Multiple imputation: dealing with missing data

Moniek C.M. de Goeij¹, Merel van Diepen¹, Kitty J. Jager², Giovanni Tripepi³, Carmine Zoccali³ and Friedo W. Dekker¹

Correspondence and offprint requests to: Moniek C.M. de Goeij; E-mail: m.c.m.de_goeij@lumc.nl

ABSTRACT

In many fields, including the field of nephrology, missing data are unfortunately an unavoidable problem in clinical/epidemiological research. The most common methods for dealing with missing data are complete case analysis—excluding patients with missing data—mean substitution—replacing missing values of a variable with the average of known values for that variable—and last observation carried forward. However, these methods have severe drawbacks potentially resulting in biased estimates and/or standard errors. In recent years, a new method has arisen for dealing with missing data called multiple imputation. This method predicts missing values based on other data present in the same patient. This procedure is repeated several times, resulting in multiple imputed data sets. Thereafter, estimates and standard errors are calculated in each imputation set and pooled into one overall estimate and standard error. The main advantage of this method is that missing data uncertainty is taken into account. Another advantage is that the method of multiple imputation gives unbiased results when data are missing at random, which is the most common type of missing data in clinical practice, whereas conventional methods do not. However, the method of multiple imputation has scarcely been used in medical literature. We, therefore, encourage authors to do so in the future when possible.

INTRODUCTION

Missing data are a problem that is often encountered when performing clinical studies. In any study design, such as a randomized controlled trial, a cohort study and a case–control study, missing data can occur. Especially in studies with routinely collected data, missing values are unfortunately unavoidable. Even after the introduction of statistical software allowing more sophisticated procedures, the methods most often used to deal with missing data are complete case analysis, mean substitution and last observation carried forward (LOCF).

In a complete case analysis, patients with missing data are excluded from the statistical analyses. Consequently, only patients with complete data will contribute to the results. Mean substitution is a method in which missing observations for a certain variable are replaced by the average of observed data for that variable in other patients. However, both of these methods have severe drawbacks. Performing a complete case analysis not only leads to a smaller sample size and, therefore, may reduce statistical power but can also lead to biased results. These biased results can occur when the patients excluded from the analysis have different patient characteristics compared with those who were included. Mean substitution will maintain the original sample size, but can still lead to biased results when the patients with missing data have different patient characteristics compared with those with available data. Furthermore, the variances obtained with mean...
The underlying reasons for data being missing in clinical research, however, are often related to several known patient characteristics. This type of missing data is called MAR, as conditional on the known patient characteristics data are missing randomly. An example of this type of missing data is the higher probability that a blood sample could not be drawn before the start of dialysis in patients who started dialysis acutely. It is highly likely that patient characteristics of subjects starting dialysis acutely are different compared with subjects not starting acutely. In that case, patients with missing blood samples are not a random subset of the study population, and their true laboratory values are probably different from those for whom a blood sample was available. However, it could be possible to predict from the available data which patients are acute starters and, therefore, have missing blood samples. Conditional on these prediction characteristics, meaning within the different strata, subjects with a missing blood sample do constitute a random subset and, therefore, this type of missing data is called MAR.

The third type of missing data, MNAR, is the most complicated type of missing data, as here the underlying reasons for data being missing are related to unknown patient characteristics. For example, subjects with a low education are more likely to skip a question about their level of education in a questionnaire. The level of education depends on many characteristics, which may not all be measured. Neither conventional methods nor the multiple imputation method are suitable when missing data are MNAR. There are some other methods available, but these methods are very complex and subject to many assumptions and are, therefore, outside the scope of this article [7]. It is common in clinical practice that the missing data type MNAR coexists with the other two types (MCAR and MAR). For example, when quality of life data are missing, some patients have not filled in the questionnaire because their pencil broke (MCAR), some because of clinical characteristics—like age—which are measured (MAR) and some patients because they have a low quality of life for reasons unrelated to measured patient characteristics (MNAR). When the MNAR type is rare in this mixture of different missing data types, multiple imputation can be used.

In the next section, we will describe the conventional methods of complete case analysis, mean substitution and LOCF and indicate for which types of missing data, MCAR and MAR, these methods give unbiased results.

**CONVENTIONAL METHODS: WHEN TO USE THEM?**

To illustrate when a complete case analysis, mean substitution and LOCF can be used, we simulated a situation where data were missing. In this example, the mean (standard error) serum urea level at the start and end of pre-dialysis care and the change (standard error) of urea were measured in 1500 incident patients starting pre-dialysis care. Standard compute commands in SPSS version 20 were used for data simulation. The first urea measurement at the start of pre-dialysis care...
(U1) was simulated by taking a random subset of 1500 patients from a normal distribution (mean 20 mmol/L, standard deviation 6, computed with the random variable function `rv.normal`). The second urea measurement at the end of pre-dialysis care (U2) had a mean of 24 and a standard deviation of 6. The correlation between U1 and U2 was specified at 0.75.

Table 1 shows the mean (standard error) U1, U2 and change in urea (U2–U1) for these 1500 patients from the main analysis of our simulation. The increase in urea (U2–U1) during pre-dialysis care was 4.0 (0.11) mmol/L, with a urea level of 20.0 (0.15) mmol/L at the start of pre-dialysis care and 24.0 (0.15) mmol/L at the end.

U2 was made MCAR for ~20% of the subjects, computed with the random number function `uniform(n)`, resulting in the missing data type MCAR. Now, for only 1218 patients both U1 and U2 were measured. A complete case analysis in these 1218 patients resulted in an increase in urea of 4.0 (0.12) mmol/L, U1: 20.0 (0.17) mmol/L and U2: 24.0 (0.17) mmol/L. All values were similar to the values from the main analysis. The standard errors were somewhat higher due to the lower sample size (n = 1218). When performing the mean substitution method, we found an increase in urea of 4.0 (0.12) mmol/L, U1: 20.0 (0.15) mmol/L and U2: 24.0 (0.14) mmol/L. Again, all values were similar to the values from the main analysis. However, the standard error of U2 is biased downwards, because all substituted data had the same value, resulting in less variance. Although the method of LOCF is not justified when only two data points are present in a study, we applied this method to our simulated situation for illustrative purposes. Using LOCF, when data were MCAR, resulted in an increase in urea of 3.3 (0.11) mmol/L, U1: 20.0 (0.15) mmol/L and U2: 23.3 (0.16) mmol/L. The mean value for U2 and thereby the increase in urea were biased downwards.

What results would the conventional methods give if data were MAR? We, therefore, simulated U2, using the uniform(n) function, with a 1% higher chance of being missing for each mmol/L higher U1. In Table 2, the results of the conventional methods are presented. In the complete case analysis, the values of U1 [19.6 (0.17) mmol/L] and U2 [23.7 (0.17) mmol/L] were biased downwards. This finding can be explained by the fact that subjects with a high mean U1 have a higher chance of having missing data for U2. In a complete case analysis, these subjects are excluded from the analysis, resulting in a lower mean U1. The mean U2 is also lower because a high correlation was specified between U1 and U2 (0.75). Had this correlation been lower, the resulting value of U2–U1 would have been even more biased, as the mean U2 would move towards the mean of the main analysis. In the mean substitution analysis, the values of U2 [23.7 (0.13) mmol/L] and thereby also U2–U1 [3.7 (0.12) mmol/L] were biased downwards. Again, this is explained by the fact that subjects with a high U2 have a higher chance of having missing data. Therefore, substitution of the mean U2 will result in an underestimation. LOCF again resulted in biased values for U2 and U2–U1, 23.2 (0.15) mmol/L and 3.2 (0.11) mmol/L, respectively.
The simulation studies presented above clearly demonstrate that the conventional method of complete case analysis gives unbiased estimates and standard errors—only slightly higher due to a smaller sample size and thereby loss of power—when data are MCAR, but not when data are MAR. With mean substitution, the standard errors are biased downwards in both types of missing data due to a higher number of values centred on the mean. Loss of statistical power and/or biased results makes neither method the method of choice. Furthermore, it is likely that, in clinical research, the two types of missing data are mixed and it is very difficult to assess which type is (most) present. When using the method of LOCF, urea levels were kept constant between the start and end of pre-dialysis care. The analysis thereby resulted in too small an increase in urea in both types of missing data and is also not the preferred method. This dilution would be less prominent when more data points were available, so that missing urea values at the end of pre-dialysis care can be replaced by a value measured at a closer data point. It can be concluded that all three conventional methods have limitations and are not preferred for handling missing data. In the next sections, we discuss the method of multiple imputation and when this method can be used.

**MULTIPLE IMPUTATION: EXPLANATION OF THE METHOD**

Multiple imputation is a method in which missing data are predicted based on known data and on the pattern of missing data [8]. In our view, this method builds on and improves the idea of mean substitution. When using mean substitution, missing values are replaced with the overall mean and imputed values are treated as ‘real’ observed values. Obviously, this approach is too simplistic and generally leads to biased results as in reality some patients have higher values than the mean and other patients have lower values. The overall mean is, therefore, not the best estimate for all of them. An improvement on simple mean substitution could be made by stratifying by age, for example, and substituting the mean among young patients for missing values among young patients and that among old patients for missing values among old patients. This line of reasoning can be followed further, by stratifying by more relevant variables, and substituting a variety of mean values in a multitude of subgroups. In principle, the method of multiple imputation thus makes strata based on the variables known by creating a prediction model for the missing variable. Each patient falls within a stratum based on his or her known characteristics. A patient’s missing value is then replaced by the mean of that stratum. However, in reality, there is variation around the mean, even in small strata. If you impute the same value for all patients within a stratum, the variation will be too low. Therefore, a random error component should be added to the imputed mean value. This error component should be randomly drawn from the error distribution (i.e. from the residuals) from the prediction model. In the above example, in patients on pre-dialysis care with missing U2, their missing U2 would be replaced by a predicted value of U2 based on the available U1. Usually, it is possible to predict the missing U2 values based on more variables, such as age and sex. If the patients with missing U2 are also old and male, it makes sense to replace their missing value with a value similar to that of old male patients. It is very important to think about the known variables you want to include in the model to predict the missing value of interest. There are several strategies you can follow in this respect [6], but generally all variables (continuous or categorical) used for your final analyses should be included. Inclusion of the outcome is also necessary because reasons for missing data are usually related to the outcome status [9]. It is important to keep in mind that depending on the statistical software used, difficulties can arise when continuous variables are not normally distributed—which can be solved with transformation—or when categorical variables need to be imputed. Another important thing to keep in mind is that the imputation model will not always perform better and may even lead to biased results when including large numbers of variables, especially, when these variables have many missing values themselves. For a more in-depth description of these two important computational issues, we refer to other literature [10, 11].

Performing the imputation procedure only once is not to be recommended. With single imputation, imputed values are treated as ‘real’ observed values and, as mentioned earlier, the standard error will be underestimated. Single imputation resembles mean substitution conditional on other known data and, therefore, will have the same problem with underestimating standard errors. The imputation procedure should be repeated several times resulting in multiple imputation data sets. Each imputation takes a random draw from the error distribution and adds this to the predicted mean value. How many times this procedure should be repeated depends on the rates of missing data. It has been claimed that 3–10 imputations are sufficient to obtain an adequate efficiency [12, 13]. Only when the percentage of missing data is very high (>30%), more imputations are recommended. After performing multiple imputations, the analyses of interest are performed within each individual imputation set. The obtained estimates and standard errors of each imputation set are then pooled together into one overall estimate and standard error. For the overall estimate, the average is taken of all separate estimates. The overall standard error consists of the within- and between-imputation variance, based on Rubin’s rules [12]. With multiple imputation, the
uncertainty of the separate imputations is taken into account by incorporating the between-imputation variance, which when compared with mean substitution or single imputation results in a less biased standard error. The combination of predicting missing values, based on other known information, and taking into account imputation uncertainty does not support the term ‘data creation’ introduced by critics of the method. Figure 1 presents a more graphical representation of the method of multiple imputation.

MUTLIPLE IMPUTATION: WHEN TO USE?

In this section, we demonstrate how multiple imputation affects the results when data are either MCAR or MAR. We used the same example of urea levels during pre-dialysis care (U1 and U2) as described before. In this example, missing U2 values were imputed based on the known U1 values (n = 1500). We performed five imputations, because U2 data were only missing for ~20% of the patients. Tables 1 and 2 show that multiple imputation resulted in the same values and standard errors when compared with the main analysis, for both MCAR and MAR. Our simulation results show that multiple imputation performs well in both types of missing data and is, therefore, the preferred method in clinical practice where mixtures of missing data types are common.

MULTIPLE IMPUTATION IN THE FIELD OF NEPHROLOGY

Studies performed in the field of nephrology often aim to estimate the association of a determinant with an outcome, such as the start of dialysis or mortality. Unfortunately, quite often the outcome cannot be established for all patients due to study drop-out or reaching the end of the study period. A complete case analysis is not the preferred method because of power problems and possible biases. Time-to-event analysis estimates an incidence rate and can handle these missing outcomes through censoring [14]. All patients can be included in this analysis because follow-up time until the event of interest or the moment of censoring is taken into account. If time-to-event data are not available, or the assumption of non-informative censoring is violated, multiple imputation could be used to impute these missing outcomes. Deo et al. [15] investigated how many trials performed in patients on renal replacement therapy without time-to-event data available, reported a method to impute missing outcome data, which they found were only 5 of the 110 trials (5%). However, none of these studies used the technique of multiple imputation.

Furthermore, when routinely collected data are used, which is often the case in the field of nephrology, the determinant of interest or the confounders you want to adjust for can have missing values. The reason for these missing data is often based on other known patient characteristics. For example, when data from a national registry of patients on dialysis are used to develop a prediction model for the risk of mortality, it is not surprising that many patients have missing data on comorbidities and laboratory values. Wagner et al. [16] found that patients with missing data (~50%) have a higher risk of mortality, were older, more likely to be white, and had more often an uncertain diagnosis of primary kidney disease, resulting in missing data type MAR. Excluding these patients from the analyses can lead to biased results. Therefore, the authors of this article performed multiple imputation as a sensitivity analysis. Another article, by Collier et al. [17], also used the technique of multiple imputation as a sensitivity analysis to check the robustness of their main complete case analysis. They investigated the association of co-morbidities with survival in patients on renal replacement therapy. The association found when using the method of multiple imputation was less strong compared with the association found when using a complete case analysis. This dilution is probably due to the fact that patients excluded from the main complete case analysis had more co-morbidities.

In summary, only a few years ago the multiple imputation method was scarcely used in the field of nephrology as shown by Deo et al. Currently, a full text search of ‘multiple imputation’ on the website of Nephrology Dialysis Transplantation resulted in seven hits between 2006 and 2008. However, this number has been rising during the last few years as there were 18 hits between 2009 and 2011. And although it has been applied as a sensitivity analysis in the study of Wagner et al. and Collier et al., hardly any study uses it in its main analyses with full attention to the imputed results. We recommend to present results the other way around, so multiple imputation results as main analysis and complete case results as a sensitivity analysis. This and other important recommendations on how to report analyses using the method of multiple imputation can be found in the article of Sterne et al. [18].

CONCLUSION

We can conclude that the multiple imputation method gives unbiased estimates and standard errors, provided that data are either MCAR or MAR, and the prediction model specification is adequate, which may require the help of a statistician. The conventional method complete case analysis only gives unbiased results—and a somewhat higher standard error due to the loss of power—when data are MCAR. In both types of missing data, mean substitution and LOCF give biased results. In clinical practice, combinations of missing data types are most common and it is difficult to assess which type(s) is present for which patient. We can only theoretically imagine which types are present. However, overall multiple imputation performs better than the conventional methods and is, therefore, the method of choice. In the field of nephrology, multiple imputation is not used very often, and we encourage authors to do so in the future. Especially, because the method is relatively easy to perform in many statistical packages, such as recent versions of SPSS, STATA and SAS. One problem about multiple imputation is that some people feel that the method is making up data, which is not scientifically ethical. With this educational article, we hope to have demonstrated that the multiple imputation method predicts data based on the known
variables with the incorporation of missing-data uncertainty, which will lead to estimates closest to the truth.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES


Received for publication: 20.11.2013; Accepted in revised form: 5.4.2013