilde{\tau}^2 = \frac{(m-2)\hat{\tau}^2}{X} \quad \text{for } X \sim \chi^2_{m-2}$$

$$\tilde{\boldsymbol{\gamma}} \sim N[\hat{\boldsymbol{\gamma}}; \tilde{\tau}^2(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F})^{-1}]$$

where m is the number of completers, and $\boldsymbol{F}$ is the $m \times 2$ matrix with rows consisting of $(1, y_i)$ from the completers. Finally, the missing x_i's are drawn using these parameter values as in equation (5).

This method is simple to extend to longitudinal data with attrition, where at each time point previously imputed values are used in the imputation model as predictors. Here, the notion of 'completer' refers, at each time point, to a subject with data up to this point, and the set of such completers changes (in fact decreases) as we move forward through the time points. More generally, such approaches can be applied with maximum-likelihood estimators from other classes of regression model, provided there is sufficient data so that the posterior distribution of the parameters can be approximated by the maximum-likelihood estimator and its covariance matrix.

These methods are best suited to monotone missing value patterns. With general patterns of missingness, for a multivariate normal imputation distribution, it is not difficult to set up MCMC samplers to provide small sample draws from the appropriate posterior. Schafer (Chapter 5 in ref. 5) provides details. Such methods have now been widely implemented, for example in SAS PROC MI[33] and in Splus.[34] Provided we are able to approximate other types of variable (e.g., ordinal and binary) by the normal, possibly after transformation, this provides an imputation tool with wide applicability. Missing nominal data needs a rather different approach, and specific samplers have been developed for this setting using the multinomial distribution (Chapter 7 in ref. 5). Large sample approximations using likelihood, as described above, are also applicable with cross-classified nominal observations using log-linear models. Flexible joint models for mixed types of variable are not as readily available as those for continuous and categorial data.

The methods so far described are all coherent in the sense that they start with a genuine multivariate distribution for the joint imputation distribution, even though the posterior from this distribution may be approximated, or the distribution itself may be regarded as only a rough approximation to the actual variables being imputed. The alternative route mentioned earlier is to start with an appropriate *univariate conditional* distribution for each variable to be imputed and impute in turn from these cyclically in a Gibbs sampling

type method, sometimes called the 'chained equations' method. The advantage of this approach is that it avoids the problem of specifying an appropriate joint imputation distribution and replaces this by the selection of appropriate univariate conditional distributions. Moreover, if there are specific constraints on particular variables, such as positivity, or joint constraints on two or more variables, such as pregnancy only occurring in females, these can be imposed comparatively simply.

Broadly the method is as follows. Suppose, $\boldsymbol{z}_{\mathrm{m}} = (z_{\mathrm{m}1}, \ldots, z_{\mathrm{m}q})$ are to be imputed conditional on $\boldsymbol{z}_{\mathrm{o}}$. For each $z_{\mathrm{m}k}$, $k \in (1, \ldots, q)$, an appropriate regression model is set up

$$f(z_{\mathrm{m}k} \mid z_{\mathrm{m}1}, \ldots, z_{\mathrm{m}(k-1)}, z_{\mathrm{m}(k+1)}, \ldots, z_q, \boldsymbol{z}_{\mathrm{o}}), \quad k = 1, \ldots, q$$

This might be a logistic regression for a binary variable, or a log-linear model for a categorical one. To get going, starting values are needed for all missing values. These can be variable means or random samples from the observed values of the variable. For each variable in turn, the model is fitted and appropriate proper imputations are drawn to impute the missing values of that variable. The approximate large sample likelihood method discussed above is convenient for this, but other methods could be used.

Thus, at the end of one cycle, that is, k going from $(1, \ldots, q)$, imputations will exist for all the missing values. In the spirit of the Gibbs sampler, the whole cycle is repeated several times, taking the draws from the last cycle to form the first imputed dataset. In contrast to genuine Gibbs samplers however, only a comparatively few cycles are used, typically 10 or 20. Then the whole process is repeated to obtain further sets of imputed values. This approach has been implemented by Raghunathan *et al.*[35,36,IVEWARE] and van Buuren *et al.*[37,MICE]. More recently MICE has been implemented in Stata.[38,39]

The method is much simpler for monotone missingness, for then starting from the variable with the least missing values, one cycle is sufficient to provide a set of imputations, and no starting values are needed. This is essentially a generalization of the method described earlier for longitudinal data. In the light of this, there is probably some advantage when using this approach in first re-ordering the variables so that the missing value pattern is as monotone as possible, so that the role of the repeated conditional sampling is minimized.

While this method has been seen to work acceptably well in some settings, it must be remembered that a joint limiting distribution for the imputations is not guaranteed to exist. In this context, Gelman and Raghunathan[40] note that '[i]t is hard to establish convergence in the general case, but simulation studies suggest that the coverage properties in some important practical cases are quite good'. This is clearly an area which requires further investigation; it would be very useful to have a more formal justification for a method with such great practical attraction.

By considering the cross-sectional setting we have, up till now, ignored possible dependence structure in the data. Longitudinal data with attrition is an exception which can be handled effectively by repeated use of cross-sectional methods. There are many settings however with dependent data, for example clustered and hierarchical data, and longitudinal data with intermediate missingness, where such simplifications do not exist. If a substantive model is to be used which reflects this structure, such as a generalized linear mixed model, then it is important that the imputation model also reflects this structure.

MI for structured substantive models is much less well developed than for cross-sectional ones. Using the multivariate normal distribution it is possible to formulate appropriately structured multivariate imputation models. Two examples of such an approach are PAN for Splus[34,41] and MLwiN,[42,43] and an application is described in ref. 44. Many problems however involve several types of variable, not just continuous, and, unless one is prepared to use the normal approximation for these, additional developments are required. The chained equation approach is very convenient for this in the cross-sectional setting, but its formulation in the structured setting is far from straightforward and as far as we are aware only the simplest settings have so far been considered.

5 Uncongenial imputation and sensitivity analysis

We have already noted that all analyses for incomplete data rest on assumptions that are not verifiable from the data under analysis. Thus many have argued for a second step in such analyses where the sensitivity of the results to key assumptions is explored. In such *sensitivity analysis*, the definition of 'key' is of course highly problem specific and open to debate. It is then important that it is clear what assumptions are being examined, and to what variations on these sensitivity is being presented. MI provides one particular, flexible, route for doing this.

A key assumption in many primary analyses based on MI is that of MAR, even if in practice issues surrounding non-monotoneness are skirted in the actual procedures used. Many examples of sensitivity analysis using MI are concerned with departures from MAR. Before exploring how these might be approached we make a distinction between two types of sensitivity analysis that might be considered. In the first, a range of MNAR models are fitted, possibly with parameters governing the non-random missingness fixed rather than estimated, and the results compared with those obtained under the MAR assumption. Such analyses might use a frequentist, for example[45,46] or Bayesian for example[17,47] paradigms. Such methods require the estimation of the model under the MNAR missingness assumption, so lose much of the inherent simplicity of MI that arises under MAR. Similar ends can be achieved in the MI setting using reweighting (see the paper by Carpenter *et al*. in the current issue).

The second approach to sensitivity analysis uses the important feature of MI, that the imputation and substantive model can be *uncongenial*,[48] that is the two models may reflect different structures. Not all uncongenial combinations will lead to valid conclusions,[48] gives details of conditions on these, but in the current setting we can legitimately use MNAR imputations in combination with the same substantive model used with the MAR imputations to provide an assessment of sensitivity to a particular non-ignorable missingness process in the imputation model. Potentially, this is comparatively simple to do, because the imputer has control over the imputation procedure and can (within reason) introduce any desired non-random process. Some examples of this procedure in a survey context are given by Longford[49,50] and Taylor *et al*.[36] Such methods seem particularly relevant to the observational data setting.

A simple, and flexible, application of this approach has an important role in sensitivity analysis for longitudinal clinical trials with dropout.[10,51] We can introduce an explicit non-random dropout model into the imputation process.[45,47] Or we can modify directly the future behaviour of dropouts conditional on the past (under the MAR assumption such conditional behaviour is the same for those who dropout and those who do not). Using pattern-mixture models this behaviour can be allowed to differ among the dropouts and completers. In the MI context, this is particularly convenient if the model for future behaviour of the dropouts can be constructed from components from the MAR model, for then the existing Bayesian draws from the MAR model can be used in the first stage of the MI procedure.[52] Little and Yau[53] use this approach to construct ITT analyses in which it is assumed that the subjects who dropout differ from completers in their compliance (under the MAR assumption compliance is assumed to be the same in both groups). Other forms of future behaviour for the dropouts might also be considered in this way as part of a sensitivity analysis. One great advantage of this approach is that it is relatively simple to communicate the implied departures from MAR to which sensitivity is being assessed.

A related use of MI in such clinical trial settings, which is not necessarily part of a sensitivity analysis but does use uncongenial substantive and imputation models, is the inclusion of variables in the imputation model which may be predictive of dropout, but on which we do not want to condition in the substantive model. An example might be degree of compliance – in some settings this is very predictive of dropout. Similar ends can be achieved by *joint* modelling of the outcome and additional variables, but this may be awkward to do in all but the simplest settings.

Finally, such uncongenial imputations may be used when it is much more convenient to have an imputation model that has a quite different structure to the substantive model. The longitudinal clinical trial again provides such an setting, this time with non-normal outcomes. For some problems, such as fitting population averaged models to repeated binary outcomes, non-likelihood methods of analysis might have advantages, especially when a full-likelihood analysis is complicated and/or very demanding computationally. In particular, generalized estimating equations are commonly used in such settings. Such methods are valid under MCAR but not MAR. One solution is to fit a model consistent with MAR, impute from this, and then fit the (not necessarily congenial) substantive model to the imputed data.

We have noted that the population averaged (or marginal) substantive model does not specify the entire joint distribution of the repeated outcomes (in particular the dependence structure is left unspecified), and so cannot be used as a basis for imputing data under MAR. Several alternatives are possible. One could 'fill-out' the specification of the population averaged model by introducing dependence structure and impute from this. But such models are not convenient imputation models for the same reason they are not convenient substantive models and, if this route is pursued, one may as well anyway use a full likelihood analysis in the first place.

Another option is to use a convenient and sufficiently rich joint model for the imputations, but one which is not parameterized in a population averaged way (and so is uncongenial). Two options are 1) a generalized linear mixed model in which the dependence is induced through subject random effects, and 2) a log-linear model. The latter

is a relatively simple choice for which imputation procedures are already implemented in widely available software.

6 Discussion and future directions

It is probably true that in most settings suitable analyses with missing data can be constructed that do not require the use of MI. Where such analyses are convenient to do, such as with likelihood-based analyses for longitudinal studies with attrition, MI offers little advantage under MAR. For smaller problems fully Bayesian analyses are often quite feasible using a tool like WinBUGS.[54] However MI has (at least) three distinct and important advantages. First, it can be applied very generally, to very large datasets with complex patterns of missingness among covariates, and uses only complete data quantities with very simple rules of combination. This makes it especially attractive for observational studies. Secondly, even when the substantive analysis under MAR is relatively straightforward, MI provides a relatively flexible and convenient route for investigating sensitivity to postulated NMAR mechanisms. Thirdly, the imputation model may include variables not in the substantive model, which can lead to additional efficiency, or most importantly in the clinical trial setting, allow post-randomization covariates in the imputation model if they are predictive of dropout. We do re-iterate at this point, that however convenient, *no* method of analysis can be expected to provide an 'automatic' solution to the problem of missing data, and any approach used must be carefully considered in the context of the problem.

Many of the issues and difficulties with MI surround the construction and use of the imputation model, in particular for mixtures of types of variable. There have been many important developments in this since the original introduction of MI, and two important routes have become delineated, joint modelling through the multivariate normal, with approximation for non-normal variables, and the use of conditional univariate models, without rigorous formal justification. Many variations on these basic themes exist.

With implementations of MI in several mainstream statistical packages, and a growing and realistic appreciation of its potential value, its use in a medical setting is very likely to increase. However, for us, there remains a wide range of issues that are open to further development. Specifically,

1) Does the semi-automatic use of MI require new forms of model selection and diagnostic procedures (e.g., to detect influential points, impossible imputations, and non-linearities in the imputation models)? If so what form should these take?
2) A more rigorous theoretical basis is needed for the chained equation approach.
3) Reliable and flexible imputation methods are needed for mixtures of types of variable from structured (i.e., hierarchical and/or longitudinal) data. This could be done using formal joint models. Using latent multivariate normal structures to tackle this is particularly appealing and would parallel developments other areas of Bayesian modelling. Alternatively, it may also be possible to tackle this using the 'chained equation' approach.
4) Methods for appropriate, communicable, sensitivity analysis need to continue to be developed and moved into the mainstream.

5) Guidelines are needed for presenting primary analyses, MAR analyses and sensitivity analyses in applied research.

The papers in this issue of *Statistical Methods in Medical Research* provide examples of recent developments in MI which bear directly on these points.

Acknowledgement

We are very grateful to Patrick Royston and Ian White for their helpful comments on an earlier draft of this paper.

References

1 Rubin DB. Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. Proceedings of the Survey Research Methods Section of the American Statistical Association, 1978, pp. 20–34.
2 Rubin DB. *Multiple imputation for nonresponse in surveys*. Wiley, 1987.
3 Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine* 1991; **10**: 585–98.
4 Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**: 473–90.
5 Schafer JL. *Analysis of incomplete multivariate data*. Chapman and Hall, 1997.
6 Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research* 1999; **8**: 3–15.
7 Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *American Statistician* 2001; **55**: 244–54.
8 Little RJA, Rubin DB. *Statistical analysis with missing data*, second edition. Wiley, 2002.
9 Allison PD. *Missing data*. Sage Publications, 1994.
10 Little RJA. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 1993; **88**: 125–34.
11 Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–92.
12 Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**: 106–121.
13 Molenberghs G, Thijs H, Jansen I, Beunkens C, Kenward MG, Mallinkrodt G, Carroll RC. Analyzing incomplete longitudinal clinical trial data. *Biostatistics* 2004; **5**: 445–64.
14 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* 1977; **39**: 1–38.
15 McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* 1997; **92**: 162–70.
16 Reilly M, Pepe M. A mean score method for missing and auxilliary covariate data in regression models. *Biometrika* 1995; **82**: 299–314.
17 Best NG, Spiegelhalter DJ, Thomas A, Brayne CEG. Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society, Series A* 1996; **159**: 232–342.
18 Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**: 663–85.
19 van der Laan MJ, Robins JM. *Unified methods for censored longitudinal data and causality*. Springer, 2002.
20 Tsiatis AA. *Semiparametric theory and missing data*. Springer, 2006.
21 Carpenter JR, Kenward MG, Vansteelandt S. A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal*

Statistical Society, Series A 2006; **169**: 571–584.
22 Li KH, Raghunathan TE, Rubin DB. Large-sample significance levels from multiply imputed data using moment-based statistics and an *F* reference distribution. *Journal of the American Statistical Association* 1991; **86**: 1065–73.
23 Li KH, Meng XL, Raghunathan TE, Rubin DB. Significance levels from repeated *p*-values with multiply-imputed data. *Statistica Sinica* 1991; **1**: 65–92.
24 Meng XL, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* 1992; **89**: 267–78.
25 Carpenter J, Kenward MG. *Missing data in clinical trials*. UK National Health Service, National Centre for Research on Methodology, 2006, in press.
26 Fay RE. When are inferences from multiple imputation valid? Proceedings of the Survey Research Methods Section of the American Statistical Association, 1992, pp. 227–32.
27 Schafer JL, Schenker N. Inference with imputed conditional means. *Journal of the American Statistical Association* 2000; **95**: 144–54.
28 Gill R, van der Laan M, Robins J. *Coarsening at random: characterizations, conjectures and counterexamples*. Proceedings of the First Seattle Symposium on Survival Analysis, 1996, pp. 255–94.
29 Robins JM, Gill RD. Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine* 1997; **16**: 39–56.
30 Herzog TN, Rubin DB. Using multiple imputations to handle nonresponse in sample surveys. *Incomplete data in sample surveys*, Volume 2: *Theory and Bibliographies*, 1983, Academic Press, pp. 115–42.
31 Heitjan DF, Little RJA. Multiple imputation for the fatal accident reporting system. *Applied Statistics* 1991; **40**: 13–29.
32 Little RJA, Rubin DB. *Statistical analysis with missing data*. Wiley and Sons, 1986.
33 SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513, USA.
34 Splus 6.1 for Windows: Insightful Corporation, 1700 Westlake Avenue N, Suite 500, Seattle, Washington 98109, USA.
35 Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 2001; **27**: 85–95.
36 Taylor JMG, Cooper KL, Wei JT, Sarma AV, Raghunathan TE, Heeringa SG. Use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among African–American men. *American Journal of Epidemiology* 1987; **82**: 528–50.
37 van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**: 681–694.
38 Stata Corporation, 702 University Drive East, College Station, Texas 77840, USA.
39 Royston P. Multiple imputation of missing values. *The Stata Journal* 2004; **3**: 227–41.
40 Gelman A, Raghunathan TE. Using conditional distributions for missing-data imputation. *Statistical Science* 2001; **15**: 268–69.
41 Schafer J. *Imputation of missing covariates under a general linear mixed model*. Technical report, Department of Statistics, Penn State University, 1997.
42 MLwiN: Centre for Multilevel Modelling, Institute of Education, 20 Bedford Way, London WC1H 0AL, UK.
43 Carpenter JR, Goldstein H. Multiple imputation in MLwiN. *Multilevel Modelling Newsletter* 2004; **16**: 9–18.
44 Liu M, Taylor JMG, Belin TR. Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics* 2000; **56**: 1157–63.
45 Kenward MG. Selection models for repeated measurements with nonrandom dropout: an illustration of sensitivity. *Statistics in Medicine* 1998; **17**: 2723–32.
46 Robins JM, Rotnitzky A, Scharfstein DO. Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association* 1998; **93**: 1321–39.
47 Carpenter JR, Pocock S, Lamm CJ. Coping with missing values in clinical trials: a model based approach applied to asthma trials. *Statistics in Medicine* 2002; **21**: 1043–66.
48 Meng X-L. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1995; **10**: 538–58.
49 Longford NT. An experiment in primary health care evaluation. *Journal of the Royal*

Statistical Society, Series A 1999; **162**: 291–302.
50 Longford NT. Handling missing data in diaries of alchohol consumption. *Journal of the Royal Statistical Society, Series A* 2000; **163**: 381–402.
51 Little RJA. A class of pattern-mixture models for multivariate incomplete data. *Biometrika* 1994; **81**: 471–83.
52 Kenward MG, Molenberghs G, Thijs H. Pattern-mixture models with proper time dependence. *Biometrika* 2003; **90**: 53–71.
53 Little RJA, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics* 1996; **52**: 1324–33.
54 WinBUGS, MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 2SR, UK.
55 Wang N, Robins JM. Large-sample thoery for parametric multiple imputation procedures. *Biometrika* 1998; **85**: 935–48.
56 Robins JM, Wang N. Inference for imputation estimators. *Biometrika* 2000; **85**: 113–24.

Appendix

An outline justification for the MI procedure

At the heart of the MI method is a Bayesian argument. Suppose we regard the missing data z_m as a 'parameter' (legitmate in a Bayesian framework) and we are interested in inferences about γ. Thus we have a statistical model with two parameters γ, z_m and observed data z_o. In a Bayesian analysis, these have a joint posterior distribution

$$f(z_m, \gamma \mid z_o)$$

Our focus is on γ with z_m being regarded as a nuisance. As the posterior can be partitioned as follows

$$f(\gamma, z_m \mid z_o) = f(z_m \mid z_o) f(\gamma \mid z_m, z_o)$$

(where we have suppressed the parameters of $f(z_m \mid z_o)$) the marginal posterior for γ can be written

$$f(\gamma \mid z_o) = E_{z_m|z_o}\{f(\gamma \mid z_m, z_o)\}$$

In particular the posterior mean and variance for γ can be expressed

$$E(\gamma \mid z_o) = E_{z_m|z_o}\{E_\gamma(\gamma \mid z_m, z_o)\}$$

$$V(\gamma \mid z_o) = E_{z_m|z_o}\{V_\gamma(\gamma \mid z_m, z_o)\} + V_{z_m|z_o}\{E_\gamma(\gamma \mid z_m, z_o)\}$$

These can be approximated using empirical moments. Let $\boldsymbol{z}_{\mathrm{m}}^{j},\ j = 1, \ldots, M$, be draws from the conditional predictive distribution $\boldsymbol{z}_{\mathrm{m}} | \boldsymbol{z}_{\mathrm{o}}$, then approximately

$$E(\gamma \mid \boldsymbol{z}_{\mathrm{o}}) \simeq \frac{1}{M} \sum_{j=1}^{M} \{E_{\gamma}(\gamma \mid \boldsymbol{z}_{\mathrm{m}}^{j}, \boldsymbol{z}_{\mathrm{o}})\} = \tilde{\gamma} \quad \text{say}$$

and

$$V(\gamma \mid \boldsymbol{z}_{\mathrm{o}}) \simeq \frac{1}{M} \sum_{j=1}^{M} V_{\gamma}(\gamma \mid \boldsymbol{z}_{\mathrm{m}}^{j}, \boldsymbol{z}_{\mathrm{o}}) + \frac{1}{M-1} \sum_{j=1}^{M} \{E_{\gamma}(\gamma \mid \boldsymbol{z}_{\mathrm{m}}^{j}, \boldsymbol{z}_{\mathrm{o}}) - \tilde{\gamma}\}^{2}$$

We assume that in sufficiently large samples the conditional posterior moments for γ can be approximated by maximum likelihood, or suitably efficient, estimators from the completed data sets. This shows why we need to use imputation (conditional predictive) draws from a proper Bayesian posterior. We also see that the MI estimates of the parameters of interest and its variance approximate first two moments of the posterior distribution in a fully Bayesian analysis. The large sample approximation of the posterior by a normal distribution also suggests that combination of estimators should be done on the scale for which the posterior is better approximated by the normal distribution, for example using log odds-ratios when the estimand of interest is the odds-ratio.

The justification for the final *frequentist* step is surprisingly subtle. We refer to the works of Rubin (ref. 2, Sections 4 and 5), Wang and Robins,[55] Robins and Wang,[56] and Tsiatis (ref. 9, Section 14) for details.