Published in final edited form as:

Acad Pediatr. 2010 ; 10(3): 205–210. doi:10.1016/j.acap.2010.01.005.

Techniques for Handling Missing Data in Secondary Analyses of Large Surveys

Diane L. Langkamp, MD, MPH¹, Amy Lehman, M.A.S.², and Stanley Lemeshow, Ph.D.³

¹Department of Pediatrics, Akron Children's Hospital, Akron, Ohio

²Center for Biostatistics, The Ohio State University, Columbus, Ohio

³College of Public Health, The Ohio State University, Columbus, Ohio

Abstract

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 more than 10% of the cases had missing data, the re-weight 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 re-weighting or multiple imputation, if a large percentage of cases are missing data.

Keywords

missing data; non-response bias; secondary analysis; hot deck imputation; weighting; multiple imputation

© 2010 Academic pediatric Association. Published by Elsevier Inc. All rights reserved.

Corresponding Author: Diane L. Langkamp, MD, MPH, One Perkins Square, Akron, OH 44308, Tel: 330-543-8952, Fax: 330-543-6045, dlangkamp@chmca.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Presented in part at the Pediatric Academic Societies Meeting in San Francisco on May 1, 2006

The authors have no conflicts of interest.

INTRODUCTION

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 non-technical 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, i.e., 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 can not 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 (NHIS),¹⁸ the National Health and Nutrition Examination Survey (NHANES)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 deletion⁵⁻⁹. 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 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.²⁹

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,²² we re-distributed 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 four 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 three 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 non-technical 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. 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 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 Center for Epidemiologic Studies Depression Scale (CES-D).²³ 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 three 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.²³ 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 Holt⁵. The LF asks mothers to report on the degree or frequency of five behaviors or emotional states of the child which were measured in 3- or 4-point scales. The five behaviors/emotional states were: difficulty managing the child, temper tantrums, happiness, fearfulness, and difficulty getting along with others. Civic and Holt reduced each variable to a dichotomous outcome. If the child had a problem in 3 or more of the 5 specified

areas, then Civic and Holt labeled this 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 that asked the mother if her child had any of 21 health problems (e.g., 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.

Analysis

Techniques for Treatment of Missing Data—We compared four techniques for handling missing CES-D data. To do so, we created models predicting chronic maternal depressive symptoms (i.e. CES-D scores ≥ 16 at both interviews) as a function of the three independent variables.

Summary of re-weighting technique—In our work on maternal chronic depressive symptoms using the NMIHS and LF, our goal was to obtain nationally-representative estimates; thus, we chose to re-distribute the statistical weights of those mothers with missing CES-D scores to mothers with similar characteristics who had CES-D scores for both the 1988 and 1991 surveys. To do this, 768 subgroups were created using the demographic variables of insurance for prenatal health care, household income in 1988, maternal educational level, mother’s marital status, whether the mother used cocaine in the twelve months before becoming pregnant, the age of the infant when the 1991 survey was conducted, the amount of weight the mother lost after the birth of the infant, mother’s race, and the child’s birth weight. The 768 subgroups were derived from all possible combinations of these demographic variables. These demographic variables were chosen because the cases with complete and incomplete CES-D scores showed significant differences in these variables, and because they each provided somewhat unique information to the relationship (as assessed by examining the correlations among the variables). The statistical weights of those cases in the missing/incomplete group that fell into a subgroup defined by these demographic variables were added up and redistributed equally among those cases in the subgroup with complete CES-D scores. For some subgroups, there were cases with missing CES-D scores but no corresponding cases with complete CES-D scores. In these cases, the cases were assigned to a similar subgroup based on a one level change in one of the variables. This means that the cases were matched on 7 of the 8 variables. The cases with missing CES-D scores then were dropped from further analysis. The 7,276 cases with both CES-D scores were re-weighted to represent the 3,898,922 mothers of children born alive in the US in 1988. After the statistical weights were redistributed, a multinomial logistic regression was performed to determine the factors associated with chronic maternal depression.

Drop Method—For the second method, complete case analysis (drop), we deleted cases where the CES-D score was missing at either or both interviews and used the original weights for the nonmissing observations.

Hot Deck Method—The third technique was the hot deck technique, a form of single imputation, performed using Stata’s hotdeck procedure. We used hot deck within adjustment cells with replacement.¹ For the hot deck technique, we used the same sub-groups or cells as we did for the reweight technique. Each mother who was missing one or both CES-D scores

was matched to a sub-group or cell of respondents with complete CES-D scores who had the same demographic characteristics (covariates) described above. For each mother with a missing CES-D score, the hot deck replaced her missing CES-D score with the recorded CES-D score of a randomly selected respondent within the same sub-group or cell. Stata's hotdeck procedure uses the approximate Bayesian bootstrap method of Rubin and Scheker.²⁴ The imputed dataset was then used in the survey logistic model with the original weights.

Multiple Imputation—The fourth technique was multiple imputation. We used Stata's ICE procedure to perform multiple imputation using chained equations (creating 5 imputed datasets) and then used the MIM procedure to perform logistic regression analysis with the original weights.²⁵ For the multiple imputation, we used the following predictors of maternal depression, all of which were identified in our previous study: child chronic illness, child behavior problems, maternal health, maternal age, maternal level of education, mother's marital status, and whether the pregnancy was planned. (Note that Stata 11 {released July 2009} contains a new, comprehensive MI command that performs the imputation, individual analysis, and pooling of individual analyses steps as the ICE and MIM procedures used in this manuscript.)

Simulation Study Method

In the current project, we performed a simulation study to determine how best to handle missing data. From the original 7847 respondents in the data set we selected all cases that had complete data on CES-D score, maternal age, maternal education, race, and poverty status. From these, we randomly selected a sample of 3,000 cases with complete data. We then removed a varying proportion of cases from this smaller data set in a non-random way so that women who were from a minority, or had a low income were more likely to be dropped. To create the data sets with varying proportions of missing data, we split the data into 4 strata based upon race/ethnicity (Black or Hispanic vs. White) and poor vs. non-poor (<100% poverty level in 1991 or not). For a fixed percentage of missing data (10%, 20%, 30%, 40%) we randomly selected 70% of the missing observations from the first stratum (consisting of individuals who were Black or Hispanic and poor), 15% from stratum 2 (Black or Hispanic and not poor), 15% from stratum 3 (White and poor), and none from the 4th stratum (White and not poor). For example, in selecting 10% of the data (300 cases), 270 cases were randomly selected from stratum 1, 45 cases each from strata 2 and 3, and no cases from stratum 4. The selected cases had their CES-D scores set to missing. The reason for selecting the majority of the observations from the first stratum and none from the fourth stratum was to create a nonmissing sample that would be substantially different in terms of race/ethnicity and poverty level compared to the original data set. We then created 81 subgroups to use for re-weighting. We performed the simulation in the following manner. First, we fit the model with the 3,000 observations that had complete data. This provided the model coefficients for the regression. Then, we fit the same model after deleting 10%, 20%, 30% or 40% of the data and using one of the 4 methods for handling missing data. For each percentage of missing data, we used the four missing data techniques and repeated the simulation 1,000 times to generate 1,000 missing datasets for each technique. Third, we calculated the difference between the parameter estimates from each of the 1,000 simulations and the original parameter estimates from all 3,000 complete observations. The results of the 1,000 trials were averaged. The 2.5th, 50th, and 97.5th percentiles from the distribution were then plotted. All analyses were performed using Stata 7.0 survey commands²⁶ to account for the complex survey design.

RESULTS

Figures 1–4 compare the difference in the model coefficients for the three independent variables (child chronic illness, child behavior problems, and maternal health) and the intercept

for each of the four methods for handling missing data. The superior technique for handling missing data will have model coefficients as close as possible to the model coefficients with the full data (i.e., the difference in the model coefficients will be close to zero and the 95% Confidence Interval for the difference will include zero). Figure 1 presents the results when 10% of the cases have missing data. With 10% missing data, the results of all four methods of handling missing data are similar. The average differences in the model coefficients are slightly larger for the hot deck technique. The confidence intervals for the differences for all four methods contained zero for the three variables. The confidence interval for the intercept did not contain zero in the drop technique. Figure 2 shows the results with 20% of the cases missing data. The hot deck technique tended to slightly underestimate the coefficient for child behavior problems and maternal health and, for maternal health, the 95% CI does not contain zero. Again, the drop technique underestimated the intercept and the confidence interval did not contain zero. Note that the estimates from the re-weight technique were closest to the true estimates from the full model for child behavior problems and maternal health. Figure 3 illustrates the results with 30% of the cases missing data. The drop technique underestimates the coefficient for child behavior problems, maternal health and the intercept. Using the hot deck technique underestimates the model coefficient for child behavior problems and maternal health and the 95% CI does not contain zero for either variable. Figure 4 shows the results when 40% of the cases have missing data. Here the model estimates using the drop technique are all lower than the original estimate. The hot deck technique underestimates the model coefficient for all three variables and the 95% CI does not contain zero for child behavior problems or maternal health. The re-weight and multiple imputation techniques produced estimates that were similar to the full model. We performed the simulation using the four different techniques for handling missing data with three other independent variables and found similar patterns.

DISCUSSION

We recommend that investigators use caution when analyzing survey data if a large percentage of values are missing for the variable of interest. When analyzing survey data with 10% or less of the values missing, there may be a slight bias by simply dropping cases with missing data (complete case analysis or listwise deletion) or using the hot deck technique, although these biases may not be severe. We found that when using the NMIHS and LF, if more than 10% of cases were missing a value for the variable of interest, then deleting cases with missing values from the analysis or using hot deck may produce biased results and the conclusions may not be valid for the larger target population. Re-distributing the statistical weights and multiple imputation appeared to be superior methods when more than 10% of the cases are missing data. The exact cutpoint may be different for different datasets.

We are not aware of any systematic reviews of the pediatric literature that assess published articles about the reporting and use of techniques for handling missing data. However, recent reviews of the psychological literature²⁷ and the education literature³ found that many investigators did not discuss management of missing data. If investigators did discuss their method for handling missing data, the most frequently reported methods were complete case analysis (listwise deletion) and available case analysis (pairwise deletion). These techniques, that drop cases with missing values, introduce bias if the missing data are not missing completely at random; furthermore, these methods also are inefficient, because they do not use all the data that the original investigator worked hard to collect, and, thus, they reduce sample size and statistical power. In 1999, the American Psychological Association Task Force on Statistical Inference specifically discouraged the use of complete case analysis (listwise deletion) and available case analysis (pairwise deletion), stating that these methods are “among the worst methods available for practical applications”.²⁸ Our simulation study emphasizes the need to assess the proportion of cases with missing data when selecting a technique to treat

missing data. In addition, it directs child health researchers to consider more complex, but more appropriate methods to deal with missing data rather than dropping cases with missing data.

The current study has several limitations. First, our review and evaluation of missing data techniques is not exhaustive. The reader is directed to the methodological literature for discussion of other techniques to handle missing data, particularly maximum likelihood estimation.^{3,10,11,29} Second, our work was limited in applying a technique for handling missing data in the NMIHS and LF because we had access only to the public-use data files. If we had access to the original data set, we may have selected different and, perhaps, more meaningful demographic variables (e.g., geographic detail) to create the subgroups or cells that were used to treat the missing data.

In a review of the education literature, Peugh and Enders described missing data as a “dirty little secret” of educational research because it was not discussed in the majority of articles evaluated.³ To move beyond this “secret,” reports of secondary analyses of existing data published in the pediatric literature need to provide explicit information about the technique employed to handle missing data. Furthermore, child health researchers must be encouraged to explore the pattern of missing data and the reasons for missing data. Investigators also must acknowledge the uncertainties that are present as a result of the data being missing. Finally, a major advantage of using nationally representative survey data sets is that they allow researchers to obtain population estimates. If missing data are not addressed in a thoughtful and appropriate manner, these estimates may be biased and the conclusions drawn may not be valid for the larger population of interest. Our work demonstrates that the re-weight technique may be an appropriate method for handling missing data, particularly when more than 10% of the cases have missing values.

Acknowledgments

Funded by grants from The National Institute of Mental Health, #1-R03-MH64060-01A1 (PI, Dr. Langkamp) and the Akron Children's Hospital Foundation (PI, Dr. Langkamp)

References

1. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2nd ed.. Hoboken, NJ: Wiley; 2002.
2. Allison, PD. Missing Data. Thousand Oaks, CA: Sage; 2002.
3. Peugh JL, Enders CK. Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research* 2004;74:525–556.
4. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials* 2004;368–376. [PubMed: 16279275]
5. Civic D, Holt VL. Maternal Depressive Symptoms and Child Behavior Problems in a Nationally Representative Normal Birthweight Sample. *Maternal and Child Health Journal* 2000;4:215–221. [PubMed: 11272341]
6. Deal LW, Holt VL. Young maternal age and depressive symptoms: Results from the 1988 National Maternal and Infant Health Survey. *Am J Public Health* 1998;88:266–270. [PubMed: 9491019]
7. Mandl KD, Tronick EZ, Brennan TA, Alpert HR, Homer CJ. Infant health care use and maternal depression. *Arch Pediatr Adolesc Med* 1999;153:808–813. [PubMed: 10437752]
8. Petterson SM, Albers AB. Effects of Poverty and maternal depression on early child development. *Child Development* 2001;72:1794–1813. [PubMed: 11768146]
9. Chen A, Rogan WJ. Breastfeeding and the risk of postneonatal death in the United States. *Pediatrics* 2004;113:e435–e439. [PubMed: 15121986]
10. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59:1087–1091. [PubMed: 16980149]

11. Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. *Annu Rev Public Health* 2004;25:99–117. [PubMed: 15015914]
12. Croy CD, Novins DK. Methods for addressing missing data in psychiatric and developmental research. *J Am Acad Child Adolesc Psychiatry* 2005;44:1230–1240. [PubMed: 16292114]
13. Hakim RB, Ronsaville DS. Effect of compliance with health supervision guidelines among US infants on emergency department visits. *Arch Pediatr Adolesc Med* 2002;156:1015–1020. [PubMed: 12361448]
14. O'Connor TM, Yang SJ, Nicklas TA. Beverage intake among preschool children and its effect on weight status. *Pediatrics* 2006;118:e1010–e1018. [PubMed: 17015497]
15. Merewood A, Brooks D, Bauchner H, MacAuley L, Mehta SD. Maternal birthplace and breastfeeding initiation among term and preterm infants: a statewide assessment for Massachusetts. *Pediatrics* 2006;118:e1048–e1054. [PubMed: 17015498]
16. Kim S, Egerter S, Cubbin C, Takahashi ER, Braverman P. Potential implications of missing income data in population-based surveys: an example from a postpartum survey in California. *Public Health Rep* 2007;122:753–763. [PubMed: 18051668]
17. Hille ET, Elbertse L, Gravenhorst JB, Brand R, Verloove-Vanhorick SP. Dutch POPS-19 Collaborative Study Group. Non-response bias in a follow-up study of 19-year-old adolescents born as preterm infants. *Pediatrics* 2005;116:e662–e666. [PubMed: 16263980]
18. [Accessed on 1/29/08]. at <http://www.cdc.gov/nchs/nhis.htm>
19. [Accessed on 1/29/08]. at <http://www.cdc.gov/nchs/nhanes.htm>
20. Centers for Disease Control and Prevention. National Maternal and Infant Health Survey. Available at: <http://www.cdc.gov/nchs/nvss/nmihs.htm>.
21. Bennett DA. How can I deal with missing data in my study? *Aust N Z J Public Health* 2001;25:464–469. [PubMed: 11688629]
22. Langkamp DL, Hoshaw-Woodard S, Boye ME, Lemeshow S. Delays in receipt of immunizations in low-birthweight children. *Arch Pediatr Adolesc Med* 2001;155:167–172. [PubMed: 11177092]
23. Radloff LS. The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement* 1977;1:385–401.
24. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J Am Stat Assoc* 1986;81:366–374.
25. For more information, see UCLA Stata Library: Multiple imputation using ICE accessed on 5/13/09 at <http://www.ats.ucla.edu/stat/Stata/library/ice.htm>
26. Statacorp. Stata Statistical Software: Release 7.0. College Station, TX: Stata Corporation; 2001.
27. Bodner TE. Missing data: Prevalence and reporting practices. *Psychological Reports* 2006;99:675–680. [PubMed: 17305182]
28. Wilkinson L. Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist* 1999;8:594–604.
29. 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. [PubMed: 17401454]

Figure 1.

Figure 2.

Figure 3.

Figure 4.