American Journal of Epidemiology
Published by the Johns Hopkins Bloomberg School of Public Health 2008.

Vol. 168, No. 4
DOI: 10.1093/aje/kwn071
Advance Access publication June 30, 2008

Special Article

Use of Multiple Imputation in the Epidemiologic Literature

Mark A. Klebanoff[1] **and Stephen R. Cole**[2]

[1] Division of Epidemiology, Statistics, and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Department of Health and Human Services, Bethesda, MD.
[2] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

Received for publication January 14, 2008; accepted for publication March 4, 2008.

The authors attempted to catalog the use of procedures to impute missing data in the epidemiologic literature and to determine the degree to which imputed results differed in practice from unimputed results. The full text of articles published in 2005 and 2006 in four leading epidemiologic journals was searched for the text *imput*. Sixteen articles utilizing multiple imputation, inverse probability weighting, or the expectation-maximization algorithm to impute missing data were found. The small number of relevant manuscripts and diversity of detail provided precluded systematic analysis of the use of imputation procedures. To form a bridge between current and future practice, the authors suggest details that should be included in articles that utilize these procedures.

expectation; imputation; missing data; probability weighting

Missing data are ubiquitous in epidemiologic research; the traditional approach to handling missing data has been ‘‘complete-case’’ analysis (1, 2). In brief, complete-case analyses delete observations with missing information on any studied covariate. Not only is the resultant precision diminished (e.g., widened confidence intervals) because of the reduction in sample size, but bias may be introduced if the data are not missing completely at random. Simple approaches, such as including an indicator variable for missingness, do not in general correct this bias (1). Simple single imputation (deterministic or stochastic) is appropriate in rare cases when the between-imputation variance is vanishing in relation to the within-imputation variance, as well as in other specific situations (3). Methods such as multiple imputation exist that are more generally appropriate and allow asymptotically unbiased estimation under the weaker assumption of missing at random conditional on measured variables. These are now widely available in standard software (e.g., SAS procedure MI (multiple imputation); SAS Institute, Inc., Cary, North Carolina) (4).

Our goal was to catalog the use of multiple imputation in the epidemiologic literature and to determine the degree to which imputed results differed in practice from unimputed results. We searched the full text of articles published over 2 years in the January 2005 to December 2006 issues of the *American Journal of Epidemiology*, the *Annals of Epidemiology*, *Epidemiology*, and the *International Journal of Epidemiology* for the text *imput*. All articles identified were reviewed by the first author to determine whether multiple imputation was used, considering only relevant articles, defined as those presenting original research results, imputing exposures, outcomes, or covariates, and not based on simulated data. The number of eligible articles was derived from the annual summary of articles (considering Original Contributions and Practice of Epidemiology articles) for the *American Journal of Epidemiology* (5, 6) and by manual review of the table of contents for the other three journals.

Numerous articles used a variety of ad hoc methods to impute missing data, such as imputing half the assay detection limit for values below that limit. Several articles studied had such ambiguous descriptions of the methods used for missing data that we were unable to determine the method used (7, 8). Among the 99 articles containing the text *imput* we found 12 relevant articles that used

Correspondence to Dr. Mark A. Klebanoff, Division of Epidemiology, Statistics, and Prevention Research, National Institute of Child Health and Human Development, National Institutes of Health, Department of Health and Human Services, 6100 Executive Boulevard, Room 7B05F, MSC 7510, Bethesda, MD 20892-7510 (e-mail: mk90h@nih.gov).

multiple imputation (9–20). In addition, we also found articles that used inverse probability weighting (7, 21, 22) or the expectation-maximization (EM) algorithm (23) to account for missing data. The degree of detail reported in the 12 papers utilizing multiple imputation was highly variable. Seven papers stated the variables used to impute missing data (9, 10, 13–16, 19). Five papers provided some measure of how imputation changed the results (9, 10, 13, 16, 19); four presented only results obtained by imputed data (12, 14, 15, 17), and three presented unimputed results but stated that imputed results were similar (11, 18, 20). Eight of the 12 papers utilizing multiple imputation stated the number of data sets imputed (9, 10, 12, 13, 15–17, 19). The 16 identified papers that used multiple imputation, inverse probability weighting, or the expectation-maximization algorithm represented less than 2 percent of the 1,105 original research articles published during these 2 calendar years in these journals (i.e., 8/465 papers (1.7 percent) in the *American Journal of Epidemiology*, 3/220 (1.4 percent) in the *Annals of Epidemiology*, 1/172 (0.6 percent) in *Epidemiology*, and 4/248 (1.6 percent) in the *International Journal of Epidemiology*).

We were surprised at how infrequently multiple imputation appeared in published epidemiologic manuscripts given the well-described shortcomings of simpler approaches (1, 2) and relatively easy implementation with widely used statistical software (4). We excluded specific types of papers from the numerator but were unable to apply the same exclusions to the denominator of all published papers; therefore, we have underestimated the use of imputation methods in published papers. Moreover, papers using the methods listed above may not have included the text *imput*. Nevertheless, even if only half of papers published in the epidemiologic literature are relevant by our definition and only half of papers using the listed methods used the text *imput*, the use of these methods is still quite rare. Perhaps use of these methods is more common than our survey found, but journal editors and reviewers are uncomfortable with them, and manuscripts imputing data are less likely to be accepted for publication.

The small number of relevant manuscripts and diversity of detail provided precluded systematic analysis of the use of multiple imputation procedures. To increase the field's comfort with the procedures, we suggest the following considerations for future manuscripts using these methods. Authors who utilize multiple imputation or a similar method should state the fraction or number of observations deleted from the unimputed analysis because of missing data and the fraction or number recovered by imputation. The variables used to impute missing data should be stated. Revealing the set of variables upon which the missing-at-random assumption rests is akin to revealing the set of confounders upon which the assumption of no unmeasured confounding rests. We were surprised that eight of the 12 papers using multiple imputation stated the number of data sets imputed, while only seven stated the variables used to perform the imputations. Finally, we suggest that authors provide primary results from their imputed and complete case analyses, along with the corresponding confidence intervals. We understand that the few investigators routinely incorporating these procedures might view this recommendation as a ''step backward'' in methodological sophistication. However, we see this suggestion as lending a bridge between current and future practice.

The assumptions required by imputation methods are the same as the assumptions of methods already routinely relied upon by epidemiologists. For instance, survival analysis generally requires the assumption that had they been followed, individuals lost to follow-up would have had the same outcome as those who remained under observation. This missing-at-random assumption is enacted implicitly and within levels of exposure in the estimation of Kaplan-Meier curves or conditional on covariates for the estimation of a proportional hazards model. Theory suggests that, if correctly and thoughtfully applied, imputation methods should reduce bias and increase precision in everyday use. We were unable to assess the impact of these methods in practice because of the rarity of use and lack of detail in description. We remain hopeful that inclusion of our suggested details in future publications will demonstrate to the field at large how imputation can reduce bias and increase precision in everyday use and that investigators will become more likely to utilize these methods.

ACKNOWLEDGMENTS

Dr. Klebanoff was supported by intramural funds from the National Institute of Child Health and Human Development.

Conflict of interest: none declared.

REFERENCES

1. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142:1255–64.
2. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol 1991;134:895–907.
3. Weinberg CR, Moledor ES, Umbach DM, et al. Imputation for exposure histories with gaps, under an excess relative risk model. Epidemiology 1996;7:490–7.
4. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat 2007;61:79–90.
5. The Editors. What happens to your manuscript: characteristics of papers published in volumes 161 and 162. Am J Epidemiol 2005;162:1235–6.
6. The Editors. What happens to your manuscript: characteristics of papers published in volumes 163 and 164. Am J Epidemiol 2006;164:1251–2.
7. Mustard CA, Kalcevich C, Frank JW, et al. Childhood and early adult predictors of risk of incident back pain: Ontario Child Health Study 2001 follow-up. Am J Epidemiol 2005; 162:779–86.
8. Parai JL, Kreiger N, Tomlinson G, et al. The validity of the certification of manner of death by Ontario coroners. Ann Epidemiol 2006;16:805–11.

9. Bodnar LM, Tang G, Ness RB, et al. Periconceptional multivitamin use reduces the risk of preeclampsia. Am J Epidemiol 2006;164:470–7.
10. Colt JS, Severson RK, Lubin J, et al. Organochlorines in carpet dust and non-Hodgkin lymphoma. Epidemiology 2005;16: 516–25.
11. Farhang L, Weintraub JM, Petreas M, et al. Association of DDT and DDE with birth weight and length of gestation in the Child Health and Development Studies, 1959–1967. Am J Epidemiol 2005;162:717–25.
12. Ferrie JE, Martikainen P, Shipley MJ, et al. Self-reported economic difficulties and coronary events in men: evidence from the Whitehall II study. Int J Epidemiol 2005;34:640–8.
13. Grievink L, van der Velden PG, Yzermans CJ, et al. The importance of estimating selection bias on prevalence estimates shortly after a disaster. Ann Epidemiol 2006;16:782–8.
14. Maty SC, Everson-Rose SA, Haan MN, et al. Education, income, occupation, and the 34-year incidence (1965–99) of type 2 diabetes in the Alameda County Study. Int J Epidemiol 2005;34:1274–81.
15. Molitor J, Molitor NT, Jerrett M, et al. Bayesian modeling of air pollution health effects with missing exposure data. Am J Epidemiol 2006;164:69–76.
16. Schildcrout JS, Sheppard L, Lumley T, et al. Ambient air pollution and asthma exacerbations in children: an eight-city analysis. Am J Epidemiol 2006;164:505–17.
17. Webb AL, Conlisk AJ, Barnhart HX, et al. Maternal and childhood nutrition and later blood pressure levels in young Guatemalan adults. Int J Epidemiol 2005;34: 898–904.
18. Witt CM, Jena S, Selim D, et al. Pragmatic randomized trial evaluating the clinical and economic effectiveness of acupuncture for chronic low back pain. Am J Epidemiol 2006; 164:487–96.
19. Wood AM, White IR, Hillsdon M, et al. Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. Int J Epidemiol 2005;34: 89–99.
20. Zeka A, Zanobetti A, Schwartz J. Individual-level modifiers of the effects of particulate matter on daily mortality. Am J Epidemiol 2006;163:849–59.
21. Frederiksen H, Hjelmborg J, Mortensen J, et al. Age trajectories of grip strength: cross-sectional and longitudinal data among 8,342 Danes aged 46 to 102. Ann Epidemiol 2006;16: 554–62.
22. Rao RS, Sigurdson AJ, Doody MM, et al. An application of a weighting method to adjust for nonresponse in standardized incidence ratio analysis of cohort studies. Ann Epidemiol 2005;15:129–36.
23. De Stavola BL, Nitsch D, dos Santos Silva I, et al. Statistical issues in life course epidemiology. Am J Epidemiol 2006; 163:84–96.