RESEARCH METHODOLOGY SERIES

Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis

Rolf H.H. Groenwold MD PhD, Ian R. White MSc, A. Rogier T. Donders PhD, James R. Carpenter DPhil, Douglas G. Altman DSc, Karel G.M. Moons PhD

Missing data are a frequently encountered problem in epidemiologic and clinical research.[1,2] One approach is to include in the analysis only those participants without missing observations (complete or available case analysis).[1–4] However, in addition to reducing statistical power, this approach will often result in biased estimates of the associations between covariates and outcomes.[2,3,5,6] Another popular method is to replace missing values using imputation methods.[2] These methods can be applied equally for missing outcomes, missing exposures and missing covariates. A third method, the missing-indicator method, is specifically proposed for missing confounder data in etiologic research.[7,8] This method uses a dummy (1/0) variable in the statistical model to indicate whether the value for that variable is missing, and all missing values are set to the same value. Accordingly, each participant can still be included in the analysis, reducing the loss of statistical power.

In 2005 and 2006, two papers on the missing-indicator method were published, with conflicting conclusions.[3,4] Donders and colleagues focused on missing covariate data in nonrandomized studies and argued that the missing-indicator method would very likely produce biased results.[3] The direction and size of the bias depended on the reason or mechanism of missingness. In contrast, White and Thompson focused on missing baseline covariate data in randomized trials and found that the missing-indicator method produced unbiased estimates of the treatment effect.[4]

Given the popularity of the missing-indicator method among medical researchers, we aim to clarify this apparent discrepancy. We review the missing-indicator method and illustrate its validity, using real data with incomplete covariates from randomized and nonrandomized studies.

Methods to handle missing covariate data

Complete case analysis

The simplest method to handle missing covariate data is to omit from the analysis participants with any missing data (i.e., perform an analysis of available or complete data only). Although this results in loss of statistical power, complete case analysis generally gives unbiased estimates when the participants without complete observations are a representative subset of the study population, a situation known as "missing completely at random."[2,3,5] Most often, however, it is unlikely that data are missing completely at random, but rather missingness of data depends (partly) on observed patient characteristics. For example, in a study of diagnostic accuracy, information on an invasive test can be missing if the diagnosis was already clear enough based on preceding (less invasive) tests. In such situations, complete case analysis may result in biased estimates.

Complete case analysis is unbiased only if missingness is conditionally independent of the

Competing interests:
Please see end of article.

This article has been peer reviewed.

Correspondence to:
Rolf H.H. Groenwold,
r.h.h.groenwold@umcutrecht.nl

CMAJ 2012. DOI:10.1503/cmaj.110977

KEY POINTS

- The missing-indicator method is a popular and simple method to handle missing data in clinical research but has been criticized for introducing bias.
- In nonrandomized studies, the factor or test under study is often related to variables with missing values, in which case the missing-indicator method typically results in biased estimates.
- In randomized trials, the distribution of baseline covariates with missing values is likely balanced across treatment groups, which means the missing-indicator method will give unbiased estimates and obeys the intention-to-treat principle.

This is one in an occasional series that examines controversial aspects of research methods and reporting.

outcome,[9,10] which means that given other patient variables, missingness is independent of the outcome. This is unlikely in the given example.

Imputation

If missingness of a variable is related to observed characteristics but not to unobserved characteristics, the data are (confusingly) called "missing at random."[2,5,6] If data are missing at random, one may use the observed data to estimate the missing value and subsequently replace (impute) the missing value by that estimate. This is usually done using a multivariable regression model, which imputes the missing value with the most likely value, based on all observed patient characteristics, including the outcome.[11] In multiple imputation, uncertainty from the fact that the imputed values were not actually observed, but rather estimated, is accounted for.[2,3,5,6,9] Multiple imputation provides valid estimates and standard errors in many circumstances when missing data are missing at random.[2,3,5,6,11] However, it is a complex technique requiring expertise and appropriate software.[2] Hence, simpler approaches, such as the missing-indicator method, are more appealing.

Missing-indicator method

The missing-indicator method was proposed for missing confounder data in etiologic research[7,8] and has since received much attention in the medical literature.[3–6,10,12] The missing-indicator method does not impute missing values. Instead, missing observations are set to a fixed value (usually zero, but other numbers will give the same results), and an extra indicator or dummy (1/0) variable is added to the analytical (multivariable) model to indicate whether the value for that variable is missing. Consequently, each participant can still be included in the analysis to maintain statistical power.

When using the missing-indicator method to adjust for an incomplete covariate, the estimated association between the independent variable under study (e.g., treatment, risk factor or predictor) and outcome is a weighted average of two associations representing (a) the association between the independent variable and outcome, adjusted for all covariates, among the participants for whom all data were observed; and (b) the association between the independent variable and outcome, adjusted only for complete covariates, among the participants for whom the covariate was not observed. For nonrandomized studies, the second association will typically be biased because it is only partially adjusted for confounding. Furthermore, the first association is based on a complete case analysis, so this association is unbiased only if missingness is conditionally independent of the outcome.[9,10] But, given the nature of nonrandomized studies, in which covariates are commonly mutually related, the missing-indicator method will almost always give biased results.[3]

In randomized trials, however, randomization implies that baseline covariates are balanced across treatment groups and therefore not related to the treatment under study. Hence, unadjusted treatment effects from randomized trials are unbiased. Because of randomization, the distribution of missing values is likely to be balanced across treatment groups as well. Consequently, both the association between treatment and outcome among the participants for whom all data were observed, and the association between treatment and outcome among the participants for whom not all data were observed, will be unbiased.[9] Hence, both complete case analysis and the missing-indicator method will give unbiased estimates. In trials on continuous outcomes, the major reason for covariate adjustment is to increase precision. An important issue, irrespective of the proportion of missingness, is that including all participants for analysis is essential for estimating intention-to-treat effects. Therefore, estimates obtained by using the missing-indicator method will be more precise than those obtained by complete case analysis,[4] and they will also obey the intention-to-treat principle by including all participants randomly assigned to treatment groups.[13]

Missingness of baseline covariates in a randomized trial is not necessarily the same as missing completely at random. In a randomized trial on the effects of a certain treatment for depression, participants who are severely depressed could be more likely to have missing baseline covariates. If the baseline covariate indicates severity of the depression, however, missingness will likely also depend on the value of the baseline variable itself, which is called "missing not at random."[4] But, even if baseline covariate data are missing not at random, randomization implies that missingness is still not related to treatment, so the observed treatment effect will still be unbiased with application of the missing-indicator method.

We have shown that the design of a study, rather than the mechanism of missingness, determines whether the missing-indicator method is valid to handle missing data. A detailed explanation of bias when using the missing-indicator method is provided in Appendix 1 (available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.110977/-/DC1).

Examples

In this section, we illustrate the pros and cons of the missing-indicator method using two case

studies. In both examples, we started with a complete dataset. The results obtained from these complete datasets were considered true associations. New outcome data were created using the true associations. Missingness was then created using a specified mechanism, and three methods to handle missing data were applied: the missing-indicator method, complete case analysis and multiple imputation. We focused on situations with only one covariate with missing values. It is likely that differences between the methods will be more pronounced when more than one covariate has missing values. All analyses were performed in R for Windows (version 2.8.1)[14] or Stata (version 11). Multiple imputation was implemented using multiple imputation by chained equations in R[15] and Stata.[16] This entire process (creating missing values and addressing missing values with the three approaches of analysis) was repeated 1000 times to reduce random variation. The choice for 1000 replicates means that "correct" 95% coverage will likely be between 93.6% and 96.4%.

Example 1: diagnostic study

In a study involving adults in whom deep venous thrombosis was suspected, the diagnostic value of several index tests was assessed.[17] The available dataset consisted of 795 participants with two index tests to predict the presence or absence of deep venous thrombosis: a difference in calf circumference of at least 3 cm (yes/no) and plasma D-dimer level (continuous, log-transformed). The Pearson correlation between the two index tests was 0.32 (95% confidence interval 0.25–0.38).

We created 25% missing values on the variable D-dimer level either in a random sample of the study population (missing completely at random), or with missingness related to calf circumference and deep venous thrombosis (missing at random). In the latter case, the probability of missingness of the D-dimer level was doubled if a participant had either a large difference in calf circumference in combination with deep venous thrombosis or a small difference in calf circumference without deep venous thrombosis. This choice resembled clinical practice, because in an

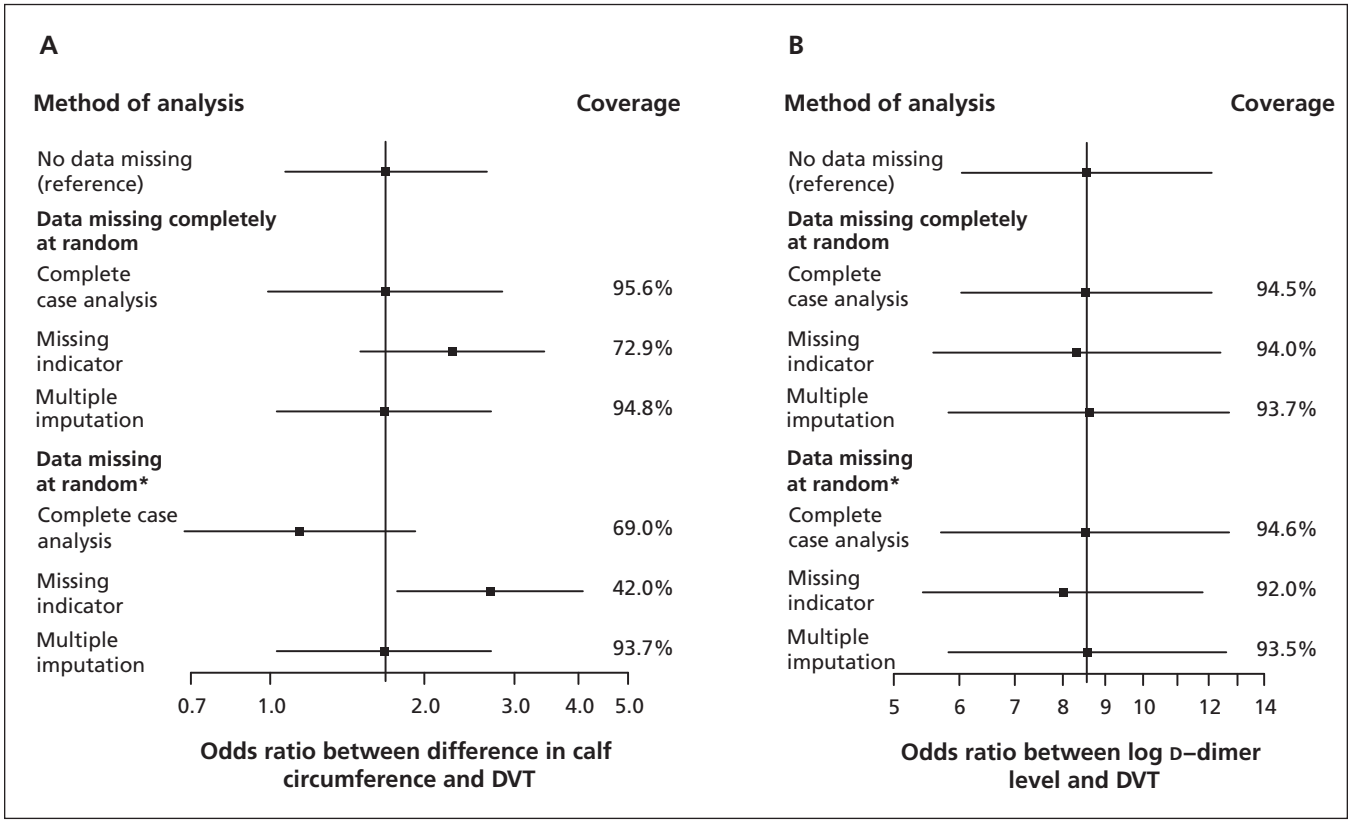

Figure 1: Comparison of analytical methods (complete case analysis, missing indicator, multiple imputation) to handle missing data in a diagnostic study of two index tests — difference in calf circumference (A) and D-dimer level (B) — with missing data on one diagnostic predictor. See Example 1 for more details. DVT = deep venous thrombosis.
Note: Odds ratios (ORs) and 95% confidence intervals (CIs) are based on log-transformed ORs and their standard errors, which were averaged over 1000 simulations, except for the reference method, where they resulted from analyzing the completely observed dataset. Coverage indicates the proportion of estimated 95% CIs in which the true value (based on the reference method) is included (ideally, the coverage is 95%).
***Missingness of the D-dimer level was related to the difference in calf circumference and the diagnosis of DVT.**

instance of a "normal" calf circumference in combination with a "healthy" clinical presentation (low probability of deep venous thrombosis), additional measurement of D-dimer level is likely omitted. Alternatively, a large difference in calf circumference in combination with a clinical presentation of deep venous thrombosis may directly result in referral for reference testing (ultrasonography) and skipping D-dimer measurement.

For multiple imputation we used predictive mean matching with the dichotomized difference in calf circumference and deep venous thrombosis status included in a linear regression imputation model, and 25 imputed datasets were produced.[18] We analyzed each imputed dataset using logistic regression of deep venous thrombosis status on log D-dimer level and dichotomized difference in calf circumference. We combined the estimated regression coefficients and their standard errors using the standard procedures before presenting them as odds ratios.[18]

Use of the missing-indicator method resulted in biased associations between calf circumference and outcome whether missingness was missing completely at random or missing at random (Figure 1A). Complete case analysis provided correct estimates of the associations between both index tests and outcome, and coverage close to the ideal 95% when data were missing completely at random. The results were, however, less precise compared with the other methods (indicated by the larger confidence intervals), because fewer participants were included in the analyses. Complete case analysis yielded biased estimates for calf circumference when missingness was missing at random. Finally, multiple imputation provided unbiased estimates with good coverage regardless of whether the data were missing at random or completely at random. When the proportion of missingness increased, the difference between the methods became larger (results not shown), as shown by others.[19] The observation that the association between the variable with missing values (D-dimer level) and the outcome (deep venous thrombosis) is apparently unbiased (Figure 1B) suggests that using the missing-indicator method for one variable predominantly affects coefficients of the other variables.

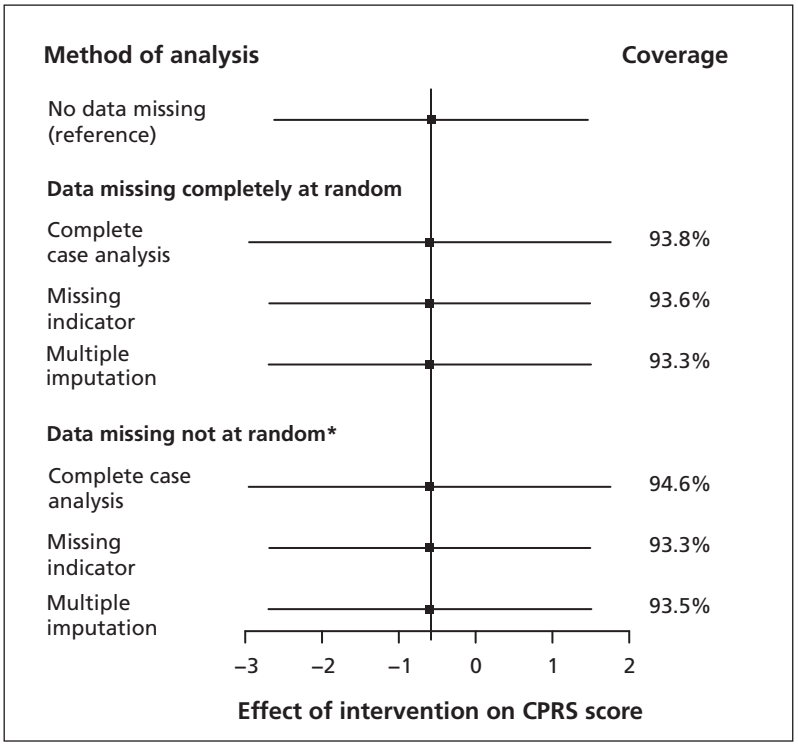

Figure 2: Comparison of methods (complete case analysis, missing indicator, multiple imputation) to handle missing data in a randomized trial on the effect of intervention on Comprehensive Psychopathological Rating Scale (CPRS) score with missing data on one baseline variable.
Note: Estimated effects and 95% confidence intervals (CIs) were averaged over 1000 simulations, except for the reference method, where they resulted from analyzing the completely observed dataset. Coverage indicates the proportion of estimated 95% CIs in which the true value (based on the reference method) is included (ideally, the coverage is 95%).
***Missingness of the baseline variable was related to the value of the baseline variable itself.**

Example 2: randomized trial

A randomized trial compared the effectiveness of intensive management (intervention) and standard management for severely mentally ill patients in the community.[20] For this example, we considered as outcome a measure of psychopathology, the Comprehensive Psychopathological Rating Scale score, and used 595 patients with scores observed at baseline and at two-year follow-up. We estimated the effect of the intervention on the score at two-year follow-up adjusted for the baseline score, using linear regression modelling.

Missingness was created on the covariate baseline score, which was missing completely at random in a 25% random sample of the study population. This reflects the idea that missingness in a randomized trial is likely to be balanced across treatment groups. Alternatively, a situation was created in which patients with more severe psychopathology (indicated by a score higher than the median) were twice as likely to be noncompliant and hence have a missing score at baseline than patients with milder psychopathology (data missing not at random). The procedure was as before, except that the imputation model was a linear regression of baseline score on the two-year follow-up score and randomized group.

Results are shown in Figure 2. For both data missing completely at random and missing not at random, all methods, including the missing-indicator method, yielded correct effect estimates and reasonable coverage. However, confidence intervals

were wider for complete case analysis, reflecting its loss of statistical power. Again, the differences among the methods increased with increasing proportions of missingness (results not shown).

Conclusion

As shown previously, complete case analysis is not a valid method to handle missing data in nonrandomized studies if data are missing at random.[3] In this situation, multiple imputation is the recommended alternative.[2] Although easier to implement, the missing-indicator method typically results in biased estimates in nonrandomized studies (both when data are missing at random or missing completely at random). In randomized trials, the missing-indicator method is a valid method to handle missing baseline covariate data, irrespective of the mechanism of missingness. Even if the proportions of missingness on baseline covariates is low, a complete case analysis does not obey the intention-to-treat principle when adjusting for covariates. An intention-to-treat analysis can also be conducted by simply omitting the incomplete baseline covariate from the model, but this will likely yield estimates that are less precise. The missing-indicator method has the important advantage of obeying the intention-to-treat principle. Although the missing-indicator method was originally proposed for missing confounder data in etiologic research, its use should be limited to randomized trials only.

References

1. Altman DG, Bland JM. Missing data. *BMJ* 2007;334:424.
2. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.
3. Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59:1087-91.
4. White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. *Stat Med* 2005;24:993-1007.
5. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analysis. *Am J Epidemiol* 1995;142:1255-64.
6. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York (NY): John Wiley & Sons; 1987.
7. Miettinen OS. *Theoretical epidemiology: principles of occurrence research*. New York (NY): John Wiley & Sons; 1985.
8. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br J Cancer* 2004;91:4-8.
9. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010;29:2920-31.
10. Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J Am Stat Assoc* 1996;91:222-30.
11. Moons KG, Donders AR, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006;59:1092-101.
12. Knol MJ, Janssen KJ, Donders AR, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J Clin Epidemiol* 2010;63:728-36.
13. White IR, Horton NJ, Carpenter J, et al. Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ* 2011;342:d40.
14. R Development Core Team. *R: a language and environment for statistical computing*. Vienna (Austria): R Foundation for Statistical Computing; 2008. ISBN 3-900051-07-0.
15. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations. *J Stat Softw* 2011;45(3).
16. Royston P. Multiple imputation of missing values. *Stata J* 2004; 4:227-41.
17. Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. *Thromb Haemost* 2005;94:200-5.
18. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011;30:377-99.
19. Janssen KJ, Donders AR, Harrell FE Jr, et al. Missing covariate data in medical research: To impute is better than to ignore. *J Clin Epidemiol* 2010;63:721-7.
20. Burns T, Creed F, Fahy T, et al. Intensive versus standard case management for severe psychotic illness: a randomised trial. UK 700 Group. *Lancet* 1999;353:2185-9.

Competing interests: None declared by Rolf H.H. Groenwold, Ian R. White and A. Rogier T. Donders. Douglas G. Altman is supported by a grant from Cancer Research UK (C5529). James R. Carpenter declares that he or his institution have received funds from the Economic and Social Research Council and the Medical Research Council (MRC) for missing data research, Novartis for statistical consultancy, and GlaxoSmithKline, Pfizer and Boehringer Ingelheim for leading courses on missing data. Karel G.M. Moons is supported by the Netherlands Organisation for Scientific Research (grants 917.46.360 and 918.10.615).

Affiliations: From the Julius Center for Health Sciences and Primary Care (Groenwold, Moons), University Medical Center Utrecht, Utrecht, the Netherlands; MRC Biostatistics Unit (White), Cambridge, UK; the Department of Epidemiology, Biostatistics and HTA (Donders), Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands; the Department of Medical Statistics (Carpenter), London School of Hygiene & Tropical Medicine, London, UK; the Centre for Statistics in Medicine (Altman), University of Oxford, Oxford, UK

Contributors: All authors contributed to the concept and design of the paper. Rolf H.H. Groenwold, Ian R. White, A. Rogier T. Donders and Karel G.M. Moons wrote the first draft of the paper, which James R. Carpenter and Douglas G. Altman critically reviewed. All authors contributed to revisions of the paper. Rolf H.H. Groenwold, Ian R. White and Karel G.M. Moons will act as guarantors.