Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls

Jonathan A C Sterne,[1] Ian R White,[2] John B Carlin,[3] Michael Spratt,[1] Patrick Royston,[4] Michael G Kenward,[5] Angela M Wood,[6] James R Carpenter[5]

Most studies have some missing data. **Jonathan Sterne and colleagues** describe the appropriate use and reporting of the multiple imputation approach to dealing with them

RESEARCH, p 144

[1]Department of Social Medicine, University of Bristol, Bristol BS8 2PR
[2]MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 0SR
[3]Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, and University of Melbourne, Parkville, Victoria 3052, Australia
[4]Cancer and Statistical Methodology Groups, MRC Clinical Trials Unit, London NW1 2DA
[5]Medical Statistics Unit, London School of Hygiene and Tropical Medicine London, WC1E 7HT
[6]Department of Public Health and Primary Care, Institute of Public Health, Cambridge
Correspondence to: J A C Sterne
jonathan.sterne@bristol.ac.uk

Accepted: 30 January 2009

Cite this as: *BMJ* 2009;338:b2393
doi: 10.1136/bmj.b2393

Missing data are unavoidable in epidemiological and clinical research but their potential to undermine the validity of research results has often been overlooked in the medical literature.[1] This is partly because statistical methods that can tackle problems arising from missing data have, until recently, not been readily accessible to medical researchers. However, multiple imputation—a relatively flexible, general purpose approach to dealing with missing data—is now available in standard statistical software,[2-5] making it possible to handle missing data semiroutinely. Results based on this computationally intensive method are increasingly reported, but it needs to be applied carefully to avoid misleading conclusions.

In this article, we review the reasons why missing data may lead to bias and loss of information in epidemiological and clinical research. We discuss the circumstances in which multiple imputation may help by reducing bias or increasing precision, as well as describing potential pitfalls in its application. Finally, we describe the recent use and reporting of analyses using multiple imputation in general medical journals, and suggest guidelines for the conduct and reporting of such analyses.

Box 1 | Types of missing data*

- *Missing completely at random*—There are no systematic differences between the missing values and the observed values. For example, blood pressure measurements may be missing because of breakdown of an automatic sphygmomanometer
- *Missing at random*—Any systematic difference between the missing values and the observed values can be explained by differences in observed data. For example, missing blood pressure measurements may be lower than measured blood pressures but only because younger people may be more likely to have missing blood pressure measurements
- *Missing not at random*—Even after the observed data are taken into account, systematic differences remain between the missing values and the observed values. For example, people with high blood pressure may be more likely to miss clinic appointments because they have headaches

Consequences of missing data

Researchers usually address missing data by including in the analysis only complete cases—those individuals who have no missing data in any of the variables required for that analysis. However, results of such analyses can be biased. Furthermore, the cumulative effect of missing data in several variables often leads to exclusion of a substantial proportion of the original sample, which in turn causes a substantial loss of precision and power.

The risk of bias due to missing data depends on the reasons why data are missing. Reasons for missing data are commonly classified as: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (box 1).[6] This nomenclature is widely used, even though the phrases convey little about their technical meaning and practical implications, which can be subtle. When it is plausible that data are missing at random, but not completely at random, analyses based on complete cases may be biased. Such biases can be overcome using methods such as multiple imputation that allow individuals with incomplete data to be included in analyses. Unfortunately, it is not possible to distinguish between missing at random and missing not at random using observed data. Therefore, biases caused by data that are missing not at random can be addressed only by sensitivity analyses examining the effect of different assumptions about the missing data mechanism.

Statistical methods to handle missing data

A variety of ad hoc approaches are commonly used to deal with missing data. These include replacing missing values with values imputed from the observed data (for example, the mean of the observed values), using a missing category indicator,[7] and replacing missing values with the last measured value (last value carried forward).[8] None of these approaches is statistically valid in general, and they can lead to serious bias. Single imputation of missing values usually causes standard errors to be too small, since it fails to account for the fact that we are uncertain about the missing values.

When there are missing outcome data in a randomised controlled trial, a common sensitivity analysis is to explore "best" and "worst" case scenarios by replacing missing values with "good" outcomes in one group and "bad" outcomes in the other group. This can be useful if there are only a few missing values of a binary outcome, but because imputing all missing values to good or bad is a strong assumption the sensitivity analyses can give a very wide range of estimates of the intervention effect, even if there are only a moderate number of missing outcomes. When outcomes are quantitative (numerical) such sensitivity analyses are not possible because there is no obvious good or bad outcome.

If we assume data are missing at random (box 1), then unbiased and statistically more powerful analyses (compared with analyses based on complete cases) can generally be done by including individuals with incomplete data. Sometimes this is possible by building a more general model incorporating information on partially observed variables—for example, using random effects models to incorporate information on partially observed variables from intermediate time points[9 10] or by using bayesian methods to incorporate partially observed variables into a full statistical model from which the analysis of interest can be derived.[11] Other approaches include weighting the analysis to allow for the missing data,[12 13] and maximum likelihood estimation that simultaneously models the reasons for missing data and the associations of interest in the substantive analysis.[11] Here, we focus on multiple imputation, which is a popular alternative to these approaches.

What is multiple imputation?

Multiple imputation is a general approach to the problem of missing data that is available in several commonly used statistical packages. It aims to allow for the uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining results obtained from each of them.

The first stage is to create multiple copies of the dataset, with the missing values replaced by imputed values. These are sampled from their predictive distribution based on the observed data—thus multiple imputation is based on a bayesian approach. The imputation procedure must fully account for all uncertainty in predicting the missing values by injecting appropriate variability into the multiple imputed values; we can never know the true values of the missing data.

The second stage is to use standard statistical methods to fit the model of interest to each of the imputed datasets. Estimated associations in each of the imputed datasets will differ because of the variation introduced in the imputation of the missing values, and they are only useful when averaged together to give overall estimated associations. Standard errors are calculated using Rubin's rules,[14] which take account of the variability in results between the imputed datasets, reflecting the uncertainty associated with the missing values. Valid inferences are obtained because we are averaging over the distribution of the missing data given the observed data.

Consider, for example, a study investigating the association of systolic blood pressure with the risk of subsequent coronary heart disease, in which data on systolic blood pressure are missing for some people. The probability that systolic blood pressure is missing is likely to decrease with age (doctors are more likely to measure it in older people), increasing body mass index, and history of smoking (doctors are more likely to measure it in people with heart disease risk factors or comorbidities). If we assume that data are missing at random and that we have systolic blood pressure data on a representative sample of individuals within strata of age, smoking, body mass index, and coronary heart disease, then we can use multiple imputation to estimate the overall association between systolic blood pressure and coronary heart disease.

Multiple imputation has potential to improve the validity of medical research. However, the multiple imputation procedure requires the user to model the distribution of each variable with missing values, in terms of the observed data. The validity of results from multiple imputation depends on such modelling being done carefully and appropriately. Multiple imputation should not be regarded as a routine technique to be applied at the push of a button—whenever possible specialist statistical help should be obtained.

Pitfalls in multiple imputation analyses

A recent *BMJ* article reported the development of the QRISK tool for cardiovascular risk prediction, based on a large general practice research database.[15] The researchers correctly identified a difficulty with missing data in their database and used multiple imputation to handle the missing data in their analysis. In their published prediction model, however, cardiovascular risk was found to be unrelated to cholesterol (coded as the ratio of total to high density lipoprotein cholesterol), which was surprising.[16] The authors have subsequently clarified that when they restricted their analysis to individuals with complete information (no missing data) there was a clear association between cholesterol and cardiovascular risk. Furthermore, a similar result was obtained after using a revised, improved, imputation procedure.[17] It is thus important to be aware of problems that can occur in multiple imputation analyses, which we discuss below.

Omitting the outcome variable from the imputation procedure

Often an analysis explores the association between one or more predictors and an outcome but some of the predictors have missing values. In this case, the outcome carries information about the missing values of the predictors and this information must be used.[18] For example, consider a survival model relating systolic blood pressure to time to coronary heart disease, fitted to data that have some missing values of systolic blood pressure. When missing systolic blood pressure values are imputed, individuals who develop coronary heart disease should have larger values, on average, than those who remain disease free. Failure to include the coronary heart disease outcome and time to this outcome when imputing

the missing systolic blood pressure values would falsely weaken the association between systolic blood pressure and coronary heart disease.

Dealing with non-normally distributed variables

Many multiple imputation procedures assume that data are normally distributed, so including non-normally distributed variables may introduce bias. For example, if a biochemical factor had a highly skewed distribution but was implicitly assumed to be normally distributed, then imputation procedures could produce some implausibly low or even negative values. A pragmatic approach here is to transform such variables to approximate normality before imputation and then transform the imputed values back to the original scale. Different problems arise when data are missing in binary or categorical variables. Some procedures[19] may handle these types of missing data better than others,[11] and this area requires further research.[20 21]

Plausibility of missing at random assumption

"Missing at random" is an assumption that justifies the analysis, not a property of the data. For example, the missing at random assumption may be reasonable if a variable that is predictive of missing data in a covariate of interest is included in the imputation model, but not if the variable is omitted from the model. Multiple imputation analyses will avoid bias only if enough variables predictive of missing values are included in the imputation model. For example, if individuals with high socioeconomic status are both more likely to have their systolic blood pressure measured and less likely to have high systolic blood pressure then, unless socioeconomic status is included in the model used when imputing systolic blood pressure, multiple imputation will underestimate mean systolic blood pressure and may wrongly estimate the association between systolic blood pressure and coronary heart disease.

It is sensible to include a wide range of variables in imputation models, including all variables in the substantive analysis, plus, as far as computationally feasible, all variables predictive of the missing values themselves and all variables influencing the process causing the missing data, even if they are not of interest in the substantive analysis.[22] Failure to do so may mean that the missing at random assumption is not plausible and that the results of the substantive analysis are biased.

Data that are missing not at random

Some data are inherently missing not at random because it is not possible to account for systematic differences between the missing values and the observed values using the observed data. In such cases multiple imputation may give misleading results. For example, consider a study investigating predictors of depression. If individuals are more likely to miss appointments because they are depressed on the day of the appointment, then it may be impossible to make the missing at random assumption plausible, even if a large number of variables is included in the imputation model. When data are missing not at random, bias in analyses based on multiple imputation may be as big as or bigger than the bias in analyses of complete cases. Unfortunately, it is impossible to determine from the data how large a problem this may be. The onus rests on the data analyst to consider all the possible reasons for missing data and assess the likelihood of missing not at random being a serious concern.

Where complete cases and multiple imputation analyses give different results, the analyst should attempt to understand why, and this should be reported in publications.

Computational problems

Multiple imputation is computationally intensive and involves approximations. Some algorithms need to be run repeatedly in order to yield adequate results, and the required run length increases when more data are missing. Unforeseen difficulties may arise when the algorithms are run in settings different from those in which they were developed—for example, with high proportions of missing data, very large numbers of variables, or small numbers of observations. These points are discussed more fully elsewhere.[23]

Practical implications

The imputation models that were used in the original and revised versions of the QRISK cardiovascular risk prediction tool discussed above have been clarified.[24] The main reasons for the unexpected finding of a null association between cholesterol level and cardiovascular risk were omission of the cardiovascular disease outcome when imputing missing cholesterol values and calculation of the ratio of cholesterol to HDL based on imputed cholesterol and HDL values, which led to extreme values of the ratio being included in estimations. The impact of these pitfalls was increased by the high proportion of missing data (70% of HDL cholesterol values were missing).

Suggested reporting guidelines

In the era of online supplements to research papers, it is feasible and reasonable for authors to provide sufficient detail of imputation analyses to facilitate peer review, without distracting from the substantive research question. Box 2 lists the information that should be provided, either as supplements or within the main paper. This extends guidance provided as part of the STROBE initiative to strengthen the reporting of observational studies,[25] and complements suggestions for reporting of analyses using multiple imputation in the epidemiological literature.[26]

Box 3 on bmj.com relates the suggested guidelines to the use of multiple imputation in a published paper that examined the cost effectiveness of chemotherapy with that of standard palliative care in patients with advanced non-small cell lung cancer.

Summary

We are enthusiastic about the potential for multiple imputation and other methods[12] to improve the validity of medical research results and to reduce the

Box 2 | Guidelines for reporting any analysis potentially affected by missing data

- Report the number of missing values for each variable of interest, or the number of cases with complete data for each important component of the analysis. Give reasons for missing values if possible, and indicate how many individuals were excluded because of missing data when reporting the flow of participants through the study. If possible, describe reasons for missing data in terms of other variables (rather than just reporting a universal reason such as treatment failure)
- Clarify whether there are important differences between individuals with complete and incomplete data—for example, by providing a table comparing the distributions of key exposure and outcome variables in these different groups
- Describe the type of analysis used to account for missing data (eg, multiple imputation), and the assumptions that were made (eg, missing at random)

For analyses based on multiple imputation

- Provide details of the imputation modelling:
  - Report details of the software used and of key settings for the imputation modelling
  - Report the number of imputed datasets that were created (Although five imputed datasets have been suggested to be sufficient on theoretical grounds,[9] a larger number (at least 20) may be preferable to reduce sampling variability from the imputation process[27])
  - What variables were included in the imputation procedure?
  - How were non-normally distributed and binary/categorical variables dealt with?
  - If statistical interactions were included in the final analyses, were they also included in imputation models?
- If a large fraction of the data is imputed, compare observed and imputed values
- Where possible, provide results from analyses restricted to complete cases, for comparison with results based on multiple imputation. If there are important differences between the results, suggest explanations, bearing in mind that analyses of complete cases may suffer more chance variation, and that under the missing at random assumption multiple imputation should correct biases that may arise in complete cases analyses
- Discuss whether the variables included in the imputation model make the missing at random assumption plausible
- It is also desirable to investigate the robustness of key inferences to possible departures from the missing at random assumption, by assuming a range of missing not at random mechanisms in sensitivity analyses. This is an area of ongoing research[28 29]

waste of resources caused by missing data. The cost of multiple imputation analyses is small compared with the cost of collecting the data. It would be a pity if the avoidable pitfalls of multiple imputation slowed progress towards the wider use of these methods. It is no longer excusable for missing values and the reason they arose to be swept under the carpet, nor for potentially misleading and inefficient analyses of complete cases to be considered adequate. We hope that the pitfalls and guidelines discussed here will contribute to the appropriate use and reporting of methods to deal with missing data.

We thank Lucinda Billingham for checking our description of the article described in box 3.

Contributors: JACS, IRW, JBC, and JRC wrote the first draft of the paper. MS conducted the review of the use of multiple imputation in medical journals and analysed the data. All authors contributed to the final draft and subsequent redrafts of the paper. JACS, IRW, and JRC will act as guarantors

Funding: Funded by UK Medical Research Council grant G0600599. IRW was supported by MRC grant U.1052.00.006 and JBC by NHMRC (Australia) grant 334336.

Competing interests: None declared.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Wood A, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomised controlled trials. *Clin Trials* 2004;1:368-76.
2 Royston P. Multiple imputation of missing values. *Stata J* 2004;4:227-41.
3 Royston P. Multiple imputation of missing values: update of ice. *Stata J* 2005;5:527-36.
4 Multiple Imputation Online. *Software*.www.multiple-imputation.com.
5 SAS Institute. *The MI procedure*. http://support.sas.com/rnd/app/papers/miv802.pdf.
6 Little RJ, Rubin DB. *Statistical analysis with missing data*. 2nd ed. New York: Wiley, 2002.
7 Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *Am J Epidemiol* 1991;134:895-907.
8 Carpenter JR, Kenward MG. A critique of common approaches to missing data. In: *Missing data in randomised controlled trials— a practical guide*. Birmingham: National Institute for Health Research, 2008. www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml.
9 Carpenter JR, Kenward MG. MAR methods for quantitative data. In: *Missing data in randomised controlled trials— a practical guide*. Birmingham: National Institute for Health Research, 2008. www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml.
10 Goldstein H, Carpenter J, Kenward MG, Levin K. Multilevel models with multivariate mixed response types. *Stat modelling* (in press).
11 Schafer JL. *Analysis of incomplete multivariate data*. London: Chapman and Hall, 1997.
12 Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for non-ignorable drop-out using semiparametric non-response models. *J Am Stat Assoc* 1999;94:1096-120.
13 Carpenter JR, Kenward MG, Vansteelandt S. A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *J R Stat Soc [Ser A]* 2006;169:571-84.
14 Rubin D. *Multiple imputation for nonresponse in surveys*. New York: Wiley, 1987.
15 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. *BMJ* 2007;335:136.
16 Peto R. Doubts about QRISK score: total/HDL cholesterol should be important [electronic response to Hippisley-Cox J, et al]. *BMJ* 2007 www.bmj.com/cgi/eletters/335/7611/136#172067.
17 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. QRISK— authors' response [electronic response]. *BMJ* 2007 www.bmj.com/cgi/eletters/335/7611/136#174181.
18 Moons KG, Donders RA, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006;59:1092-101.
19 Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999;18:681-94.
20 Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *Am Stat* 2007;61:79-90.
21 Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat Med* 2007;26:1368-82.
22 Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001;6:330-51.
23 Carpenter J, Kenward M. *Brief comments on computational issues with multiple imputation*. www.missingdata.org.uk/mi_comp_issues.pdf.
24 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 1*. 2007. www.qresearch.org/Public_Documents/QRISK1%20Technical%20Supplement.pdf.
25 Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE initiative. strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335:806-8.
26 Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. *Am J Epidemiol* 2008;168:355-7.
27 Horton NJ, Lipsitz SR. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *Am Stat* 2001;55:244-54.
28 Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Stat Med* 2003;22:2553-75.
29 Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Stat Methods Med Res* 2007;16:259-75.