American Journal of Epidemiology
Vol. 176, No. 4

Letters to the Editor

RE: "DEALING WITH MISSING OUTCOME DATA IN RANDOMIZED TRIALS AND OBSERVATIONAL STUDIES"

Groenwold et al. (1) offer advice on ways to handle missing outcome data in randomized experiments and observational studies. We agree that, regardless of the amount of effort invested in collecting complete data, it is difficult to avoid some missing values. However, we are concerned about the conclusion, stated in the abstract and the discussion, that "complete case analysis with covariate adjustment can and should be used as the analysis of choice more often" (1, pp. 210 and 217). In support of this conclusion, Groenwold et al. presented a simulation study that considered a very basic scenario: binary treatment with a constant treatment effect, 1 fully observed covariate, and 1 partially missing outcome, either binary or continuous, in which the parameters of the missingness mechanism were distinct from those that governed the distribution of the outcomes (2). In the simulations, complete case analysis with covariate adjustment (CCA-CA) yielded results similar to those from the multiple imputation method implemented using multivariate imputation by chained equations (3).

For the simulations under missing-not-at-random mechanisms, CCA-CA and multiple imputation had essentially identical bias and interval coverage. This agreement is obvious because in this basic scenario, both methods used effectively the same inappropriate model. However, the fact that both methods achieved the same wrong result cannot justify a preference for one over the other. Moreover, as the authors note (1), multiple imputation has the advantage of being flexible enough to use alternative assumptions about the missing data mechanism, thereby performing sensitivity analyses (4), which is especially useful when the data may be missing not at random.

For the simulations under missing-at-random mechanisms, CCA-CA performed as well as the multiple imputation method, with both demonstrating zero bias and the correct coverage for intervals. As explained by Little and Rubin (5), if the missingness mechanism is ignorable (i.e., missing at random with distinct parameters), likelihood inference on the observed data produces efficient and consistent estimates. Because in the basic scenario the maximum likelihood estimate of the constant treatment effect is identical to the CCA-CA estimate, the similarity of the results is again obvious. However, this is essentially the only scenario in which the results from CCA-CA correspond to those from multiple imputation. For a more realistic scenario with multiple covariates and/or multivariate outcomes with missingness in more than 1 variable, CCA-CA would not be optimal and could be badly biased. Perhaps this crucial observation could be inferred from the discussion in the article by Groenwold et al. (1); however, we feel that it is important to state this fact explicitly to avoid potential misuse of the stated conclusions.

Moreover, the arguments offered by Groenwold et al. (1) do not apply to situations in which marginal population characteristics are of interest and individual causal effects depend upon covariates. In those cases, CCA-CA is inappropriate unless the missing data are missing completely at random, and even then it is generally not efficient (6).

CCA-CA should be used with caution because it is apposite only in simple missingness cases (5). To us, multiple imputation is a more transparent method because it requires the analyst to be explicit about the procedures used to create the imputations and therefore to acknowledge the underlying assumptions. Hence, in any setting, researchers would be advised to use the multiple imputation method, for both gaining efficiency and reducing bias, with the added advantage of having the flexibility to perform sensitivity analyses. Finally, a recent substantial report by a panel of the US National Academy of Sciences reached conclusions consistent with ours: "Modern statistical analysis tools—such as maximum likelihood, multiple imputation, Bayesian methods, and methods based on generalized estimating equations—can reduce the potential bias arising from missing data by making principled use of auxiliary information available for nonrespondents. The panel encourages increased use of these methods" (7, p. 2).

ACKNOWLEDGMENTS

Conflict of interest: none declared.

REFERENCES

1. Groenwold RH, Donders ART, Roes KCB, et al. Dealing with missing outcome data in randomized trials and observational studies. *Am J Epidemiol*. 2012;175(3):210–217.
2. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581–592.
3. van Buuren S, Groothuis-Oudshoorn K. *MICE: Multivariate Imputation by Chained Equations. R Package, Version 1.21*. Leiden, the Netherlands: Stef van Buuren; 2009.
4. Rubin DB. *Multiple Imputation for Non Response in Surveys*. New York, NY: John Wiley & Sons, Inc; 1987.
5. Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. 2nd ed. New York, NY: John Wiley & Sons, Inc.; 2002.

6. Rubin DB. Characterizing the estimation of parameters in incomplete data problems. *J Am Stat Assoc*. 1974;69(346): 467–474.
7. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press; 2010. (http://www.nap.edu/openbook.php?record_id=12955). (Accessed April 17, 2012).

Victoria Liublinska and Donald B. Rubin
(e-mail: vliublin@fas.harvard.edu)
Statistics Department, Graduate School of Arts and Sciences, Harvard University, Cambridge, MA 02138

DOI: 10.1093/aje/kws215; Advance Access publication: July 31, 2012

THE AUTHORS REPLY

We thank Liublinska and Rubin (1) for their interest in our article on missing outcome data in randomized trials and observational research (2). In medical research, different methods are being used to handle missing data. In our article, we illustrated that in some settings, multiple imputation and complete case analysis with covariate adjustment (CCA-CA) yield the same unbiased results. We evaluated settings in which multiple imputation and CCA-CA utilized exactly the same amount of information on covariates or predictors of missingness. It therefore does not come as a surprise that results were the same for the 2 methods, as Liublinska and Rubin rightfully point out.

Furthermore, we focused on situations in which only outcome data were missing (i.e., no missing data on covariates or treatment status), and these missing outcome data were also missing at random, meaning that missingness depended on observed variables only. We showed that in cases conditional on these observed variables (e.g., after covariate adjustment), missingness is completely at random and a CCA-CA thus yields unbiased estimates, similar to those from multiple imputation.

However, when covariate data are missing as well, it may not be possible to adequately condition on covariates using CCA-CA, whereas the multiple imputation method can address such settings (as long as both missing outcome and missing covariate data are missing at random). We fully agree with Liublinska and Rubin that these settings are actually rather common in nonrandomized epidemiologic research. In randomized trials, outcome data in particular will be missing (due to, for example, loss to follow up), whereas information on baseline covariates or treatment allocation status is commonly not missing.

As we noted in our discussion, postrandomization measurements and secondary end points are typically not included in a regression model relating treatment to the outcome when using CCA-CA because adjustments for such variables typically eliminate the impact. Such variables could, however, be included in the multiple imputation model without necessarily affecting the estimated treatment-outcome association. Hence, when missingness of the outcome also depends on variables that are not routinely included in the adjustment model, the multiple imputation method is preferred. Similarly, when outcome data are missing not at random, assumptions on the missing-not-at-random mechanism can be incorporated in the multiple imputation algorithm, whereas such assumptions cannot be taken into account when performing CCA-CA.

Obviously, CCA-CA yields a conditional treatment effect that can differ from marginal effects when using non-linear models (3). If the interest lies in marginal effects, CCA-CA does not directly provide the desired effect estimate, whereas multiple imputation does. However, marginal effects can be derived from conditional models. Hence, an interest in marginal effects is not an argument against CCA-CA.

Another argument Liublinska and Rubin put forward in favor of multiple imputation is that it is more explicit about the underlying assumptions and is more transparent. Multiple imputation is in essence a 2-stage modeling process in which arguably the joint model is not immediately evident. A CCA-CA, for example, a logistic regression model (binary outcome data), or a mixed model (longitudinal data setting) including all relevant covariates, is often more transparent. In the settings that we and others (4) evaluated, multiple imputation and CCA-CA performed similarly, and arguably the latter method was more transparent.

Nowadays, multiple imputation increasingly becomes the standard approach when confronted with missing data in medical research, often without any reflection on the missingness mechanism. Because of this, one advantage of multiple imputation over CCA-CA, namely the potential to incorporate assumptions on missing-not-at-random mechanisms, is often not exploited. Interestingly, multiple imputation is often applied in situations in which direct modeling (including the same covariates as the imputation model) would perform similarly. Our aim was to draw attention to this issue, even though the range of settings in which CCA-CA could be applied appears limited.

ACKNOWLEDGMENTS

Conflict of interest: none declared.

REFERENCES

1. Liublinska V, Rubin DB. Re: "Dealing with missing outcome data in randomized trials and observational studies" [letter]. *Am J Epidemiol*. 2012;176(4):357–358.
2. Groenwold RHH, Donders ART, Roes KCB, et al. Dealing with missing outcome data in randomized trials and observational studies. *Am J Epidemiol*. 2012;175(3):210–217.
3. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika.* 1984;71(3): 431–444.
4. Siddiqui O. MMRM versus MI in dealing with missing data—a comparison based on 25 NDA data sets. *J Biopharm Stat*. 2011;21(3):423–436.

Rolf. H. H. Groenwold[1], A. Rogier T. Donders[2], Kit C. B. Roes[1], Frank E. Harrell, Jr.[3], and Karel G. M. Moons[1] (e-mail: r.h.h.groenwold@umcutrecht.nl)