Intensive Care Med
DOI 10.1007/s00134-014-3624-x

STATISTICAL EDITORIAL

S. Chevret
S. Seaman
M. Resche-Rigon

Multiple imputation: a mature approach to dealing with missing data

Received: 9 December 2014
Accepted: 16 December 2014

Electronic supplementary material The online version of this article (doi:10.1007/s00134-014-3624-x) contains supplementary material, which is available to authorized users.

S. Chevret (✉) · M. Resche-Rigon
Service de Biostatistique et Information Médicale,
Hôpital Saint-Louis, AP-HP, ECSTRA Team, Inserm UMR-1153,
Université Paris Diderot, 1 rue Claude Vellefaux,
Paris 75010, France
e-mail: sylvie.chevret@univ-paris-diderot.fr

S. Seaman
MRC Biostatistics Unit, Institute of Public Health,
Robinson Way, Cambridge CB2 0SR, UK

Missing values in clinical studies are almost unavoidable. When analyzing such data, the standard response is to exclude the patients with missing data. This is known as 'complete case analysis' (CCA) and has been shown to be the leading strategy in the epidemiology [1] and intensive care unit (ICU) literature [2]. However, if the excluded patients are not a representative subsample from the whole sample, their exclusion can lead to bias and loss of precision in estimation, both of which can, for example, adversely affect the performance of predictive risk models in the ICU (Supplementary 1). To deal with this issue, numerous imputation methods have been developed. The simplest method is "simple imputation." This involves replacing each missing value with a single value, such as the mean of the observed data [3]. Thereafter, all patients present in the sample can be included in the analysis. The simplicity and ease of implementation of this method make it attractive. However, it tends to distort the distributions of variables and the associations between them, which can lead to biased estimation. Also, because the analysis treats imputed missing values in the same way as observed values, uncertainty is underestimated, leading to confidence interval (CI) and p values that are too narrow/small. To address these limitations and to make full use of all the information in the sample, multiple imputation (MI) methods have been proposed [4, 5]. These are nowadays considered one of the best methods for analyzing data sets with missing values. Nevertheless, a gap exists between the methods the biostatistical literature recommends and those that are actually applied in the ICU literature [6]. This is illustrated by the finding in a recent review that only one article had used MI [2]. Recently, another ICU paper reported the use of MI, but the only detail given was the software used [7]. Here, we aim to provide practitioners with a short, accessible insight into the theoretical underpinnings of this powerful and flexible approach, at least to understand the basics of MI. As with all statistical techniques, it is important to understand the underlying assumptions and limitations of MI in order better to be able to assess the credibility of results obtained from it and to implement it well.

The aim of MI is to provide unbiased estimates and valid standard errors and CIs for these estimates. It is thus named because each missing value is replaced by a set of plausible values, thus giving rise to multiple imputed data sets. This use of a set of plausible values for each missing variable, rather than just a single value, as in simple imputation, allows the uncertainty in the true, unknown value of the missing variable to be reflected [4, 5]. These multiple imputed data sets are then analyzed in a way that accounts for this uncertainty, thus giving valid standard errors and CIs. MI involves three steps: (1) imputation of missing values from a so-called "imputation model" repeated m times, which results in the m complete imputed data set; (2) the fitting of an "analysis model" (i.e., the model of interest) to each of the m imputed data

Published online: 13 January 2015

sets separately; (3) pooling of the m sets of estimates thus obtained to give an overall set of estimates and corresponding standard errors (Fig. 1). When the missing data are missing completely at random (MCAR) or missing at random (MAR) [8] (see Table 1) and the imputation model is correctly specified, this three-step process results in valid statistical inferences, that is, unbiased estimates with valid CIs [5].

The first step of imputation is the most difficult. It involves specifying an imputation model and using the observed data to estimate the associations between the variables in this model. These estimated associations are then used together with the observed data on each individual to generate values for any missing variables on that individual that are consistent with his or her observed data and the assumed imputation model. The choice of imputation model depends on the type of missing data and also on the form of the analysis model (linear, logistic or Cox model, etc.) and the set of predictors included in the analysis model. One popular imputation model assumes that the set of variables with missing values are multivariate normally distributed [11]. This implies that for each variable missing on each individual, a set of imputed values should be sampled from a normal distribution with mean and variance depending on that individual's observed data. This imputation is commonly performed using a computer algorithm known as Markov chain Monte Carlo (MCMC) [9, 10]. Although the assumption that the data are multivariate normally distributed is often violated (e.g., when categorical covariates are missing), this approach is often robust to such violation. Nevertheless, to allow more flexibility, it can be convenient to separate the imputation process into a series of univariate imputations, one for each incomplete variable and applied iteratively. This method, known as "MI by chained equations" (MICE) or "fully conditional specification," involves specifying a separate model for the distribution of each variable given all the other variables [10, 12]. It has the appeal of enabling different models to be specified for different types of variable, e.g., linear regression for imputing a continuous variable, logistic regression for imputing a binary variable, Poisson regression for a count variable, etc., and it is sometimes preferred for this reason. Once the choice of imputation method has been made, selection of variables that will contribute to the imputation process should at least include variables with missing values and variables of the analysis model including the outcome (Supplementary 2).

Finally, as with any statistical method, it is important that analyses using MI are reported in a way that allows readers to assess the adequacy of the methods used [13]. Unfortunately, this is often not the case [2, 7]. Besides reporting the missing data structure (number of individuals with missing data, reasons why if possible, number of

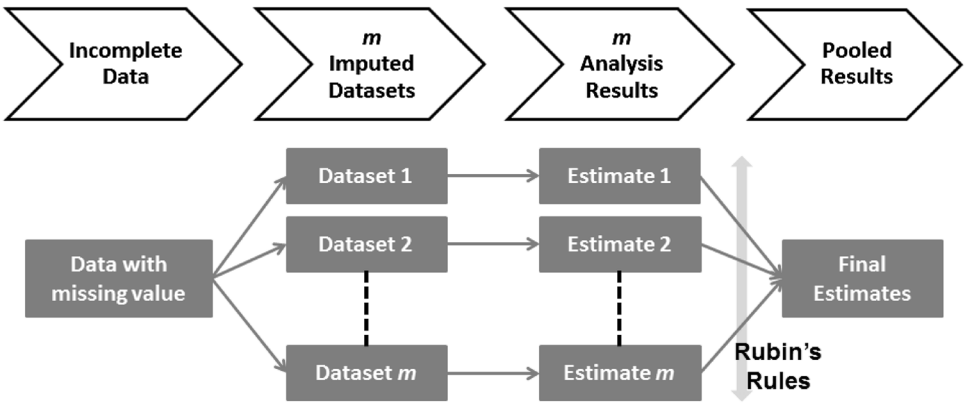

Fig. 1 Scheme of the main steps in multiple imputation. Rubin's rules give overall estimates and corresponding standard errors from the m separate analyses [5]

Table 1 Statistical assumptions for missing data analyses, from Little and Rubin [6]

Classification	Meaning	Examples
Missing completely at random (MCAR)	Missingness does not depend on the data	Laboratory test results were not available because of a machine breakdown. One can reasonably suppose that the breakdown is independent of any patient characteristics. Such missingness mechanisms are rare
Missing at random (MAR)	Missingness depends only on the observed data	Troponin might be more likely measured in patients with chest pain. If the status for chest pain is recorded for all patients, while some troponin levels are missing, then the missingness on troponin is MAR
Missing not at random (MNAR)	Missingness depends on both observed and missing data	Typically a self-report on therapeutic observance. One usually considers that patients with low observance and male are more likely not to fill in self-reports. Although the gender is available, the missingness for observance still depends on the observance value itself

missing values for each variable), it is helpful to show a table comparing the distribution of key variables in individuals with complete and incomplete data and to report (and reference) the MI approach and the imputation models used as well as the number of imputations (Supplementary 3). Moreover, it is interesting to discuss any discrepancies between the results from MI and those from complete case analysis. Finally, ideally, sensitivity analyses would also be carried out to assess the robustness of the results to violation of the MAR assumption [14]. There is room for improvement in such reporting.

Conflicts of interest None.

References

1. Eekhout I, de Boer RM, Twisk JW, de Vet HC, Heymans MW (2012) Missing data: a systematic review of how they are reported and handled. Epidemiology 23(5):729–732
2. Vesin A, Azoulay E, Ruckly S, Vignoud L, Rusinova K, Benoit D, Soares M, Azeivedo-Maia P, Abroug F, Benbenishty J, Timsit JF (2013) Reporting and handling missing values in clinical studies in intensive care units. Intensive Care Med 39:1396–1404
3. Lee CH, Arzeno NM, Ho JC, Vikalo H, Ghosh J (2012) A imputation-enhanced algorithm for ICU mortality prediction. Comput Cardiol 39:253–256
4. Rubin DB (1976) Inference and missing data. Biometrika 63:581–592
5. Rubin DB (1987) Multiple imputation for nonresponse in surveys. John Wiley, New York
6. Pérez A, Dennis RJ, Gil JF, Rondón MA, López A (2002) Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. Stat Med 21(24):3885–3896
7. Debaty G, Maignan M, Savary D, Koch FX, Ruckly S, Durand M, Picard J, Escallier C, Chouquer R, Santre C, Minet C, Guergour D, Hammer L, Bouvaist H, Belle L, Adrie C, Payen JF, Carpentier F, Gueugniaud PY, Danel V, Timsit JF (2014) Impact of intra-arrest therapeutic hypothermia in outcomes of prehospital cardiac arrest: a randomized controlled trial. Intensive Care Med 40(12):1832–1842
8. van Buuren S. Flexible imputation of missing data. Chapman & Hall/CRC, 2012, Boca Raton. ISBN 9781439868249. CRC Press
9. Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, Hoboken
10. van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 18:681–694
11. Schaffer JL (2008) Software for multiple imputation. http://www.stat.psu.edu/~jls/misoftwa.html. Accessed 28 Dec 2014
12. White I, Royston P, Wood AM (2011) Multiple imputation using chained equations: issues and guidance for practice. Stat Med 30(4):377–399
13. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 29(338):b2393
14. Carpenter JR, Kenward MG, White IR (2007) Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat Methods Med Res 16(3):259–275