Techniques for Handling Missing Data in Secondary Analyses of Large Surveys

Diane L. Langkamp, MD, MPH; Amy Lehman, MAS; Stanley Lemeshow, PhD

Objective.—Using an appropriate method to handle cases with missing data when performing secondary analyses of survey data is important to reduce bias and to reach valid conclusions for the target population. Many published secondary analyses using child health data sets do not discuss the technique employed to treat missing data or simply delete cases with missing data. Missing data may threaten statistical power by reducing sample size or, in more extreme situations, estimates derived by deleting cases with missing values may be biased, particularly if the cases with missing values are systematically different from those with complete data. The aim of this study was to determine which of 4 techniques for handling missing data most closely estimates the true model coefficient when varying proportions of cases are missing data.

Methods.—We performed a simulation study to compare model coefficients when all cases had complete data and when 4 techniques for handling missing data were employed with 10%, 20%, 30%, or 40% of the cases missing data.

Results.—When >10% of the cases had missing data, the reweight and multiple imputation techniques were superior to dropping cases with missing scores or hot deck imputation.

Conclusions.—These findings suggest that child health researchers should use caution when analyzing survey data if a large percentage of cases have missing values. In most situations, the technique of dropping cases with missing data should be discouraged. Investigators should consider reweighting or multiple imputation if a large percentage of cases are missing data.

KEY WORDS: hot deck imputation; missing data; multiple imputation; nonresponse bias; secondary analysis; weighting

Academic Pediatrics 2010;10:205–10

As more researchers use secondary analyses of existing survey data to address child health issues, the question of how to handle missing data has become increasingly important. Although the problem of missing data has been recognized and increasingly debated in the statistical literature,[1–4] many child health researchers do not directly address questions about treatment of missing data when performing secondary analyses.[5–9] Discussions about procedures for handling missing data are available in the statistical literature,[1–3,10,11] but there are not many resources written in a nontechnical fashion for substantive child health researchers.[12] In many studies of existing survey data in the pediatric literature, researchers either do not discuss the method used to handle missing data or use complete case analysis (ie, they simply drop cases with missing data).[13–15] In fact, many of the frequently used statistical packages will automatically drop or delete observations with missing values from an analysis. If systematic differences exist between the complete and incomplete cases, reducing the data set in this manner can produce biased results,[16,17] and the conclusions drawn may not be valid for the larger population of interest. Even when the results are not biased, missing data reduces sample size and thus may threaten statistical power. Using different approaches to treatment of missing data can result in different values of key statistics and may result in different conclusions from the analyses. Researchers must be aware of the limitations (or default settings) of their statistical software. Due to the nature of survey data, incomplete responses often cannot be avoided. Participants may skip or refuse to answer a question, or, in a longitudinal survey, participants may not be available or may refuse to participate in subsequent waves of data collection. Hence, an appropriate method for dealing with nonresponse bias should be incorporated into any analysis of this type of data.

Many large national surveys, including the National Health Interview Survey,[18] the National Health and Nutrition Examination Survey,[19] and the National Maternal and Infant Health Survey (NMIHS),[20] use multistage sampling schemes, where different persons in the population have unequal probabilities of being selected to participate in the survey. Data analyses from these samples use sampling weights to take into account the probability of selecting a given person. When using data sets that have been developed with multistage sampling schemes, missing data techniques also must address the issue of complex survey design so that the results may be generalizable to the larger target population.

Several previous studies that analyzed data from the NMIHS used complete case analysis, also known as listwise deletion.[5–9] This technique excludes cases from the analysis if any of the variables under consideration have missing values. A related method for handling missing data, known as available case analysis or pairwise

From the Department of Pediatrics, Akron Children's Hospital, Akron, Ohio (Dr Langkamp); Center for Biostatistics (Ms Lehman) and College of Public Health (Dr Lemeshow), The Ohio State University, Columbus, Ohio.

Presented in part at the Pediatric Academic Societies' Annual Meeting, San Francisco, California, April 29–May 2 2006.

Address correspondence to Diane L. Langkamp, MD, MPH, One Perkins Square, Akron, Ohio 44308 (e-mail: dlangkamp@chmca.org).

Received for publication February 9, 2009; accepted January 20, 2010.

deletion,[1–3] uses all available data to compute each statistic. In other words, different observations may be used in calculating different statistics so that the number of cases varies from one analysis to the next. Pairwise deletion reduces statistical power and increases the risk of bias in a similar way to complete case analysis.

A second group of techniques for handling missing data involves imputation, where a researcher replaces a missing value with either a single estimate (single imputation) or with multiple estimates (multiple imputation).[1] Several commonly used techniques of single imputation include mean substitution, conditional mean estimation, and hot deck imputation.[1,10,11,21] Commonly used techniques of multiple imputation include conditional Gaussian, predictive mean matching, and chained equations.[22]

A third approach to missing data is to attach weights to each subject included in the analysis to represent subjects who were excluded due to missing data. In our previous work using the NMIHS,[23] we redistributed the statistical weights of individuals with missing or incomplete records to individuals with similar demographic characteristics who had complete information. We refer to this as the reweight technique.

The purpose of this investigation was to perform a simulation study to determine which of 4 methods for handling missing data most closely approximates the results using the full data, with varying percentages of cases missing. We chose to compare the reweight technique with 3 commonly used methods for handling missing data: complete case analysis (drop technique), a form of single imputation known as the hot deck technique, and multiple imputation using chained equations. Because the need for using more complex approaches to handle missing data may vary depending upon the percentage of cases with missing observations, the analyses were performed with 10%, 20%, 30%, and 40% of the cases missing. Although procedures for dealing with missing data are familiar to statisticians and research methodologists,[1–3] the lack of such reviews in the child health literature underscores the need to clarify the limitations of techniques that frequently have been used to treat missing data and to disseminate this information in a nontechnical fashion to child health researchers.

METHODS

Data Source

This investigation analyzed data from the live birth component of the 1988 NMIHS (in which birth certificate data are linked to mothers' survey data) and the 1991 Longitudinal Follow-up (LF) Live Birth survey.[20] The 1988 NMIHS used a nationally representative sample of 9953 children born in the United State that year and linked birth certificate data to interviews of mothers. African American and low birth weight children were oversampled. The mothers of 8285 children participated in the both the 1988 and 1991 surveys. We excluded 198 cases where the child was no longer living with the mother and 240 cases where the mother's race/ethnicity was not white, black, or Hispanic. The remaining 7847 cases were available for analysis, but included 571 cases that were missing one or both Center for Epidemiologic Studies Depression Scale (CES-D) scores.

The LF asked the participating mothers about the child's health, behavior, and development, as well as about the mother's own health since the initial interview. At both interviews, the mothers were asked to complete the CES-D.[24] To illustrate the use of these techniques for dealing with missing data, we compared the results for predicting the CES-D score as a function of 3 variables: child chronic illness, child behavior problems, and maternal health.

Variables

Dependent Variable

The CES-D is a 20-item self-administered survey designed to assess current symptoms of depression. The possible range of scores is zero to 60, with a population mean of 9.25 and a standard deviation of 8.58; higher scores indicate greater depressive symptoms. A CES-D score of 16 or higher is associated with substantial depressive symptoms.[24] If the mother's CES-D score was 16 or greater at both interviews, we considered her to have chronic depressive symptoms. The primary dependent variable of our analysis was the presence of chronic depressive symptoms. Approximately 8.8% of the mothers who participated in both surveys were missing one or both CES-D scores.

Independent Variables

The child behavior problems variable was derived from the LF data as described by Civic and Holt.[5] The LF asks mothers to report on the degree or frequency of 5 behaviors or emotional states of the child, which were measured in 3- or 4-point scales. The 5 behaviors/emotional states were as follows: difficulty managing the child, temper tantrums, happiness, fearfulness, and difficulty getting along with others. Civic and Holt[5] reduced each variable to a dichotomous outcome. If the child had a problem in 3 or more of the 5 specified areas, it was labeled as "composite behavior problems." We used this definition of "composite behavior problems" as our variable "child behavior problems."

The maternal health variable was derived from the LF survey that asked the mother to describe her own health as "excellent, very good, good, fair or poor." We recoded this into a dichotomous variable as "excellent, very good, good" or "fair, poor."

Finally, the child chronic illness variable was derived from the LF survey, which asked the mother if her child had any of 21 health problems (eg, deafness, asthma, sickle cell anemia, developmentally delayed, or mental retardation) or "any other serious disorder." If the mother answered affirmatively to any of these questions, then we coded the child as having a chronic illness.

Statistical Analysis

Techniques for Treatment of Missing Data

We compared 4 techniques for handling missing CES-D data. To do so, we created models predicting chronic