Published in final edited form as:

J Spec Pediatr Nurs. 2017 April ; 22(2): . doi:10.1111/jspn.12179.

Examining Solutions to Missing Data in Longitudinal Nursing Research

Mary B. Roberts, MS*, Mary C. Sullivan, PhD, RN, FAAN§, and Suzy B. Winchester, MA§

*Mary Roberts is a senior data analyst, Memorial Hospital of Rhode Island, Center for Primary Care and Prevention, Pawtucket, RI

§Mary Sullivan is a professor and principal investigator of the longitudinal study, University of Rhode Island, College of Nursing, Kingston, RI

§Suzy Winchester is the project director of the longitudinal study, University of Rhode Island, College of Nursing, Kingston, RI

Abstract

Purpose—Longitudinal studies are highly valuable in pediatrics because they provide useful data about developmental patterns of child health and behavior over time. When data are missing, the value of the research is impacted. The study's purpose was to: (1) introduce a 3-step approach to assess and address missing data; (2) illustrate this approach using categorical and continuous level variables from a longitudinal study of premature infants.

Methods—A three-step approach with simulations was followed to assess the amount and pattern of missing data and to determine the most appropriate imputation method for the missing data. Patterns of missingness were Missing Completely at Random, Missing at Random, and Not Missing at Random. Missing continuous-level data were imputed using mean replacement, stochastic regression, multiple imputation, and fully conditional specification. Missing categorical-level data were imputed using last value carried forward, hot-decking, stochastic regression, and fully conditional specification. Simulations were used to evaluate these imputation methods under different patterns of missingness at different levels of missing data.

Results—The rate of missingness was 16–23% for continuous variables and 1–28% for categorical variables. Fully conditional specification imputation provided the least difference in mean and standard deviation estimates for continuous measures. Fully conditional specification imputation was acceptable for categorical measures. Results obtained through simulation reinforced and confirmed these findings.

Practice Implications—Significant investments are made in the collection of longitudinal data. The prudent handling of missing data can protect these investments and potentially improve the scientific information contained in pediatric longitudinal studies.

Corresponding Author: Mary C. Sullivan, PhD, RN, FAAN, mcsullivan@uri.edu.

Conflicts of Interest:

The authors report no actual or potential conflicts of interest.

Keywords

Methods; Missing Data; Longitudinal Research

Longitudinal studies are highly valuable in pediatrics because children with their families are repeatedly observed, which provides useful data about developmental patterns of health and behavior over time. However, longitudinal studies are frequently riddled with missing data. When data are missing, the value of the research is impacted because the number of observations is reduced which directly effects statistical power, analytical options, confidence in the results, and generalizability of findings (Graham, 2009; Hall et al., 2001; Ibrahim, Chu, & Chen, 2012; Saunders et al., 2006). Research participants in longitudinal studies are asked to complete multiple measurements over many months or years, so missing data occur when children and families miss or partially complete assessments and questionnaires (Hall et al., 2001; Schlomer, Bauman, & Card, 2010). When longitudinal studies are published, missing data are generally relegated to a few sentences, if mentioned at all (Saunders et al., 2006). While researchers carefully strategize approaches to reduce the occurrence of missing data before and during data collection, statisticians have developed techniques to accurately estimate and replace missing data. When appropriate missing data approaches are employed, statistical power and confidence in the results are increased. In this paper, we first outline a 3-step approach to assess and address missing data. Next, we use this approach to tackle missing data from our 23-year longitudinal study of premature infants (Author) and to illustrate how nurse scientists and statisticians determine the best approaches to the missing data problem. The missing data that we illustrate include categorical and continuous level variables that are commonly part of pediatric research and require different analytic approaches.

A practical three-step approach that research teams can follow is the framework for this paper (Schlomer et al., 2010). The three sequential steps are: (Step 1) to determine the amount of missing data; (Step 2) to determine the pattern(s) of missing data; and (Step 3) to compare and model the most appropriate method for missing data. Following recent recommendations, we use mathematical models called simulations for different patterns and levels of missingness in order to better understand the data, assess the different imputation methods, and select the best solution for our sample (Janssen, 2010).

Step 1. Determine the Amount of Missing Data

Attention to the rate of missing data is critical to its impact on statistical power, analysis integrity, estimation bias, and generalizability of the findings (Bennett, 2001; Hall et al., 2001). Acceptable rates of missing data in longitudinal research vary from 5% to 20% (Bennett, 2001; Peng, Harwell, Liou, & Ehman, 2006; Schafer, 1999; Schlomer et al., 2010), although the acceptable consensus among most researchers is 10% (Langkamp, Lehman, & Lemeshow, 2010). The amount of missing data is easily calculated by running frequencies and descriptive statistics using common statistical programs. In general, routines that tabulate frequencies or calculate means and standard deviations will provide the number of observations the statistics are based upon. The difference between the sample size and the

number of observations used for the statistics (mean or frequencies) is the amount of missingness for that particular variable.

Step 2. Determine the Pattern(s) of Missing Data

Scientists refer to three common patterns to classify missing data values in order to show whether the missingness pattern is systematically related to study variables. These patterns have long been known as Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR). For MCAR, no specific pattern of missingness is found (Ibrahim et al., 2012; Myers, 2011; Saunders et al., 2006; Schlomer et al., 2010). The missing data occur randomly and independently from the study variables of interest. For example, in our longitudinal dataset of problem behaviors and health outcomes of former preterm infants, missing data may be independent of the participants' behaviors or health complications (variables/outcomes of interest), and are simply due to transportation issues (i.e., participant moved out of the area).

Missing data values described as MAR indicate that the missingness is related to other variable(s) obtained during data collection, but not related to the measured outcomes of interest (Ibrahim et al., 2012; Myers, 2011; Saunders et al., 2006; Schlomer et al., 2010). For example, MAR missing data in our longitudinal study may be due to a change of parent or caregiver who facilitates their child's continued participation in the study (a collected measure), but not due to problem behaviors or health complications (outcomes of interest).

Finally, missing data values described as NMAR is most concerning because the missingness is related to the outcome variable(s) (Ibrahim et al., 2012; Myers, 2011; Saunders et al., 2006; Schlomer et al., 2010). For example, NMAR missing data in our longitudinal study may be due to participant levels of depression and anxiety (a problem behavior outcome measure) that preclude the completion of this assessment measure.

Once the pattern of missingness is determined, we integrate a technique called simulation for Step 2. Simulations are an analysis technique to enable an examination of the missing data patterns and to improve understanding and raise confidence in the study results. This step involves multiple imputation. Multiple imputation is a statistical technique that uses all observed data to fill in plausible data values for those that are missing (Janssen et al., 2010). It is called multiple because missing values are replaced with several different values. A new data set is created for each of the imputations that includes the original data plus the newly imputed ones. Then, each new dataset is analyzed as a complete case dataset with no missing values. When compared, these simulations are used to evaluate the appropriateness of the imputation method as well as their strengths and weaknesses.

Step 3. Compare and Model the Most Appropriate Method to Address Missing Data

Methods to address missing data fall into one of two categories – variable-level methods and analytic methods. In the variable-level method, a complete dataset(s) is created that can be analyzed using traditional statistical methods such as analysis of variance, repeated measures

analysis, and logistic/linear regression. Although some of these statistical methods have alternative analysis procedures (i.e. repeated measures can be analyzed with MANOVA using a complete dataset while generalized linear models can accomplish this same analysis using a dataset with missing time point observations). However, methods such as structural equation modeling require a complete dataset for the computation of various indicator statistics.

In the analytic methods approaches, missing data imputation are considered nonstochastic (nonrandom method of imputation) or stochastic (random imputation method) (Saunders et al., 2006; Schlomer et al., 2010) and vary in computational complexity. Some approaches invoke bias & reduce variability in the dataset (i.e., last value carried forward) while other imputation methods reduce bias & maximize variability [i.e., stochastic regression (Dong & Peng, 2013)].

Nonstochastic methods

The most straightforward nonstochastic method is last value carried forward (LVCF). LVCF replaces a missing value by using the data value at a previous time point (Bennett, 2001). This method introduces bias and reduces variability with the assumption that change is nonexistent between two time points.

Mean replacement is a commonly used approach due to its simplicity and general acceptability in clinical research (Farhangfar, Kurgan, & Pedrycz, 2007; Myers, 2011; Saunders et al., 2006; Schlomer et al., 2010). The substitution of the missing data value is done by calculating the group mean score of the non-missing values of a measure and replacing the mean score for those that are missing. This method also reduces statistical variability.

Regression imputation uses traditional regression methods to estimate or predict missing values (Saunders et al., 2006; Schlomer et al., 2010). One advantage to this analytic method is that covariates can be added to the regression model to better estimate the missing value. Although this may reduce bias, it also reduces statistical variability.

In hot-decking, missing values are matched to non-missing values based on designated covariates (i.e., gender) with the minimal absolute distance used to determine the best matching. The minimal absolute distance from an a priori determined number of matches is computed. Then, the values for the variable of interest in the non-missing cells are averaged. Finally, the imputed value is used to replace the missing data value (Farhangfar et al., 2007; Hayes, Slater, & Snyder, 2008; Myers, 2011; Saunders et al., 2006).

The above nonstochastic methods, LCVF, mean replacement, regression imputation, and hot-decking assume that missing data are either MCAR or MAR and may introduce statistical bias and reduce variability. Also, potential concerns for Type I or Type II errors should be considered when using these approaches.

Stochastic methods

These methods generate random and multiple datasets based on observed values (Dong & Peng, 2013). These methods reduce the potential for statistical bias and maximize variability. In stochastic regression, a random value is added to an imputed value (Dong & Peng, 2013; Schlomer et al., 2010). The random value is generated from a conditional distribution with maximum likelihood estimates for the unknown distribution parameters based upon the non-missing data (Graham, 2009; Myers, 2011; Schlomer et al., 2010).

Multiple imputation is similar to the aforementioned approaches, except that in this method, multiple data sets are generated through a Monte Carlo technique that fills in missing data. These datasets are analyzed using traditional statistical methods and the results are averaged across the multiple data sets (Dong & Peng, 2013; Ibrahim et al., 2012; Myers, 2011; Saunders et al., 2006; Schlomer et al., 2010). A more contemporary method of multiple imputation, fully conditional specification (FCS) imputation generates a complete data set after a sampling “learning phase” for each variable specified with missingness. Using this method of imputation, the analyst can model interactions between variables, if the analysis of missingness warrants it. The process is repeated in order to obtain multiple complete data sets that are then combined, and analyzed as multiply imputed data sets (Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). FCS imputation can accommodate a variety of data types (White, Royston, & Wood, 2011).

There are many nonstochastic and stochastic analytic methods for scientists to use for missing data values. The non-stochastic methods are LCVF, mean replacement, regression imputation, and hot-decking. The stochastic methods are stochastic regression, expectation-maximum likelihood estimation, multiple imputation, and FCS. The selection of the optimal method for the data is made after consideration of Steps 1 and 2. Research teams consider these approaches in consultation with a data analyst. There are many advantages to implementing a missing data strategy to improve and strengthen research findings.

The Longitudinal Study

The sample of 215 infants and their parents have participated in ten longitudinal follow-up studies since the infants' birth (Author). The a priori study design was a 5-group classification based on neonatal morbidity. The 5 neonatal groups were: full term (healthy with no medical or neurological problems; n=55), healthy preterm (< 37 weeks gestational age (GA) with no medical or neurological problems; n=33), medical preterm (<37 weeks GA with medical diagnosis of bronchopulmonary dysplasia, necrotizing enterocolitis, sepsis, etc.; n=60), neurological preterm (<37 weeks GA with diagnosis of intraventricular hemorrhage grade III or IV, meningitis, hydrocephalus; n=36), and small for gestation age preterm (<37 weeks GA, birth weight below the 10th percentile for gestational age with/without neonatal complications [n= 31; (Lubchenco, Hansman, & Boyd, 1966)]).

Even with high sample retention from birth [4y (N=183, 86%), 8y (N=189, 89%), 12y (N=186, 87%), 17y (N=180, 85%)] there are missing data. Here, we use problem behavior and physical health data from five age points: birth, 18 months, 30 months, and ages 4 years,

8 years, 12 years and 17 years to illustrate the 3-step approach to missing data. The full sample characteristics have been described elsewhere (Author).

The problem behaviors and physical health measures were part of each follow-up assessment where physical assessments were completed on the infants/children and parents completed questionnaires. Neonatal data were collected from medical records. Parents gave informed consent at each age point, while child assent was obtained at ages 8, 12, and 17 years. University and hospital Institutional Review Boards approved the studies at each age.

Measures

Problem behaviors—The Child Behavior Checklist (CBCL) was used as the continuous level variable (Achenbach, 1991; Achenbach, 2003). In the CBCL, problem behaviors are scored on 3-point scale ranging from 0 (not true), 1 (somewhat true), and 2 (very true) by the parents. The internalizing problem behaviors scale consists of withdrawn, somatic complaints, anxious/depressed subscales, while the externalizing scale is comprised of delinquent or rule-breaking behaviors and aggressive behaviors (Achenbach, 1991). The CBCL measure was collected at ages 4 and 8 with internal consistency from .68 to .93 and test-retest reliability of .82 to .95 for scales and subscales (Achenbach, 1991). For the CBCL at ages 12 and 17 years, internal consistency ranged from .78 to .94 and test-retest reliability was .82 to .92. Content, criterion-related, and construct validity have been supported in multiple ways (Achenbach & Rescorla, 2001). Internalizing and externalizing scores are the continuous level outcome variables for this report.

Health status—Health status was the categorical variable used. A complete medical and neurological assessment by a research nurse was conducted at each age point. Health status was classified as normal (no physical abnormalities), suspect (continued respiratory distress, cardiac conditions, referral for hearing or orthopedic), or abnormal (any preexisting physical abnormality, i.e., asthma, allergies, diabetes, autoimmune deficiencies, or orthopedic conditions). Inter-rater reliability was conducted on a quarterly basis and established to be 85% or greater at each age point. Health status (normal versus suspect/abnormal) is the categorical level outcome variable for this report.

Data Analysis for Missing Data

We generated means and standard deviations for the CBCL measures to determine the overall amount of useable data at each assessment. For health status, frequencies were tabulated for all useable data. The difference between our sample size and the number of useable observations is the amount of missingness for each variable.

To determine the patterns of missingness as MCAR, MAR, or NMAR per Step 2, the longitudinal data were examined by selected neonatal and demographic covariates of neonatal group, infant gender, acuity, race/ethnicity and socioeconomic status. These were examined through Chi-square analysis or ANOVA. Little's MCAR Test (Little, 1988) was utilized to test overall data missingness as MCAR.

Data imputation approaches were mean replacement, stochastic regression imputation, multiple imputation, and fully conditional specification (FCS) imputation for the continuous

measures of CBCL scores at ages 4, 8, 12, and 17 years. Mean replacement was calculated using SPSS for Windows (version 21.0) while both the stochastic regression imputation and multiple imputation values were obtained using SPSS AMOS (version 6.0).

LVCF, hot-decking, stochastic regression imputation with rounding, and FCS imputation were used for the categorical measure of health status at ages 18 months, 30 months, ages 4, 8, 12, and 17 years. Hot-decking values were obtained by using the average of the closest 5 complete case observations matched on gender, neonatal group classification, neonatal acuity, and SES. Stochastic regression values were obtained using SPSS AMOS and then rounded to the nearest category level. All imputation values were calculated within each time point. FCS imputation was performed for both the categorical and continuous variables simultaneously using PROC MI in SAS (version 9.4)

Simulations—Simulations are used to evaluate the appropriateness of the imputation method. When comparing missing data imputation approaches, a “reference” dataset is required. The “reference” data sets serves as the true values which are compared to the imputed values in the simulation (Kalaycioglu, 2016). A “reference” data set was generated using the means and standard deviations for the CBCL internalizing and externalizing scores over 4 assessments while preserving the correlations for these measures over time. Similarly, a “reference” data set was generated for the health status measures across 5 assessments using the probability of a suspect/abnormal status at each assessment also while preserving the across time correlations.

Observations were selected as missing and then imputed using the above-mentioned methods. These values were then compared to the reference data for bias, percent relative bias, coverage of 95% confidence intervals, estimate based standard error (SE), empirical SE, and relative mean-squared error (RMSE). Formulas for these measures can be found in the supplemental materials (Appendices). For assessment purposes, target values for bias and relative bias were to be nearly zero whereas coverage and RMSE should be near 100%. Empirical SE and estimate based SE should be of similar magnitude to the complete case data.

For Step 3, imputation methods were compared at both the age-point level and the repeated measures analysis level. Changes in the means and variance estimates were compared to form complete case data estimates for each imputation method used for the CBCL variables. Positive changes indicated an increase in the estimate, while a negative change indicated a reduction in the estimate. For health status, we examined the change in the percent at each level for each imputation method. In both examinations, values near zero indicated little difference from complete case data. At the repeated measures analysis level, imputation methods were compared through the use of effect size and estimates of variance for time. Gender, SES, neonatal acuity, race/ethnicity and health status at one-month were covariates in these analyses. Race/ethnicity was dichotomized into minority (African-American, Hispanic/Latino, and Other) and non-minority (White) due to small sample sizes in several groups. CBCL and health status were analyzed using Multiple Analysis of Covariance (MANCOVA). Health status was analyzed as a bivariate measure (normal versus suspect/abnormal) assuming a normal approximation in order to utilize the MANCOVA framework.

IBM SPSS version 21 (IBM Corp., Released 2012) was used to analyze and compare data imputation approaches.

Results

In Step 1, determination of the amount of missing data, both internalizing and externalizing CBCL scores had 16% – 23% missing data across the four time points (see Figure 1a). Missing data for the health status variable ranged from 0% at 18 months to 28% at 30 months (see Figure 1b). We further investigated the extent of missing data by examining the missingness by the selected covariates (neonatal group, gender, race/ethnicity, SES, neonatal acuity, and health status at 1 month). For internalizing and externalizing CBCL scores, there was not statistically significant missingness associated with the covariates across the time points (data not shown). However, for the categorical-level health status variable, missingness was related to neonatal acuity and 1-month health status through the 8-year time point, but was not significant for ages 12 and 17 years. Because the amount of missing data was greater than the acceptable rate of 10%, Step 2 was initiated (Peng et al., 2006; Schlomer et al., 2010).

For Step 2, determination of the patterns of missing data, internalizing CBCL scores and health status were determined to be MCAR using Little's MCAR Test (see Table 1). However, the missing values for externalizing CBCL scores were significantly related to all covariate variables, thus, the pattern of missing data was determined to be MAR.

Three levels of missingness (10%, 20%, and 30%) were simulated under the following patterns of missingness: MCAR, MAR and NMAR. Each level/pattern of missingness combination was simulated 500 times and compared to the reference dataset.

Observations were selected under MCAR using a uniform random number generator in SAS (version 9.4). For MAR, probabilities of missingness were generated using logistic regression for missingness. It was assumed that the missingness in the variable could be attributed to the baseline characteristics including gender, non-white race, neonatal Hobel, and SES. Since NMAR can never really be known, the data for MAR without covariate adjustment were used as a proxy for NMAR.

Within the simulated data, as the amount of missingness increases, we generally saw an increase in the estimate of variance for time. A notable exception is the CBCL scores where the error term decreased across the levels of missingness for the two single imputation methods (mean replacement, stochastic replacement), but increased for the multiply imputed methods (stochastic, FCS imputation; Tables 2a, 2b, 2c). For the categorical imputation methods, the multiply imputed method, FCS, most closely compared to the complete case data. NMAR had the highest estimates of error across all levels of missingness and methods of imputation, while MCAR had the lowest.

In Step 3, to compare and model the most appropriate method for missingness, minor mean differences were found for mean replacement, regression imputation, multiple imputation and FCS imputation approaches on internalizing and externalizing CBCL scores across all age points (see Table 3a). For the single imputation methods (mean replacement and

stochastic regression), the estimates of variation were reduced as much as 11.8%. Changes for multiple imputation estimates of variation were mostly positive with the greatest change seen at the 4-year age point. Similar results were also seen in the FCS imputation but the overall increase in variance estimation is smaller than with the multiple imputation approach.

With only slight exceptions at the age point level, each imputation method for the health status variable resulted in a decrease in the proportion of participants in the “normal” level (10% or less) and an increase the proportion of participants in the “suspect/abnormal” level (see Table 3b). None of the imputation methods used for health status were clearly superior although the FCS imputation did have the least amount of variation across the levels of health status at each age point.

Using the imputed data, repeated measures ANOVA results are displayed in Tables 2a–2c. Model variance was reduced for both CBCL measures using the mean replacement and stochastic regression imputation methods (see Tables 2a, 2b). When examining the time effect for the CBCL measures, the effect size was largest with the complete case data followed by the multiple and FCS imputation methods. Both methods of single imputation had time effect sizes, which were smaller than the complete case or multiple imputed estimates. For health status, model variance was reduced by both the LCVF, hot-decking, and stochastic regression imputation methods (see Table 2c). FCS imputation, in this case, only mildly inflated the model variance. The effect size was reduced using the four imputation methods.

In addition to the simulation results presented in Tables 2a–2c, means and error estimates for the CBCL measures and medical status variable were evaluated under different imputation methods, levels and patterns of missingness. These results, in general, show increasing relative bias with increasing levels of missingness with the computed coverage decreasing as missingness increased. RMSE also decreased as missingness increased and was lower for the single imputed methods. Multiply imputed data methods (multiple imputation, FCS imputation) showed higher levels of relative bias than single imputation methods but overall had greater levels of coverage and more acceptable RMSE values (see supplemental information for analysis results).

Discussion

The 3-step approach proposed by Schlomer and colleagues (Schlomer et al., 2010) provides a framework to examine missing data in a longitudinal study of premature infants. The missing data variables included the categorical-level variable of health status and two continuous level variables of internalizing and externalizing problem behaviors. We modified the Schlomer et al. (2010) framework to include simulation scenarios to model missing data and examine its patterns. The goal of simulation is not to create data but to prevent the exclusion of observed data (Janssen, 2010). Frequently used data imputation approaches (mean replacement, regression imputation, multiple imputation, LVCF, FCS imputation, and hot-desking) were analyzed and compared using commonly available statistical software packages.

The FCS imputation approach provided the least difference in mean and standard deviation estimates for the CBCL measures. Similar to Finch and colleagues (Finch, 2010), stochastic regression performed reasonably well for the categorical variable of health status. FCS proved to be a flexible approach as it can be used with continuous and categorical data. Its underlying algorithms allow for additional covariates. Little's Test indicated differences in the patterns of missingness between the CBCL internalizing and externalizing scores. This might be understood by considering the parent's perspective in completing the CBCL questions. Externalizing behaviors that direct emotional problems outward (e.g., aggressive behaviors such as bragging, arguing, teasing) would be more observable, while internalizing behaviors that direct emotional problems inward (e.g., withdrawn, sadness, fear, loneliness, worries) may not be as clearly observed.

Patterns of missingness dictate how data should be analyzed. Most traditional statistical analysis methods have an underlying assumption of MCAR (Nakai & Ke, 2011; National Research Council of the National Academies, 2010). The same statistical methods may be used for data that are MAR as long as those variables identified as being related to the missingness pattern are included in the analysis model either through covariates or as potential weighting variables in an inverse probability weighting model (Baraldi & Enders, 2010; National Research Council of the National Academies, 2010).

There are no specific statistical tests for distinguishing the MAR and NMAR patterns (Nakai & Ke, 2011). Although most imputation methods are not appropriate for data in the NMAR pattern, the multiple imputation approach has been found to be the least biased for the NMAR pattern (Baraldi & Enders, 2010). Selection models or pattern mixture models with auxiliary variables to characterize data missingness can be used to analyze NMAR data. To minimize any errors in the pattern determination, sensitivity analyses with various assumptions should be conducted and compared (National Research Council of the National Academies, 2010).

Recent studies have examined the performance of various imputation methods for different patterns of missingness using simulation and have found that findings vary based on the combination of imputation method and pattern of missingness (Ferro, 2014; Jones, Mishra, & Dobson, 2015; McPherson et al., 2014; Romaniuk, Patton, & Carlin, 2014). As demonstrated here, simulations have a place in missing data approaches depending on the level of missingness. In our results, despite model comparisons of 10%, 20% and 30% missingness, the patterns remained the same (Tables 2a–2c). Therefore, one can estimate the amount of missingness and model it for the analysis to answer the research question. These findings underscore the need to carefully consider the patterns of missingness in longitudinal data while instilling a sense of confidence in results arising from the choice of an imputation method. It also suggests that Schomer et al's (2010) 3-step approach should incorporate simulation to better understand how the choice of imputation method is effected by levels and patterns of missingness. As shown in the analysis of the actual data, stochastic regression performed as well as other imputation methods for categorical data. However, under simulation, stochastic regression failed to produce acceptable levels of relative bias, coverage and RMSE. Thus, leading to the conclusion that stochastic regression is not an acceptable imputation method for these data.

In practical application, we will never know the true pattern of missing data. Analysis using responsibly applied imputation techniques and/or alternative analysis methods can, however, capitalize on the information contained in the data as well as inform investigators about future directions for collection of data in longitudinal studies.

How Might This Information Affect Nursing Practice?

A systematic procedure to minimize missing data should be considered a priori (Graham, Taylor, Olchowski, & Cumsille, 2006), and reviewed again when data are analyzed after collection (Baraldi & Enders, 2010; Nakai & Ke, 2011; National Research Council of the National Academies, 2010). When planned in conjunction with a statistician, various scenarios of data missingness are considered and a proposed method for handling missing data can be determined. Significant investments of time, cost, and intellectual capital are made in the collection of longitudinal data. The prudent handling of these missing data can protect these investments and potentially improve the scientific information contained in longitudinal data sets.

This report underscores the need for careful consideration in characterizing and handling missing data in conjunction with analysis methods. Knowledge and expanding statistical software capabilities enable nurse scientists to confidently manage missing data. Examination of longitudinal data sets for missing data and appropriate imputation approaches strengthen study findings, generalizability, and statistical power. Presentation of primary analysis findings and findings from sensitivity analyses are essential to understanding the effect of missingness on study results (Allison; National Research Council of the National Academies, 2010).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This project was funded by the National Institutes of Health, National Institute of Nursing Research [R01 NR003695].

This study would not have been possible without the commitment of the participating children and their families.

References

- Achenbach, TM. Manual for the Child Behavior Checklist/4–18 and 1991 Profile. Burlington, VT: University of Vermont Department of Psychiatry; 1991.
- Achenbach, TM. Manual for the Adult Self-Report for Ages 18–59. Burlington, VT: University of Vermont, Department of Psychiatry; 2003.
- Achenbach, TM., Rescorla, LA. Manual for the ASEBA school-age forms & profiles. Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families; 2001.
- Allison, P. Sensitivity analysis for not missing at random. 2014 Sep 25. <http://www.statisticalhorizons.com/sensitivity-analysis>

- Baraldi AN, Enders CK. An introduction to modern missing data analyses. *Journal of School Psychology*. 2010; 48:5–37. DOI: 10.1016/j.jsp.2009.10.001 [PubMed: 20006986]
- Bennett DA. How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*. 2001; 25(5):464–469. [PubMed: 11688629]
- Dong Y, Peng CYJ. Principled missing data methods for researchers. *Springer Plus*. 2013; 2(222):1–17. <http://www.springerplus.com/content/2/1/222>. [PubMed: 23419944]
- Farhangfar A, Kurgan LA, Pedrycz W. A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*. 2007; 37(5):692–709.
- Ferro MA. Missing data in longitudinal studies: Cross-sectional multiple imputation methods provides similar estimates to full-information maximum likelihood. *Annals of Epidemiology*. 2014; 24:75–77. <http://dx.doi.org/10.1016/j.annepidem.2013.10.007>. [PubMed: 24210708]
- Finch WH. Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science*. 2010; 8:361–378.
- Graham JW. Missing data analysis: making it work in the real world. *Annual Review of Psychology*. 2009; 60:549–576. DOI: 10.1146/annurev.psych.58.110405.085530
- Graham JW, Taylor BJ, Olchowski AE, Cumsille PE. Planned missing data designs in psychological research. *Psychological Methods*. 2006; 11:323–343. [PubMed: 17154750]
- Hall SM, Delucchi KL, Velicer WF, Kahler CW, Ranger-Moore J, Hedeker D, ... Niaura R. Statistical analysis of randomized trials in tobacco treatment: Longitudinal designs with dichotomous outcome. *Nicotine and Tobacco Research*. 2001; 3(3):193–202. DOI: 10.1080/14622200110050411 [PubMed: 11506764]
- Hayes, AF, Slater, MD., Snyder, LB., editors. *The SAGE sourcebook of advanced data analysis methods for communication research*. Thousand Oaks, CA: Sage Publications; 2008.
- IBM Corp. *IBM SPSS Statistics for Windows, Version 21*. Armonk, NY: IBM Corp; (Released 2012)
- Ibrahim JG, Chu H, Chen MH. Missing data in clinical studies: Issues and methods. *Journal of Clinical Oncology*. 2012; 30(26):3297–3303. DOI: 10.1200/JCO.2011.38.7589
- Janssen KJM, Donders ART, Harrell FE, Vergouwe Y, Chen Q, Grobbee DE, Moons KGM. Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology*. 2010; 63:721–727. DOI: 10.1016/j.jclinepi.2009.12.008 [PubMed: 20338724]
- Jones M, Mishra GD, Dobson A. Analytical results in longitudinal studies depended on target of inference and assumed mechanism of attrition. *Journal of Clinical Epidemiology*. 2015; 1:11. <http://dx.doi.org/10.1016/j.jclinepi.2015.03.011>.
- Kalaycioglu O, Copas A, King M, Omar RZ. A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies. *Journal of the Royal Statistical Society*. 2016; 179(3):683–706. DOI: 10.1111/rssa.12140
- Langkamp DL, Lehman A, Lemeshow S. Techniques for handling missing data in secondary analyses of large surveys. *Academic Pediatrics*. 2010; 10(3):205–210. [PubMed: 20338836]
- Little RJA. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*. 1988; 84(404):1198–1202.
- Lubchenco LO, Hansman C, Boyd E. Intrauterine growth in length and head circumference as estimated from live births at gestational ages from 26 to 42 weeks. *Pediatrics*. 1966; 37(3):403–408. [PubMed: 5906365]
- McPherson S, Barbosa-Leiker C, Mamey MR, McDonell M, Enders CK, Roll J. A ‘missing not at random’ (MNAR) and ‘missing at random’ (MAR) growth model comparison with a buprenorphine/naloxone clinical trial. *Addiction*. 2014; 110:51–58. DOI: 10.1111/add.12714 [PubMed: 25170740]
- Myers TA. Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures*. 2011; 5(4):297–310. DOI: 10.1080/19312458.2011.624490
- Nakai M, Ke W. Review of the methods for handling missing data in longitudinal data analysis. *International Journal of Mathematical Analysis*. 2011; 5(1):1–13.

- National Research Council of the National Academies. Committee on National Statistics. Panel on handling missing data in clinical trials. Washington, DC: National Academy of Sciences; 2010. The prevention and treatment of missing data in clinical trials.
- Peng, CYJ., Harwell, M., Liou, SM., Ehman, LH. Advances in missing data methods and implications for educational research. In: Sawilowsky, SS., editor. Real data analysis. Charlotte, North Carolina: Information Age Pub; 2006.
- Romaniuk H, Patton GC, Carlin JB. Multiple imputation in a longitudinal cohort study: A case study of sensitivity to imputation methods. *American Journal of Epidemiology*. 2014; 180(9):920–932. DOI: 10.1093/aje/kwu224 [PubMed: 25301814]
- Saunders JA, Morrow-Howell N, Spitznagel E, Dore P, Proctor EK, Pescarino R. Imputing missing data: A comparison of methods for social work researchers. *Social Work Research*. 2006; 30(1):1–31.
- Schafer JL. Multiple imputation: A primer. *Statistical Methods in Medical Research*. 1999; 8:3–15. [PubMed: 10347857]
- Schlomer GL, Bauman S, Card NA. Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*. 2010; 57(1):1–10. DOI: 10.1037/a0018082.supp [PubMed: 21133556]
- Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*. 2006; 76(12):1049–1064. DOI: 10.1080/10629360600810434
- White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. 2011; 30:377–399. DOI: 10.1002/sim.4067 [PubMed: 21225900]

Figure 1a.

Figure 1b.

Figure 1.

Figure 1a. Step 1. Frequency of Missing Data in Child Behavior Checklist Internalizing and Externalizing Scores by Time Point

Figure 1b. Step 1. Frequency of Missing Data in Health Status by Time Point

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Step 2. Determination and Testing for Pattern of Missingness for Outcome Measures by Covariates and Time Point

Covariates						
Outcome Measures	All Birth Variables	Neonatal Group Classification	Gender	Neonatal Acuity	SES	Race
Internalizing CBCL Score *	0.408	0.789	0.789	0.764	0.542	0.789
Externalizing CBCL Score *	0.008	0.001	0.001	0.004	0.009	0.001
Health Status **	0.111	0.416	0.416	0.714	0.154	0.154

Note. Values in cells are Little's MCAR Test *p*-value.

* Child Behavior Checklist (CBCL) internalizing and externalizing scores assessed at 4, 8, 12, and 17 years.

** Health status assessed at 18 and 30 months, 4, 8, 12, and 17 years

Table 2a
Model Level Results for Repeated Measures ANOVA for Simulated and Actual Continuous Data
Step 2. Determination of Appropriateness of Data Imputation Method for CBCL Outcome Measures by Repeated Measures Analysis

CBCL Models	Complete Case	Imputation Method						FCS Imputation
		Mean Replacement	Stochastic Regression		Multiple Imputation			
Internalizing CBCL								
Actual Data								
ES for time	0.044	0.017	0.018	0.018	0.029	0.022	0.022	
Error (time)	61.63	57.03	56.86	56.86	66.54	70.93	70.93	
Error df	342	615	615	615	615	615	615	
Simulated Data Percent Missing								
	10%	20%	30%	10%	20%	30%	10%	30%
Missing Completely at Random								
ES for time	0.128	0.131	0.136	0.131	0.139	0.156	0.175	0.123
Error (time)	58.19	58.43	58.65	55.71	52.56	48.56	51.87	58.91
Error df	425	261	154	642	642	642	642	642
Missing at Random								
ES for time	0.040	0.044	0.048	0.034	0.034	0.035	0.037	0.035
Error (time)	58.95	60.25	60.50	55.97	52.57	48.45	52.46	59.76
Error df	425	276	169	630	630	630	630	630
Not Missing at Random								
ES for time	0.120	0.115	0.111	0.122	0.121	0.126	0.129	0.108
Error (time)	60.09	61.21	61.15	56.62	53.14	48.82	52.92	61.53
Error df	437	288	181	642	642	642	642	642

Table 2b
Model Level Results for Repeated Measures ANOVA for Simulated and Actual Continuous Data
Step 2. Determination of Appropriateness of Data Imputation Method for CBCL Outcome Measures by Repeated Measures Analysis

CBCL Models	Complete Case	Imputation Method					
		Mean Replacement		Stochastic Regression		Multiple Imputation	
		10%	20%	10%	20%	30%	10%
Externalizing CBCL							
<i>Actual Data</i>							
ES for time	0.014		0.006		0.006	0.013	0.006
Error (time)	52.18		53.58		53.65	61.06	67.33
Error df	342		615		615	615	615
Simulated Data Percent Missing							
	10%	20%	30%	10%	20%	30%	10%
Missing Completely at Random							
ES for time	0.008	0.012	0.020	0.007	0.009	0.011	0.007
Error (time)	53.31	53.47	53.50	51.75	49.30	46.06	47.59
Error df	425	261	154	642	642	642	642
Missing at Random							
ES for time	0.021	0.027	0.031	0.013	0.013	0.012	0.013
Error (time)	52.47	51.41	49.67	51.23	48.13	44.30	47.32
Error df	425	276	169	630	630	630	630
Not Missing at Random							
ES for time	0.005	0.007	0.012	0.006	0.008	0.012	0.005
Error (time)	53.84	52.73	50.85	52.16	48.89	44.85	48.02
Error df	437	288	181	642	642	642	642

Note. CBCL = Child Behavior Checklist; ES = effect size; df = degrees of freedom; FCS Imputation = Fully conditional specification imputation

Table 2c

Model Level Results for Repeated Measures ANOVA for Simulated and Actual Categorical Data

Step 2. Determination of Appropriateness of Data Imputation Method for CBCL Outcome Measures by Repeated Measures Analysis

Status Models	Imputation Method											
	Complete Case			LVCF			Hot-decking			* Stochastic Regression		
	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%
Medical Status												
Actual Data												
ES for time		0.015		0.006			0.008			0.008		0.004
Error (time)		0.172		0.142			0.155			0.155		0.179
Error df		605		1025			1025			1025		1025
Simulated Data Percent Missing												
	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%
Missing Completely at Random												
ES for time	0.024	0.031	0.043	0.018	0.015	0.013	0.021	0.022	0.023	0.021	0.020	0.024
Error (time)	0.179	0.180	0.180	0.173	0.165	0.157	0.184	0.189	0.193	0.191	0.207	0.186
Error df	638	347	180	1070	1070	1070	1070	1070	1070	1070	1070	1070
Missing at Random												
ES for time	0.005	0.009	0.016	0.002	0.003	0.003	0.003	0.004	0.005	0.003	0.003	0.005
Error (time)	0.175	0.177	0.175	0.168	0.159	0.149	0.179	0.182	0.185	0.187	0.198	0.179
Error df	726	484	319	1050	1050	1050	1050	1050	1050	1050	1050	1050
Not Missing at Random												
ES for time	0.036	0.041	0.047	0.023	0.021	0.019	0.024	0.023	0.024	0.024	0.022	0.024
Error (time)	0.177	0.178	0.176	0.171	0.161	0.150	0.182	0.186	0.189	0.190	0.201	0.182
Error df	746	504	339	1070	1070	1070	1070	1070	1070	1070	1070	1070

Note. ES = effect size; df = degrees of freedom; LVCF = Last Value Carried Forward;

* Stochastic Regression = Stochastic regression estimates with rounding; FCS Imputation = Fully conditional specification imputation

Table 3a

Step 3. Comparison of Data Imputation Methods by Age Point: Descriptive Statistics for Continuous CBCL Outcome Scores

CBCL Scores by Time Point	Imputation Method											
	Complete Case			Mean Replacement			Stochastic Regression			Multiple Imputation		
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Internalizing CBCL Scores												
4 years	44.9	7.9	44.8	7.5	-0.3	-5.4	44.8	7.5	-0.5	-5.3	44.4	8.7
8 years	50.6	9.6	50.3	9.1	-0.5	-5.7	50.3	9.1	-0.5	-5.7	50.3	9.9
12 years	50.7	11.2	49.9	9.9	-1.5	-11.8	49.9	9.9	-1.5	-11.8	50.1	11.0
17 years	50.9	10.9	50.5	9.8	-0.8	-10.2	50.5	9.8	-0.8	-10.2	50.2	11.2
Externalizing CBCL Scores												
4 years	47.9	8.4	47.8	8.2	-0.2	-2.8	47.8	8.2	-0.2	-2.7	47.6	9.5
8 years	47.6	10.4	48.9	9.7	+2.6	-6.4	48.9	9.8	+2.8	-5.9	48.8	10.6
12 years	48.4	10.6	48.7	9.6	+0.8	-9.4	48.8	9.6	+0.9	-9.2	49.3	10.8
17 years	47.7	10.1	48.3	9.2	+1.3	-9.1	48.4	9.2	+1.5	-8.7	48.6	10.6
FCS Imputation												
4 years											44.7	8.6
8 years											50.5	10.4
12 years											50.1	11.6
17 years											49.9	11.5
% Change M												
4 years											+9.6	+8.7
8 years											+3.4	+7.6
12 years											-1.6	+3.3
17 years											+2.0	+5.1
% Change SD												
4 years											+13.2	+10.0
8 years											+2.2	+0.5
12 years											+1.8	+6.1
17 years											+4.8	+7.4

Note. CBCL = Child Behavior Checklist; FCS Imputation = Fully conditional specification imputation; M(SD) = mean (standard deviation); % Change = percent change in mean and standard deviation from complete case data ("–," "+," change indicates a decrease in estimate from complete case and "+," change indicates an increase in estimate from complete case).

Table 3b

Step 3. Comparison of Data Imputation Methods by Age Point: Descriptive Statistics for Categorical Health Status Outcome

Health Status by Time Point		Imputation Method															
		Complete Case			LVCF			Hot-decking			Stochastic Regression *			FCS Imputation			
		N	%		N	%	% Change	N	%	% Change	N	%	% Change	N	%	% Change	
18 months																	
Normal		91	68.9	137	63.7	-7.6	137	63.7	-7.6	137	63.7	-7.6	137	63.7	-7.6	137	63.7
Suspect/Abnormal		41	31.1	78	36.3	+16.7	78	36.3	+16.7	78	36.3	+16.7	78	36.3	+16.7	78	36.3
30 months																	
Normal		95	72.0	144	67.0	-6.9	153	71.2	-1.1	169	78.6	+9.2	150	69.8	-3.1	150	69.8
Suspect/Abnormal		37	28.0	71	33.0	+17.9	62	28.8	+2.9	46	21.4	-23.6	65	30.2	+7.9	65	30.2
4 years																	
Normal		94	71.2	141	65.6	-7.9	144	67.0	-5.9	148	68.8	-3.4	142	66.0	-7.3	142	66.0
Suspect/Abnormal		38	28.8	74	34.4	+19.4	71	33.0	+14.6	67	31.2	+8.3	73	34.0	+18.1	73	34.0
8 years																	
Normal		76	57.6	120	55.8	-3.1	114	53.0	-8.0	113	52.6	-8.7	124	57.7	+0.2	124	57.7
Suspect/Abnormal		56	42.4	95	44.2	+4.3	101	47.0	+10.9	102	47.5	+12.0	91	42.3	-0.2	91	42.3
12 years																	
Normal		91	68.9	138	64.2	-6.8	133	61.9	-10.2	137	63.7	-7.6	138	64.2	-6.8	138	64.2
Suspect/Abnormal		41	31.1	77	35.8	+15.5	82	38.2	+23.2	78	36.3	+17.1	77	35.8	+15.1	77	35.8
17 years																	
Normal		86	65.2	134	62.3	-4.5	135	62.8	-3.7	134	62.3	-4.5	143	66.5	+2.0	143	66.5
Suspect/Abnormal		46	34.8	81	37.7	+8.3	80	37.2	+6.9	81	37.7	+8.3	72	33.5	-3.7	72	33.5

Note. LVCF = Last Value Carried Forward;

* Stochastic Regression = Stochastic regression estimates with rounding;

FCS Imputation = Fully conditional specification imputation; % Change = percent change in distribution percent from complete case data ("-" change indicates a decrease in estimate from complete case and "+" change indicates an increase in estimate from complete case).