ELSEVIER

Journal of Clinical Epidemiology 59 (2006) 1087−1091

Journal of Clinical Epidemiology

SPECIAL SERIES: MISSING DATA

Review: A gentle introduction to imputation of missing values

A. Rogier T. Donders[a,b,c,*], Geert J.M.G. van der Heijden[c], Theo Stijnen[d], Karel G.M. Moons[c]

[a]*Center for Biostatistics, Utrecht University, Utrecht, The Netherlands*
[b]*Department of Innovation Studies, Copernicus Institute, Utrecht University, P.O. Box 80125, 3508 TC Utrecht, The Netherlands*
[c]*Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, The Netherlands*
[d]*Department of Epidemiology and Biostatistics, Erasmus University Medical School, Rotterdam, The Netherlands*

Accepted 10 January 2006

Abstract

In most situations, simple techniques for handling missing data (such as complete case analysis, overall mean imputation, and the missing-indicator method) produce biased results, whereas imputation techniques yield valid results without complicating the analysis once the imputations are carried out. Imputation techniques are based on the idea that any subject in a study sample can be replaced by a new randomly chosen subject from the same source population. Imputation of missing data on a variable is replacing that missing by a value that is drawn from an estimate of the distribution of this variable. In single imputation, only one estimate is used. In multiple imputation, various estimates are used, reflecting the uncertainty in the estimation of this distribution. Under the general conditions of so-called missing at random and missing completely at random, both single and multiple imputations result in unbiased estimates of study associations. But single imputation results in too small estimated standard errors, whereas multiple imputation results in correctly estimated standard errors and confidence intervals. In this article we explain why all this is the case, and use a simple simulation study to demonstrate our explanations. We also explain and illustrate why two frequently used methods to handle missing data, i.e., overall mean imputation and the missing-indicator method, almost always result in biased estimates.

Keywords: Missing data; Single imputation; Multiple imputation; Indicator method; Bias; Precision

1. Introduction

Missing data are a common problem in all types of medical research. There are various methods of handling missing data. Simple and frequently used methods include complete or available case analysis, the missing-indicator method [1], and overall mean imputation. However, these methods lead to inefficient analyses and, more seriously, commonly produce severely biased estimates of the association(s) investigated [2−6]. There are more sophisticated (imputation) techniques to handle missing data, such as multiple imputation, that give much better results [2−6]. With these techniques, missing data for a subject are imputed by a value that is predicted using the subject's other, known characteristics. Presently, these sophisticated techniques are easy accessible and available in standard software such as SAS and S-Plus. Nevertheless, there seems to be a general lack of understanding that has limited their use in epidemiological research.

In this short report we will give a gentle introduction into the logic behind these sophisticated imputation techniques of missing data. We will not go into technical details, nor into details on how to perform these analyses. For this we refer to the literature [2−8]. Instead, to assist medical researchers in their future data analyses we aim to clarify in simple wording *why* (more sophisticated) imputation is a better, more valid method than the simple and frequently used techniques for handling missing data. We will start with a brief introduction on different types of missing data and the principles of imputation in general, followed by explaining single and multiple imputation, and why frequently used methods fail. All this will be illustrated using data from a simple simulation study.

2. Types of missing data

If subjects who have missing data are a random subset of the complete sample of subjects, missing data are called missing completely at random (MCAR) [9]. Typical examples

Abbreviations: MCAR, missing completely at random; MNAR, missing not at random; MAR, missing at random.

* Corresponding author. Tel.: +31-30-2534419; fax: +31-30-2532746.
E-mail address: R.Donders@geo.uu.nl (A.R.T. Donders).

0895-4356/06/$ − see front matter
doi: 10.1016/j.jclinepi.2006.01.014

of MCAR are when a tube containing a blood sample of a study subject is broken by accident (such that the blood parameters can not be measured) or when a questionnaire of a study subject is accidentally lost. The reason for missingness is completely random, i.e., the probability that an observation is missing is not related to any other patient characteristics. When missing data are MCAR, evidently the set of subjects with no missing data is also a random sample from the source population. Hence, most simple techniques for handling missing data, including complete and available case analyses, give unbiased results [2]. Obviously, estimating associations using a complete or available case analysis remains less efficient (i.e., imprecise), because part of the data is not used.

If the probability that an observation is missing depends on information that is not observed, like the value of the observation itself, missing data are called missing not at random (MNAR) [9]. For example, when asking a subject for his or her income level it might well be that missing data are more likely to occur when the income level is relatively high. Here, the reason for missingness obviously is not completely at random, but is related to unobserved patient characteristics. If missing data are MNAR, valuable information is lost from the data and, there is no universal method of handling the missing data properly [1–9].

Mostly, missing data are neither MCAR nor MNAR [5]. Instead, the probability that an observation is missing commonly depends on information for that subject that *is* present, i.e., reason for missingness is based on other observed patient characteristics. This type of missing data is confusingly called missing at random (MAR), because missing data can indeed be considered random conditional on these other patient characteristics that determined their missingness and that are available at the time of analysis [9]. For example, suppose we want to evaluate the predictive value of a particular diagnostic test, and the test results are known for all diseased subjects but unknown for a random sample of nondiseased subjects. In this case the missing data would be MAR: conditional on a patient characteristic that is observed (here the presence or absence of the disease) missing data are random. Of course, it need not be that missingness only depends on the outcome variable. But it is the simplest situation, with one dependent or outcome variable and only one independent or predictor variable. Hence, we will use this example throughout the remainder of the article to explain the method of imputation. Moreover, situations where the missing predictor values depend (directly or indirectly) on the outcome status regularly occur in epidemiological research [10,11]. When missing data are MAR, a complete or available case analysis is no longer based on a random sample from the source population and selection bias likely occurs [2–8,12]. Generally, when missing data are MAR, all simple techniques for handling missing data, i.e., complete and available case analyses, the indicator method and overall mean imputation, give biased results. However, more sophisticated techniques like single and multiple imputations give unbiased results when missing data are MAR [2–8,12]. In the next sections we will explain and illustrate both issues.

3. Imputation is replacement

We start this section by noting that in the classical (frequentistic) statistical view, conclusions drawn from any study should not depend on the sample that is involved in the study. Should the study be repeated with a different sample, nearly identical results should be obtained. The conclusions do not depend on the given set of subjects in the sample. This implies that every subject in a randomly chosen sample can be replaced by a new subject that is randomly chosen from the same source population as the original subject, without compromising the conclusions. Imputation techniques are also based on this basic principle of replacement.

In our diagnostic study example we can replace any nondiseased subject with a missing value for the test result in the sample, by a newly chosen, nondiseased subject from the source population for whom the test result is known. Note that the subject characteristics of this new replaced (which can also be read as imputed) subject will and also need not be the same as the characteristics of the original subject. The result is a new random sample from the source population and conclusions drawn from analyzing this sample will be valid for the source population.

Of course, if only nondiseased *male* subjects would have missing values, these should be replaced by subjects randomly chosen from the source population including only nondiseased males. Accordingly, and as explained above, subjects with missing data based on (other) known characteristics—i.e., MAR—are by definition a random subset from the sample given these other known characteristics. Hence, they could be replaced by randomly selected subjects from the part of the source population that we *can* identify by these characteristics. Analyzing the thus completed study sample would lead to valid results, both with respect to bias and precision of the estimated associations because we still analyze a random sample from the source population.

4. Single imputation

Direct replacement of subjects by new subjects from an identifiable source population based on observed subject characteristics may be feasible when the number of study variables is limited, as in our diagnostic example study where only two variables, the test result and disease status, were used. Commonly, however, the number of covariates is large. Suppose a nondiseased male subject, aged 39, with a body mass index of 24.5, and a systolic blood pressure of 110 has a missing test result. If the missingness of the test

result is dependent on all these known covariates (thus MAR), we should replace this subject by a randomly chosen subject from the source population of nondiseased males with the same age, body mass index, and systolic blood pressure. Because this is rather burdensome if not impossible, one can instead use the observed or available data of the other subjects to make an *estimation* of the distribution of the test result (that was initially missing) in the source population. In our example with only one other variable (disease status), one can estimate the distribution of the test result for the nondiseased subjects in a number of ways. For example, assuming that the continuous test result is normally distributed, we can calculate the sample mean (and standard deviation) of the observed nondiseased subjects—i.e., with no missing test results—and use this as estimate for the source population values. For more complex situations in which more subject characteristics are known that may have determined the reason for missingness, this univariable approach does not suffice. One rather needs to use a multivariable approach, e.g., a multivariable regression model, to better estimate the underlying distribution of the test result in the source population. Subsequently, this multivariably estimated distribution can be used to randomly draw a test result to impute the missing test results for the subjects in the study sample. This more sophisticated imputation procedure will be called the *single imputation procedure*. The procedure more sophisticated, because the imputation of the test result is based on various other—but known—characteristics of the subjects, rather than only on the estimated mean of the test result in the observed subjects (i.e., overall mean imputation which will be described below). It is single because we only impute each missing once.

If the estimated distribution of the test results based on the observed subjects in the study sample would be identical to the ''true'' underlying distribution in the source population, the single imputation procedure would be equivalent to direct replacement as described above. This of course will seldomly be the case, but the estimated distribution can certainly be an unbiased estimate of the population distribution. Therefore, the associations under study estimated after missing data have been completed (imputed) by the more sophisticated single imputation—and using standard analytical techniques and software—are unbiased. Doing so, however, one analyzes the completed data set as if all data were indeed observed. Because this was not the case, the single imputation procedure commonly results in an underestimation of the standard errors or too small *P*-values, i.e., overestimation of the precision of the study associations [2–5]. We will illustrate this in a following paragraph using simulated data.

5. Multiple imputation

To obtain correct estimates of the standard errors and *P*-values, we should take into account the imprecision caused by the fact that the distribution of the variables with missing values is estimated. This can be done by creating not a single imputed data set, but several or multiple imputed data sets in which different imputations are based on a random draw from different estimated underlying distributions [4,5]. There are various approaches to creating these multiple imputed data sets. However, because this is an introductory article we refer to a few accessible sources [5,15]. Each imputed data set can again be analyzed using standard analytical techniques. Each analysis will produce an association with standard error, resulting in multiple regression coefficients (or odds ratios) and corresponding standard errors. Because each estimated association is unbiased (assuming that data are MAR), the estimates can simply be averaged to get a pooled estimate of the association. These averaging will generally lower the variance of the combined estimate. The multiple standard errors can also be averaged. The mean of the standard errors is a measure for the uncertainty in the estimated associations caused by sampling the study subjects from a source population. Additionally, the standard deviation of the multiple estimated associations (e.g., regression coefficients) reflects the differences between the imputed data sets, i.e., the uncertainty in the estimated underlying distributions of the variables with missing values. Combining both sources of uncertainty—sampling and imputation—using a simple formula results in a single corrected standard error of the estimated association [4]. Because this formula tends to produce too large and, thus, too conservative standard errors, a more precise formula is available, which, however, is not commonly used in multiple imputation [13].

6. Simulation study

We performed a simulation study based on our diagnostic example to illustrate that single imputation yields unbiased estimates with too narrow confidence intervals and multiple imputation indeed yields unbiased estimates with correct standard errors and *P*-values. We simulated 1,000 samples of 500 subjects using R [14]. The samples were drawn from a population consisting of equal numbers of diseased and nondiseased subjects. The true regression coefficient in a logistic regression model linking the diagnostic test to the probability of disease was 1 (odds ratio = 2.7), with an intercept of 0. The diagnostic test was normally distributed with a mean of 0 and a standard deviation of 2. No other tests or subject characteristics were considered. Of the nondiseased subjects, 80% were given a missing value on the test. The diseased subjects had no missing data. Accordingly, missing data were MAR because they were based on other observed variables, here the true disease status, and overall approximately 40% were missing.

Using the procedure `mice` (for details about the software we refer to the literature [15]), 10 multiple imputed data sets were created. Then the association between the test

and the disease status plus standard error was estimated in each data set using logistic regression. Subsequently, all associations with standard errors were analyzed within each of the 10 multiply imputed data sets and the estimates were combined as described above. One extra data set was imputed and analyzed as a single imputed data set.

The results are given in Table 1. For both the single and the multiple imputation procedures, the estimates of the association are indeed unbiased. The single imputation procedure appears more precise because of the smaller standard error thus leading to smaller confidence intervals, but the 90% confidence interval does not contain the true parameter as often (only 63.6%) as it should (i.e., 90%). Multiple imputation leads to a larger standard error and wider confidence intervals, but the estimated standard errors are more correct and the confidence interval has the correct coverage (i.e., 90.3%). Hence, in contrast to single imputation, multiple imputation gives sound results both with respect to bias and precision.

7. Why do frequently used methods fail?

7.1. Indicator method

A still popular method for handing missing values is the so-called missing-indicator method [1]. For each independent variable with missing values a new dummy or indicator (0/1) variable is created with ''1'' indicating a missing on the original variable and ''0'' indicating an observed value. For the original variable the missing values are recoded as ''0.'' For (original) categorical variables this in fact means, creating an extra value category for the missing values. When estimating the association between the independent variable and the outcome in a multivariable analysis, the indicator is always included together with the original (though recoded) variable. The main advantage of the indicator method is that all subjects are used in the multivariable analysis. Although no subjects need to be excluded, we would not recommend this method even when missing data are MCAR. We illustrate this by extending our simple (two variable) diagnostic example from the previous section.

Suppose again that the population consists of equal numbers diseased and nondiseased subjects and that the diagnostic test is associated with the true disease status. But now we also consider a second test, which is a proxy for the first test. This means that the second test is not directly related to the true disease status, but only via the first test. If we would draw a sample from this population and formulate a logistic regression model to predict disease status on basis of the first test only we would expect a positive regression coefficient (case 1). If we would predict disease status only using the second test we would again expect a positive association, because of the indirect relation between disease status and the second test (case 2). If we would predict disease status using both tests, we would expect only a positive association for the first test, comparable to case 1, and a regression coefficient near 0 for the second test (case 3).

Table 1
Results from a simulation study with true regression coefficient of 1 in which missing data were created according to ''MAR'' and imputed using either single or multiple imputation (for details see text)

Method	Regression coefficient	Standard error	Coverage of the 90% confidence intervals
Single imputation	0.98904	0.090186	63.6
Multiple imputation	0.98920	0.136962	90.3

Suppose now that there are missing values on the first test but not on the second test, and that these missing data are—even—MCAR, i.e., equal proportion of missing values in diseased and nondiseased subjects. When using a logistic model to predict the true disease status based on both tests and using the indicator method for handling the missing values of the first test, the regression coefficient of the second test will now not be ''0'' as should be. For the subjects with no missing data indeed case 3 will apply. But for the subjects who do have missing values on the first test, case 2—rather than case 3—suddenly applies because there are no observations for the first test. Hence, the estimate for the regression coefficient of the second test is biased and will be somewhere between 0, the true estimate (case 3), and the value of case 2. Moreover, if the regression coefficient of the second test is biased then so is the regression coefficient of the first test given the mutual adjustment in multivariable modeling.

To illustrate this, we performed a second simulation study similar to the first simulation study. We again simulated 1,000 samples of 500 subjects drawn from a population consisting of equal numbers of diseased and nondiseased subjects (using R [14]). The true logistic regression coefficient of the (first) test—which was again normally distributed with mean 0 and standard deviation 2—was 1 (odds ratio = 2.7) with intercept 0. We now also simulated a proxy for this diagnostic test result with a mean of 0 and a standard deviation of 2 and a correlation of 0.75 with the first diagnostic test. About 40% missing values were created for the first test completely at random (MCAR), i.e., 20% for the diseased and 20% for the nondiseased subjects.

We used the missing-indicator method to analyze this data set with a logistic regression model with diagnostic test, the proxy of this test, and the indicator variable as predictor variables. Table 2 shows that the regression weights of the diagnostic test are indeed heavily biased (because the true value is 1) and also of the proxy variable (because the true value is 0).

Thus, although the indicator method has the appealing property that all available information and subjects can be used in the analyses, the fact that it can lead to biased

Table 2
Results from a simulation study with true regression coefficient of 1 in which missing data were created according to "MCAR" and where the data were imputed with the overall mean or the indicator method was used (for details see text)

	Diagnostic test	Proxy
Method	Regression coefficient (standard error)	Regression coefficient (standard error)
Indicator method[a]	0.55 (0.14)	0.51 (0.08)
Overall mean	0.55 (0.14)	

[a] The logistic model in this analysis was $\ln\{P(\text{Disease})/(1-P(\text{Disease}))\} = \text{Intercept} + b_1 \times \text{Diagnostic test} + b_2 \times \text{Proxy} + b_3 \times \text{Indicator}$, where the Indicator = 1 if the value for diagnostic test was missing and 0 otherwise, and where Diagnostic test is 0 if the value for diagnostic test was missing.

associations of the original variables and outcome is reason enough to discard this method even when missing data are MCAR, let alone when data are MAR.

7.2. Imputation using the overall sample mean

In this section we will use our simulation study to show the effect of imputation using the overall sample mean. In the simplest (two variable) case where we only consider the association between disease status and a continuous diagnostic test, the magnitude and significance of the association (regression coefficient) of the test with the outcome are based on the difference in overlap of the test result distributions between the diseased and nondiseased subjects. The less the overlap, the higher and more significant the coefficient. If the two distributions completely overlap, the regression coefficient will be "0." We used the same simulation study of the previous paragraph. Because the missing values were MCAR, we have an equal proportion of missing values for the diseased and nondiseased subjects. Imputing these missing values of the test result by the overall sample mean of the test result of the observed subjects—i.e., calculated on the nondiseased and diseased subjects together—will obviously increase the amount of overlap. Accordingly, the association between the test and the outcome will dilute and the regression coefficient will be biased toward "0" and nonsignificance. This is again illustrated in Table 2. The regression coefficient is not 1 but 0.55017.

Like the indicator method, the overall mean imputation of missing values should also be discarded because it will lead to biased associations even when missing data are MCAR.

8. Final comments

Our purpose was to provide insight into how sophisticated imputation works, to facilitate the understanding and cooperation between medical researchers and statisticians, and to make the data analysis a success. Complete and available case analyses provide inefficient though valid results when missing data are MCAR, but biased results when missing data are MAR, which is the more common form of missingness in epidemiological research. Other frequently used methods to handle missing data such as overall mean imputation and the missing-indicator method provide biased results when the missing data are MCAR, let alone when data are MAR. More sophisticated imputation techniques, where imputations are based on other known subject characteristics, are relatively easy to use and allow for the use of standard software to analyze the data once the imputations are made. Moreover, doing such imputation using the multiple imputation approach leads to unbiased results with correct standard errors, in situations where missing data are MCAR or MAR.

Acknowledgments

We gratefully acknowledge the support by The Netherlands Organization for Scientific Research (ZON-MW 904-10-006 and 917-46-360).

References

[1] Miettinen OS. Theoretical epidemiology. Principles of occurrence research in medicine. New York: Wiley; 1985.
[2] Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142:1255–64.
[3] Vach W. Logistic regression with missing values in the covariates. New York: Springer; 1994.
[4] Rubin DB. Multiple imputation for non response in surveys. New York: Wiley; 1987.
[5] Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall/CRC Press; 1997.
[6] Little RA. Regression with missing X's; a review. J Am Stat Assoc 1992;87:1227–37.
[7] Clark TG, Altman DG. Developing a prognostic model in the presence of missing data. An ovarian cancer case study. J Clin Epidemiol 2003;56:28–37.
[8] Van Buuren S, Boshuizen SC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 1999;18:681–94.
[9] Rubin DB. Inferences and missing data. Biometrika 1976;63:581–90.
[10] Oostenbrink R, Moons KGM, Bleeker SE, Moll HA, Grobbee DE. Diagnostic research on routine care data: prospects and problems. J Clin Epidemiol 2003;56:501–6.
[11] Oostenbrink R, Moons KG, Donders AR, Grobbee DE, Moll HA. Prediction of bacterial meningitis in children with meningeal signs: reduction of lumbar punctures. Acta Paediatr 2001;90:611–7.
[12] Vach W, Blettner. Missing data in epidemiological studies. In: Armitrage P, Colton T, editors. Encyclopedia of biostatistics. Chichester: Wiley; 1998. p. 2641–54.
[13] Robins JM, Wang N. Inference for imputation estimators. Biometrika 2000;87:113–24.
[14] R Development Core Team R. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2004. Available at http://www.R-project.org ISBN 3-900051-00-3, Accessed April 19, 2004.
[15] Van Buuren S, Outshoorn K. Flexible multivariate imputation by mice Technical report. Leiden, The Netherlands: TNO prevention and Health; 1999. Available at http://web.inter.nl.net/users/S.vanBuuren/mi/hmtl/mice.htm. Accessed April 1, 2004.