Missing data: a review of current methods and applications in epidemiological research

W. Todd Abraham and Daniel W. Russell

Purpose of review

Researchers inevitably confront missing data. In cross-sectional studies, nonresponse to specific items causes item-level missing data. Longitudinal studies pose a greater likelihood of item nonresponse and introduce unit nonresponse when data for an individual are missing because that person was not available for assessment. The need to adequately deal with missing data remains, regardless of whether missing data result from item nonresponse, participant attrition, or sporadic availability of respondents. The wealth of missing data techniques available to researchers often produces uncertainty regarding which to use. Our purpose is to discuss the applicability of general methods for dealing with missing data and to review current advances associated with specific missing data techniques.

Recent findings

Traditional missing data methods such as complete case analysis often produce bias and inaccurate conclusions. Similar problems extend to single imputation techniques commonly thought of as improvements over complete case methods. Research demonstrates that procedures such as multiple imputation, which incorporate uncertainty into estimates for missing data, often provide significant improvements over traditional methods.

Summary

Recent work suggests that multiple imputation and specific modeling techniques offer general methods for dealing with missing data that perform well across many types of missing data situations. In addition, advances in desktop computers and the development of user-friendly software make these techniques accessible to researchers in all fields. Future research will undoubtedly result in further refinements and extensions of these techniques, making them applicable to difficult but common situations in which missing data arise.

Keywords

missing data, dropout, attrition, multiple imputation

Curr Opin Psychiatry 17:315–321. © 2004 Lippincott Williams & Wilkins.

Institute for Social and Behavioral Research, Iowa State University, Ames, Iowa, USA

Correspondence to Todd Abraham, Institute for Social and Behavioral Research, Iowa State University, 2625 N. Loop Drive, Suite 500, Ames, IA 50010-8296, USA
Tel: +1 515 294 7416; fax: +1 515 294 3613; e-mail: abrahamt@iastate.edu

Current Opinion in Psychiatry 2004, 17:315–321

Abbreviations

FIML	full information maximum likelihood
GEE	generalized estimating equations
LOCF	last observation carried forward
MAR	missing at random
MCAR	missing completely at random
NMAR	not missing at random

0951-7367

Introduction

Any research study faces the possibility of missing data. Even those conducting cross-sectional studies often experience missing data in the form of 'item nonresponse', when participants fail to provide data on particular measures. The probability of experiencing missing data is compounded further in studies employing at least one follow-up assessment, in that data can be missing due to either item nonresponse or 'unit nonresponse', when data for a particular person are not observed because that individual did not participate in the follow-up assessment. Problems associated with both item and unit nonresponse extend naturally to longitudinal studies in that, as the number of follow-up assessments or waves of data collection increases, so does the probability of experiencing missing data. It is this association with the increased probability of missing data that is particularly relevant for epidemiological researchers who typically conduct studies employing longitudinal designs.

A recent review indicated that of the treatment studies appearing in notable psychiatry and substance abuse journals, only about half of the reports mentioned missing data due to unit nonresponse (i.e. dropouts) [1]. Furthermore, less than 20% of these studies incorporated dropouts into the data analyses. Emphasis on the need to deal with missing data has clearly increased in recent years, as has the proliferation of methods and techniques to do so. During preparation of the current review, however, we often encountered studies that dealt with item and unit nonresponse by simply deleting cases with missing data from the analyses. It is likely that the continued practice of complete case analyses results from uncertainty regarding the appropriate use of missing data techniques currently available. The current work attempts to review existing missing data methods, recent developments associated with each method, and applications of specific

methods to epidemiological research. Because advances related to some methods have not appeared during the review period (January 2003–present) we occasionally discuss prior work related to specific methods that we feel is relevant to the understanding and application of those techniques. The body of this review proceeds with a discussion of different types of missing data, specific missing data techniques, and concludes with a discussion of limited but promising options for a specific but common case of missing data.

Types of missing data

Nearly two decades ago, Little and Rubin [2,3] proposed a typology for classifying missing data based on formal statements concerning the probability related to the missing observations. Under this classification, data are missing completely at random (MCAR) if the probability of an observation being missing is unrelated to all observed and unobserved characteristics of study participants within the system under investigation. Under loose interpretations of this classification, researchers often proclaim that MCAR is reasonable based on demonstrating that the probability of missingness is independent of study measures. Although MCAR requires only that the probability of missingness is independent of study variables, the possibility of indirect associations between study variables and unobserved factors presents a problem because data needed to verify the assumption do not exist. This inability to demonstrate independence between the probability of missingness and unobserved factors often makes the MCAR assumption untenable.

Missing at random (MAR) is a second, less restrictive, assumption, wherein the probability of missingness can be dependent on observed variables but cannot depend on unobserved characteristics including the actual missing value. For example, missing data on an outcome assessed repeatedly over time in a longitudinal study can often satisfy the MAR assumption when there is a strong association between previous and current assessments of the outcome variable. Indeed, the MAR assumption is still valid when the probability of missingness depends on the actual missing value, provided some observed factor(s) exists in which the missing value and other variables have no residual association once this factor(s) is controlled [4]. It is important to note, however, that in many types of longitudinal studies missing data at some assessment (Y_i) may be strongly associated with the actual missing value and relatively unassociated with previous assessments (Y_{i-1}, Y_{i-2}, etc.). For example, studies investigating dementia [5•,6] often experience missing data due to the death of participants where previous assessments are relatively uninformative due to the natural progression of the disease. Under circumstances when the probability of missingness depends upon the actual value of the missing data point and not on other observed measures, the MAR assumption is violated and the data are considered to be not missing at random (NMAR). We return to these issues in the context of advances later in this review.

Ad hoc methods

In general, methods that fall under ad hoc procedures alter the actual structure of the data based on observed missingness after the data are collected. The two most common ad hoc methods involve complete case and available case analyses synonymous with listwise and pairwise deletion, respectively. Although complete case and available case analyses are common, use of these methods results in both conceptual and analytic biases. Conceptually, elimination of individuals with missing data points biases study conclusions in that results are obtained from a specific subsample (i.e. those with complete data) which is not representative of the intended study population. In addition, complete case and available case methods violate intention to treat principles currently endemic in epidemiological and substance use research [7–9]. Analytically, these ad hoc methods may negatively influence statistical power because of the reduction in sample size. Although available case analyses also suffer from reduced power, a larger problem involves the calculation of standard errors because the appropriate sample size is indeterminate. Sample size determination under available cases is particularly problematic for specific statistical methods such as structural equation modeling [10•] that involve an observed covariance matrix. A number of recent simulation studies comparing ad hoc methods to other missing data techniques demonstrate that complete cases and available cases analyses are rarely, if ever, appropriate [8,11••,12••,13•].

One notable exception involves the use of various weighting techniques, such as inverse probability weighting, wherein differential selection bias in parameter estimates is corrected using sample characteristics. Specifically, weights adjust parameter estimates to reflect the amount of information provided by each observed case in relation to the intended sample. Discussion of the specific details associated with various weighting techniques are beyond the scope of this work, but excellent sources exist [2,3,14••]. We choose to discuss these techniques in the context of ad hoc methods only because weighting methods necessitate the analysis of complete case data. Although it is certainly true that one often derives corrective weights after data collection, incorporation of weighting methods as a priori design features is possible. For example, researchers employing multi-phase sampling designs may specify the proportion of cases sampled at later stages based on some criterion at an earlier phase of the study before any data are collected [15]. As a general strategy, weighting methods

adequately correct bias due to missing data in complete case analyses. However, these methods provide no information about the nature or impact of missingness. We must stress that this does not imply that weighting is limited in its utility, but rather that choice of a specific missing data technique necessarily stems from one's research questions [16••].

Single imputation methods

Single imputation methods attempt to maintain the original structure of the data by filling in missing observations with plausible estimates. These methods intend to preserve marginal aspects of the data distribution and provide unbiased estimates of parameters such as the mean. Substitution of missing values based on observed means or predicted values from regression models, however, results in an underestimation of the measure's variance because all missing observations are replaced with the same value. As a result, relationships between variables are overestimated and, in the case when values are imputed for covariates, increased strength in the association between covariates and an outcome can mask treatment effects, producing an increase in type II error [17].

At present, there are a number of available techniques that fall under this heading and extensive reviews of these methods are provided elsewhere [4,5•,10•,18••–20••]. One of these techniques, however, commonly appears in the epidemiology literature and hence deserves special attention. Last observation carried forward (LOCF) deals with missing data in longitudinal designs in which a missing value (Y_i) is replaced with the preceding observed value for the same variable (Y_{i-1}). LOCF is usually defended as a conservative approach. However, recent work [21•] presents an informative argument against the conservativeness of this method. In addition to challenges of the assumptions underlying LOCF, comparative studies demonstrate that LOCF tends to overestimate treatment effects and underestimate standard errors, resulting in an increase in type I error rates [8,11••,12••]. Interestingly, these biases exist when data are generated to satisfy the MCAR assumption [11••]. An extensive simulation comparing many of these methods suggests that those methods relying on person-specific data perform better than those that do not [18••]. Although single imputation methods are preferable to complete case and available case analyses [4], additional studies comparing traditional single imputation methods with other methods discussed below suggest that better alternatives to these strategies exist [8,10•,11••,22•,23••].

Model-based methods

For the purposes of this review, we classify approaches as model based if the technique deals with missing data at the point of statistical analysis. Perhaps the earliest model-based methods are those involving generalized estimating equations (GEEs) coveted for their applicability to both continuous and discrete variables. GEE methods remain popular in some fields because they provide corrected standard errors and can account for residual correlations in the data. In addition, these methods allow dependence between the probability of missingness and observed covariates provided that the covariates are unrelated to the variable with missing data, thereby satisfying the MCAR assumption [4,23••,24]. Potential drawbacks to GEE methods involve both their focus on marginal information that, as a consequence, ignores person-specific information [25] and their lack of applicability to MAR and NMAR problems.

Recently, the missing data literature has seen a resurgence of interest in GEE methods. For example, Jung and Ahn [26•] proposed a GEE-based method for determining the sample size needed to achieve adequate statistical power when testing differences in linear slopes that corrects for the impact of MCAR data. Additional advances in this area incorporate existing weighting methods and reformulations of GEE to arrive at extensions of the method to both MAR [27•] and NMAR [28] situations.

A second broad class of model-based methods for dealing with missing data involves a number of techniques that employ maximum likelihood estimation. General maximum likelihood approaches are valued because they theoretically yield unbiased estimates under both MCAR and MAR conditions [22•]. These approaches are exceptionally popular in part because of the ease with which they extend to various missing data problems, including those involving binary variables [29••,30••]. The downside to general maximum likelihood approaches centers on their computational complexity and the fact that they tend to be problem specific, requiring different formulations for different missing data problems [31]. In addition, general maximum likelihood approaches do not impute values for the missing observations, requiring different maximum likelihood formulations for subsequent analyses using the same data set.

Although commonly thought of as a different estimation method, the expectation maximization algorithm simply applies maximum likelihood estimation to a general model [31]. Indeed, many statistical techniques including multilevel models and latent class analysis can be formulated as expectation maximization algorithm procedures [4]. Estimates obtained from the expectation maximization algorithm are typically unbiased and the method provides a data set with imputed values. The model is general enough for adequate application to

binary data [32], and may perform well when conditional independence between categorical covariates and binary outcomes does not exist [33]. A final interesting development concerning the expectation maximization algorithm deals with its applicability to non-normal data. Based on the results of a complex simulation study, an adjusted expectation maximization procedure seems to provide better estimates and more accurate standard errors than a distribution-free estimation method under both MCAR and MAR conditions when the sample size is around 500 cases [34•].

Direct or full information maximum likelihood (FIML) is a specialized case of general maximum likelihood recently developed to deal with missing data in the context of structural equation modeling. Although expectation maximization is also suited to such analyses and estimates from expectation maximization and FIML will converge in linear regression analyses [23••,35••], recent studies demonstrate that FIML generally produces better estimates and more accurate standard errors [22•,23••] than those obtained using the expectation maximization algorithm. Current debate in the literature concerning FIML concerns its ability to estimate standard errors in comparison to multiple imputation (see below). At present, findings are mixed, with some studies demonstrating compatible estimation [22•] and others indicating that standard errors obtained using multiple imputation are more accurate [23••]. Proponents of expectation maximization assert that it is more flexible, allowing for the inclusion of auxiliary covariates, whereas FIML is restricted to only those variables appearing in the model of interest. Although this was indeed the case originally, newly developed methods allow for the inclusion of auxiliary covariates in structural models estimated with FIML [36••]. In addition, recent work has extended FIML to the generation of full information correlation matrices for mixtures of continuous and categorical variables under MAR [37•].

Multiple imputation

All of the imputation methods discussed to this point share one common feature in that they generate a single data set with filled-in values. It is because of this that these methods tend to underestimate variability due to sampling. Rubin's method of multiple imputation [38] (see also Schafer [39]) is designed to address this problem by generating (k) multiple complete data sets reflecting collection of data on (k) random samples. Essentially, multiple imputation uses the expectation maximization algorithm to generate starting values for data parameters (e.g. covariances), and then regresses variables containing missing values on the completely observed variables. Coefficients from these regressions provide predicted values for missing observations that are augmented with values randomly drawn from a distribution of residuals. After imputing the missing data, the algorithm recalculates the parameters and makes random draws from a posterior distribution. This process is repeated numerous times, saving the imputed values for every kth iteration. When complete, multiple imputation yields (k) complete data sets suitable for analysis using normal methods. Analysis proceeds by applying a given statistical method to each of the imputed data sets and summing the results using simple computational formulas. Extensive descriptions of multiple imputation procedures are available in the existing literature [4,10•,19••,31,35••,40•].

The appeal of multiple imputation stems largely from simulation studies demonstrating the accuracy of imputed values [41•], parameter estimates, and standard errors [22•,23••], as well as multiple imputation's utility as a general method for dealing with different missing data situations. Recent applications of multiple imputation demonstrate its ability to handle large data sets [40•], high proportions of missing data (e.g. 40% [31] and 62% [17]), and data with high residual correlations among respondents [42••]. Additional examples of multiple imputation's flexibility include application to Cox regression models [43•] and the imputation of categorical data [40•].

Multiple imputation has generated a great deal of interest among missing data researchers, but it does not provide a panacea. Of particular concern is the fact that multiple imputation typically employs a multivariate normal model when performing data imputation. This raises questions regarding the appropriateness of the method under conditions where normality is an untenable assumption. In addition, the underlying theory of multiple imputation suggests it is a method best suited for reasonably large samples. Although these are valid issues, simulation work suggests that multiple imputation performs well with both small samples and non-normal variables [44]. A second and more recent issue related to multiple imputation concerns the appropriate number of imputations required to accurately mimic sampling variability. For example, Rubin [38] demonstrated that estimate efficiency is often achieved with as few as three imputations and that increasing the number of imputations beyond five provided little benefit. As such, most applications of multiple imputation typically employ between five and 20 imputations. Some have recently argued, however, that these guidelines are based on simulations without a solid theoretical backing and that application of random sampling theory suggests the appropriate number of imputations can easily exceed 500 [45•]. Those advocating multiple imputation have not decisively disputed this claim, nor have those making the claim provided explicit support for their

position. This issue will likely be a topic of debate in the near future.

Options when data are not missing at random

The techniques discussed to this point assume that missing data are either MCAR or MAR and are not considered appropriate when the probability of missingness depends upon the actual value of a data point had it been observed. This presents a quandary for researchers who commonly encounter NMAR data. At present, there are only two general methods for dealing with NMAR data, and neither method is entirely optimal. Selection models attempt to model simultaneously the measurement and missing data processes. For example, Touloumi *et al.* [46] employed a selection model approach examining the efficacy of a treatment for HIV positive individuals, wherein a linear random effects model examined the trajectory of the outcome and a log-normal survival time model examined the missing data process (death in this case). Although selection models have intuitive appeal and can be informative, the approach relies on unverifiable assumptions and a direct comparison between different models is not possible. As a result, selection model analyses are often accompanied by sensitivity analyses [9,47,48••].

A second approach to dealing with NMAR data involves pattern-mixture models that represent a conceptual synthesis of mixed effects models and multiple group analyses in the context of structural equation modeling. A review of mixed model analysis is beyond the scope of this paper; however, quality sources are available [49–51]. Although mixed models are generally suitable methods for handling missing data under the MAR assumption, they have features that make them attractive for analyzing NMAR data. First, random effects models do not require that all individuals possess the same number of repeated observations. Second, mixed models use data at the individual level to represent that person's deviation from general trends. Finally, within the pattern-mixture approach, these methods do not require an ignorable (i.e. MCAR or MAR) missing data assumption [24].

Under the pattern-mixture approach, one creates subsamples from the entire data set based on patterns of missing data and codes the different patterns with dummy variables. For example, in a simple treatment versus control design with repeated assessments, one can model the trajectory of the outcome over time, the effect of the treatment on that trajectory, and the interaction between treatment and dropout status [7,8,21•,24]. Of particular interest in these types of analysis are interactions between treatment, time, and dropout status. Hedeker and Gibbons [24] provide a clear discussion of these procedures, interpretation of model results, and syntax for running these analyses in SAS. In the context of missing data, such models provide useful information about how missingness influences other effects within a study. Simulation work demonstrates that mixed effect models perform as well as multiple imputation under both MCAR and MAR conditions [8,11••] and at least one simulation provides evidence suggesting that mixed effects models may perform well under NMAR conditions [11••].

As with all the currently available techniques for handling missing data, pattern-mixture models also have limitations. For example, the number of distinct data patterns can become unwieldy, resulting in too few cases within each group. Collapsing over patterns provides a simple remedy to this problem. A much more important issue concerns the inability to test the missing data assumption upon which the model is based. It is important to realize that these procedures can produce biased estimates, constrained variances, and inaccurate confidence intervals when the model is misspecified [11••].

One final note concerning techniques appropriate for NMAR data brings us back to standard multiple imputation. Recent simulation work suggests that departures from MAR toward NMAR may not dramatically alter the accuracy of estimates and standard errors obtained using multiple imputation [52]. It is important to note, however, that these conclusions are not definitive, and current consensus holds that multiple imputation is most applicable under MCAR and MAR conditions.

Conclusion

Missing data is a problem inherent to the conduct of most research, and recognition of the need to deal with missing data is clearly on the rise. As such, it is vital that researchers become aware of available techniques and their appropriate application. This is by no means an easy task as the development of new techniques and refinements to techniques that already exist proliferate. This task is further complicated by demonstrations that some techniques perform reasonably well within a specific context but are generally considered inappropriate. At present, there is no uniform solution to dealing with missing data. However, our review of the existing research suggests that multiple imputation is a dominant model because of its applicability to diverse missing data problems, and performance under realistic data conditions. In our view, the only major drawback to multiple imputation concerns the fact that the missing data problem is handled, not investigated. There are clearly instances in which information stemming from the fact that data are missing is informative. As a result, we feel pattern-mixture methods hold intuitive appeal. Clearly,

future research needs to focus on the development of methods suitable for handling NMAR data. It is our opinion that attention focused in this area will prove to be most fruitful.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:
- • of special interest
- •• of outstanding interest

1 Ladouceur R, Gosselin P, Laberge M, Blaszczynski A. Dropouts in clinical research: do results reported reflect clinical reality? The Behavior Therapist 2001; 24:44–46.

2 Little RJA, Rubin DB. Statistical analysis with missing data. New York: John Wiley & Sons; 1987.

3 Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2002.

4 Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002; 7:147–177.

5 • Harezlak J, Gao S, Hui SL. An illness–death stochastic model in the analysis of longitudinal dementia data. Stat Med 2003; 22:1465–1475.
Proposes an illness–death model in which disease incidence and mortality are modeled simultaneously using data from patients with dementia.

6 Gao S, Hui SL. Estimating the incidence of dementia from two-phase sampling with non-ignorable missing data. Stat Med 2000; 19:1545–1554.

7 Nich C, Carroll KM. 'Intention-to-treat' meets 'missing data': implications of alternate strategies for analyzing clinical trials data. Drug Alcohol Depend 2002; 68:121–130.

8 Liu G, Gould AL. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. J Biopharm Stat 2002; 12:207–226.

9 Hollis S. A graphical sensitivity analysis for clinical trials with non-ignorable missing binary outcome. Stat Med 2002; 21:3823–3834.

10 • Allison PD. Missing data techniques for structural equation modeling. J Abnorm Psychol 2003; 112:545–557.
Provides a solid review of missing data techniques as they apply to structural equation modeling analyses.

11 •• Gadbury GL, Coffey CS, Allison DB. Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond LOCF. Obesity Rev 2003; 4:175–184.
This simulation presents results suggesting that multiple imputation and mixed modeling techniques behave consistently across types of missing data. The article also contains syntax for conducting both multiple imputation and mixed model analysis in SAS.

12 •• Mallinckrodt CH, Sanger TM, Dubé S, *et al.* Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. Biol Psychiatry 2003; 53:754–760.
This article presents a nice example of mixed modeling with missing data. In addition, the authors demonstrate inadequacies associated with 'last observation carried forward' techniques.

13 • Demissie S, LaValley MP, Horton NJ, *et al.* Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. Stat Med 2003; 22:545–557.
This work provides a simulation testing bias for complete case analysis in a proportional hazards model.

14 •• Hogan JW, Lancaster T. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. Stat Methods Med Res 2004; 13:17–48.
The authors provide a detailed discussion of inverse probability weighting and the use of instrumental variables as methods to deal with missing data. Perhaps more importantly, this work specifically addresses assumptions related to each approach, offering an in-depth discussion of the need for conceptual as opposed to analytical justification for use of these methods.

15 Dunn G, Pickles A, Tansella M, Vázquez-Barquero JL. Two-phase epidemiological surveys in psychiatric research. Br J Psychiatry 1999; 174:95–100.

16 •• Dunn G, Maracy M, Dowrick C, *et al.* Estimating psychological treatment effects from a randomized controlled trial with both non-compliance and loss to follow-up. Br J Psychiatry 2003; 183:323–331.
Dunn and his colleagues present an insightful example using complier average causal effect (CACE) estimation methods to address the often-ignored difference between effects related to offering treatment, and effects related to the actual receipt of treatment. This work provides a compelling example of how one's specific research question should guide choice of methods for handling missing data.

17 Pérez A, Dennis RJ, Gil JFA, *et al.* Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. Stat Med 2002; 21:3885–3896.

18 •• Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. J Clin Epidemiol 2003; 56:968–976.
This study presents results from simulations examining 14 single imputation techniques. Results suggest that techniques which rely on person-specific information perform better than those that do not.

19 •• Oostenbrink JB, Al MJ, Rutten-van Mölken MPMH. Methods to analyse cost data of patients who withdraw in a clinical trial setting. Pharmacoeconomics 2003; 21:1103–1112.
The authors present very clear explanations of many missing data procedures. In addition, they perform multiple imputation using a novel method involving a propensity score.

20 •• Chen G, Åstebro T. How to deal with missing categorical data: test of a simple Bayesian method. Organ Res Methods 2003; 6:309–327.
The authors develop a Bayesian procedure for imputing categorical data that can extend to perform multiple imputations. In addition, the authors provide an Excel macro for performing the imputation.

21 • Mallinckrodt CH, Clark WS, Carroll RJ, Molenberghs G. Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. J Biopharm Stat 2003; 13:179–190.
The authors argue for the use of mixed models over multiple imputation based on single prespecified analyses required in the context of regulatory decisions.

22 • Newman DA. Longitudinal modeling with randomly and systematically missing data: a simulation of ad hoc, maximum likelihood, and multiple imputation techniques. Organ Res Methods 2003; 6:328–362.
This study presents results from a complex simulation study that demonstrate the comparability of full information maximum likelihood and multiple imputation techniques.

23 •• Olinsky A, Chen S, Harlow L. The comparative efficacy of imputation methods for missing data in structural equation modeling. Eur J Oper Res 2003; 151:53–79.
The authors provide an extensive review of available missing data techniques. In addition, simulation results indicate that the performance of FIML and multiple imputation vary depending on sample size within structural equation modeling analyses.

24 Hedeker D, Gibbons RD. Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychol Methods 1997; 2:64–78.

25 Hall SM, Delucchi KL, Velicer WF, *et al.* Statistical analysis of randomized trials in tobacco treatment: longitudinal designs with dichotomous outcome. Nicotine Tobacco Res 2001; 3:193–202.

26 • Jung SH, Ahn C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. Stat Med 2003; 22:1305–1315.
The authors propose a formula for calculating necessary sample size to ensure adequate statistical power in the planning stage of a study intended to test linear slopes.

27 • Matsuyama Y. Sensitivity analysis for the estimation of rates of change with non-ignorable drop-out: an application to a randomized clinical trial of the vitamin D3. Stat Med 2003; 22:811–827.
This paper presents an extension of generalized estimating equation methods to data that are missing at random.

28 FitzGerald PEB. Extended generalized estimating equations for binary familial data with incomplete families. Biometrics 2002; 58:718–726.

29 •• Lyles RH, Allen AS. Missing data in the 2×2 table: patterns and likelihood-based analysis for cross-sectional studies with supplemental sampling. Stat Med 2003; 22:517–534.
This paper presents a novel approach to dealing with binary data that are not randomly missing. The authors propose a targeted resampling method aimed at recovering information for those with missing data through assessment of supplemental information.

30 Hudgens MG. Estimating cumulative probabilities from incomplete longi-
•• tudinal binary responses with application to HIV vaccine trials. Stat Med 2003; 22:463–479.
The author demonstrates that empirical estimators for marginal, cumulative, and cumulative repeated probabilities are biased even when the data are known to be missing completely at random due to planned missingness. In addition, this work presents an alternative unbiased method for estimating probabilities and offers a SAS macro for performing the analysis.

31 Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. Psychol Methods 2001; 6:317–329.

32 Albert PS, Follmann DA, Wang SA, Suh EB. A latent autoregressive model for longitudinal binary data subject to informative missingness. Biometrics 2002; 58:631–642.

33 Horton NJ, Laird NM. Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. Biometrics 2001; 57:34–42.

34 Gold MS, Bentler PM, Kim KH. A comparison of maximum-likelihood and
• asymptotically distribution-free methods of treating incomplete nonnormal data. Struct Equation Model 2003; 10:47–79.
This paper presents an adjusted expectation maximization method that outperforms distribution free estimators in moderate samples when data are incomplete, and non-normal.

35 Graham JW, Cumsille PE, Elek-Fisk E. Methods for handling missing data. In:
•• Schinka JA, Velicer WF, volume editors. Research methods in psychology. In: Weiner IB, editor. Handbook of psychology. vol 2. New York: John Wiley & Sons; 2003. pp. 87–114.
This chapter provides a detailed review of currently available methods for handling missing data. In addition, the authors provide a step-by-step example of how to conduct multiple imputation using NORM with generalizations to other software packages.

36 Graham JW. Adding missing-data-relevant variables to FIML-based structural
•• equation models. Struct Equation Model 2003; 10:80–100.
This work demonstrates new methods for including missing data covariates that are not present in the model of interest, using FIML in the context of structural equation modeling. This advancement effectively supplants expectation maximization for structural equation modeling analyses as it is no longer a more flexible approach than FIML.

37 Song XY, Lee SY. Full maximum likelihood estimation of polychoric and
• polyserial correlations with missing data. Multivariate Behav Res 2003; 38:57–79.
The authors present an extension of FIML estimation that effectively solves problems associated with relationships among categorical variables typically encountered using the expectation maximization algorithm.

38 Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.

39 Schafer JL. Analysis of incomplete multivariate data. New York: Chapman and Hall; 1997.

40 Arnold AM, Kronmal RA. Multiple imputation of baseline data in the
• Cardiovascular Health Study. Am J Epidemiol 2003; 157:74–84.
The authors suggest an interesting approach for handling multivariate outliers resulting from known errors by setting their data to missing and performing multiple imputation.

41 Weinfurt KP, Castel LD, Li Y, *et al*. The correlation between patient
• characteristics and expectations of benefits from phase I clinical trials. Cancer 2003; 98:166–175.
Table 2 (p. 172) provides an informative comparison between original data and values obtained from multiple imputation after data were deleted.

42 Badzioch MD, Thomas DC, Jarvik GP. Summary report: missing data and
•• pedigree and genotyping errors. Genet Epidemiol 2003; 25 (Suppl 1):S36–S42.
This study presents an informative use of multiple imputation applied to nested units within families. To date, there are relatively few examples of imputation with data containing high interdependencies among respondents.

43 Clark TG, Altman DG. Developing a prognostic model in the presence of
• missing data: an ovarian cancer case study. J Clin Epidemiol 2003; 56:28–37.
This study provides an example using multiple imputed data sets in the context of Cox regression.

44 Graham JW, Schafer JL. On the performance of multiple imputation for multivariate data with small sample size. In: Hoyle R, editor. Statistical strategies for small sample research. Thousand Oaks, CA: Sage; 1999. pp. 1–29.

45 Hershberger SL, Fisher DG. A note on determining the number of
• imputations for missing data. Struct Equation Model 2003; 10:648–650.
The authors argue that common recommendations regarding the number of imputations to perform are not theory driven and may result in severe underestimations.

46 Touloumi G, Babiker AG, Kenward MG, *et al*. A comparison of two methods for the estimation of precision with incomplete longitudinal data, jointly modeled with a time-to-event outcome. Stat Med 2003; 22:3161–3175.

47 Baker SG, Ko CW, Graubard BI. A sensitivity analysis for nonrandomly missing categorical data arising from a national health disability survey. Biostatistics 2003; 4:41–56.

48 Magder LS. Simple approaches to assess the possible impact of missing
•• outcome information on estimates of risk ratios, odds ratios, and risk differences. Control Clin Trials 2003; 24:411–421.
The author provides an interesting approach to sensitivity analysis that does not require a complex model. In addition, this work proposes a parameter that quantifies the degree of departure from the missing at random assumption for binary missing data.

49 Bryk AS, Raudenbush SW. Hierarchical linear models: applications and data analysis methods. Newbury Park, CA: Sage; 1992.

50 Raudenbush SW, Bryk AS. Hierarchical linear models: applications and data analysis methods. 2nd ed. Thousand Oaks, CA: Sage; 2002.

51 Singer JD, Willett JB. Applied longitudinal data analysis: modeling change and event occurrence. New York: Oxford University Press; 2003.

52 Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001; 6:330–351.