Prev Sci (2018) 19:284–294
DOI 10.1007/s11121-016-0644-5

Principled Missing Data Treatments

Kyle M. Lang[1] · **Todd D. Little**[1]

Published online: 4 April 2016

Abstract We review a number of issues regarding missing data treatments for intervention and prevention researchers. Many of the common missing data practices in prevention research are still, unfortunately, ill-advised (e.g., use of listwise and pairwise deletion, insufficient use of auxiliary variables). Our goal is to promote better practice in the handling of missing data. We review the current state of missing data methodology and recent missing data reporting in prevention research. We describe antiquated, ad hoc missing data treatments and discuss their limitations. We discuss two modern, principled missing data treatments: multiple imputation and full information maximum likelihood, and we offer practical tips on how to best employ these methods in prevention research. The principled missing data treatments that we discuss are couched in terms of how they improve causal and statistical inference in the prevention sciences. Our recommendations are firmly grounded in missing data theory and well-validated statistical principles for handling the missing data issues that are ubiquitous in biosocial and prevention research. We augment our broad survey of missing data analysis with references to more exhaustive resources.

Keywords Missing data · Multiple imputation · Full information maximum likelihood · Auxiliary variables · Intent-to-treat · Statistical inference

✉ Kyle M. Lang
kyle.lang@ttu.edu

✉ Todd D. Little
yhat@ttu.edu; yhat@statscamp.org

[1] Institute for Measurement, Methodology, Analysis, and Policy, Texas Tech University, Lubbock, USA

Principled Missing Data Treatments

Missing data are a common problem for prevention research and improperly handling missing data can severely compromise the validity of a study's inferences. The situation, however, is not as bleak as it may seem at the outset. Though not trivial, missing data analysis is a ubiquitous component of applied data analysis for which many powerful methods have been developed (e.g., multiple imputation (MI)—Rubin 1978, 1987; full information maximum likelihood (FIML)—Anderson 1957 and multiple imputation with chained equations (MICE)—Raghunathan et al. 2001; van Buuren et al. 2006). When applied correctly, these principled missing data treatments can help recover the underlying inferential model and maximize a study's validity, even in the presence of high rates of nonresponse (Little et al. 2016).

We review current best practice in missing data analysis, by which we mean both the process of elucidating the extant missing data problem (i.e., missing data diagnostics) and the act of addressing the missing data themselves (i.e., missing data treatment). We focus our discussion on applications in the prevention sciences, although we note that the missing data problems encountered by prevention researchers are not substantially different from those encountered in other social and behavioral science research.

Treating missing data correctly is absolutely necessary to ensure the validity of scientific research because improperly handling nonresponse can substantially compromise a study's inferences. We argue that correctly addressing nonresponse is an obligation, and not a choice, of all research scientists. This view is especially true for prevention scientists because prevention research tends to directly impact a large swath of stakeholders. We have compiled the following material with three primary goals in mind: (1) to give a high-level overview

of the current state of missing data methods; (2) to illustrate why prevention researchers should prefer modern, principled missing data treatments over other antiquated, yet still common, approaches, with an emphasis on how modern missing data tools can be especially useful in addressing some of the difficult missing data problems that arise in intervention studies and randomized controlled trials; and (3) to highlight the importance of planning for missing data early in a study's design phase and to show how doing so can dramatically improve the quality of a study's inferences by maximizing the chance that important assumptions of the missing data methods are met. The methods we discuss below are easily implemented and well suited to the types of analysis that are common in prevention research (Enders 2010; Graham 2012; van Buuren 2012).

We first review missing data treatments reported in recent issues of *Prevention Science*. We then highlight important characteristics of applied missing data problems and introduce the two flagship methods of modern missing data analysis, namely, explicit model-based MI and FIML estimation. We will emphasize the superiority of these modern methods by contrasting them with four less optimal (yet still commonly employed) approaches: deletion-based techniques, single imputation methods, last observation carried forward (LOCF), and nonresponse weighting. We conclude by offering some practical guidance for researchers implementing MI and FIML.

Review of Missing Data Reporting

Our review of recent *Prevention Science* articles indicates a need for improved missing data treatment in prevention research. Specifically, we reviewed the missing data treatments reported in *Prevention Science* from February 2013 (Volume 14, Issue 1) to July 2015 (Volume 16, Issue 5). This window included 240 potential papers. We excluded articles that did not report empirical research studies (e.g., commentaries, literature reviews, errata), papers that employed qualitative methodologies, meta-analyses, and methodological papers. These exclusion criteria produced a final sample of 169 valid articles for review. All articles were coded by trained raters for two key features: (1) missing data reporting (e.g., acknowledgment of missing data, explicit reports of nonresponse rates) and (2) missing data treatment (e.g., methods used to treat the missing data, use of auxiliary variables). The raters' coding was confirmed by the first author.

The majority of articles ($n = 123$, 72.8 %) explicitly acknowledged missing data and explicitly reported some measure of the nonresponse rate (e.g., percent missing, attrition rates, covariance coverage). Several papers ($n = 19$, 11.2 %) acknowledged the presence of missing data but did not provide any explicit indication of the nonresponse rate. A total of 27 papers (16.0 %) made no mention of missing data, at all. No articles mentioned the fraction of missing information or any related quantity (e.g., relative increase in variance due to nonresponse).

The most commonly reported missing data treatment was listwise or pairwise deletion ($n = 50$, 29.6 %), followed by FIML ($n = 46$, 27.2 %), MI ($n = 22$, 13.0 %), single imputation ($n = 13$, 7.7 %), coding missing values as a categorical response level ($n = 3$, 1.8 %), nonresponse weighting ($n = 2$, 1.2 %), and LOCF ($n = 2$, 1.2 %). Surprisingly, 39 (23.1 %) studies gave no explicit description of how they addressed missing data. Of these 39 studies, 12 explicitly acknowledge the presence of nonresponse in their data. These 12 studies can reasonably be assumed to have employed a deletion-based treatment (i.e., listwise or pairwise deletion). Under this assumption, the number of deletion-based missing data treatments rises to 62 (36.7 %) and clearly dominates the distribution of approaches employed. A number of studies ($n = 12$, 7.1 %) employed multiple missing data treatments, so the counts reported above exceed 169.

Although the rates of missing data reporting and use of modern missing data treatments (i.e., MI and FIML) are higher in this review than other recently published surveys of missing data reporting practices (e.g., Bodner 2006; Little et al. 2014; Peugh and Enders 2004), our findings still suggest considerable room for improvement in the missing data methods employed by prevention researchers. First and foremost, deletion-based techniques (i.e., listwise and pairwise deletion) remain the most frequently employed treatment for nonresponse. This practice is troubling because these methods are well-known to be among the poorest choices of missing data treatment (Wilkinson and Task Force on Statistical Inference 1999). Also, although the rates of FIML usage were promisingly high, only 3 of the 46 studies that employed FIML used auxiliary variables or covariates that were explicitly included to support the fundamental assumptions of missing data analysis. The very low rate at which auxiliary variables were employed suggests that many authors may simply have relied on software defaults without giving much thought to their missing data problem (a possibility also noted by Little et al. 2014). Finally, although the rates of basic missing data reporting were high (72.8 %), this number should be 100.00 %. At a bare minimum, all authors should honestly report the extent of the missing data in their study, regardless of how they treat those missing data. If no missing data are present, this fact should be clearly stated.

Important Considerations for Missing Data Analyses

There are several critical characteristics of a missing data problem that must be considered before the missing data

themselves can be addressed. We first discuss those aspects of missing data problems that play the largest role in applied missing data analyses.

Nonresponse Pattern

One of the most basic features of a missing data problem is its *nonresponse pattern*, which simply refers to the spatial arrangement of the empty cells in an incomplete data set. The simplest of these patterns is *univariate nonresponse* in which missing data occur on only one variable. A second nonresponse pattern is the so-called *monotone nonresponse* pattern, which occurs when the rows and columns of a data set can be ordered by decreasing completeness so that the observed portions form a "staircase" pattern in which, when traversing rows or columns, every entry following the first missing datum is also missing. Such patterns are common in longitudinal research where they arise from *attrition* (i.e., participants permanently dropping out of the study), although they almost always present in addition to the final nonresponse pattern: *arbitrary nonresponse*. This pattern occurs when cells of the data set are missing in an arbitrary, apparently random, arrangement—though this evident randomness is rarely truly random, as we discuss below.

There is an alternative classification of nonresponse that can also help guide the design of a missing data analysis. This classification, which originated in the literature on sample surveys, differentiates between *item nonresponse* and *unit nonresponse*. Unit nonresponse occurs when an entire observational unit (e.g., a participant in an intervention study) fails to give any data. Unit nonresponse leads to entire rows of data sets being missing. Item nonresponse occurs when individual cells in a data set are empty but each row contains at least one observed data element. Unit nonresponse is a degenerate special case of monotone nonresponse, while item nonresponse subsumes the typical presentations of univariate, monotone, and arbitrary nonresponse.

Nonresponse Rate and Fraction of Missing Information

Another characteristic that must be accounted for when planning a missing data analysis is the actual nonresponse rate. That is, exactly how much of the anticipated sample size has been lost to missing data? There are several ways to quantify the nonresponse rate for any given missing data problem. The simplest of these measures is the percentage of missing data (or percent missing), which is the percentage of the total cells in a data set that are missing. A closely related quantity is the *attrition rate* which simply quantifies the proportion of participants in a longitudinal study who permanently "drop out" at each measurement occasion. Percent missing and attrition rate are important early screening measures that give a rough idea of the severity of the missing data problem. Yet, neither percent missing nor attrition rate give much information on how well the missing data treatment will perform or how the missing data model should be parameterized because neither of these metrics account for how well the observed data can help recover the missing values. Another simple measure of nonresponse rate is the so-called *covariance coverage*. The covariance coverage gives the proportion of observations that are available to estimate each pairwise relationship. Covariance coverage is important because low coverage indicates that the observed data offer little information to help the estimation process. Relationships with low coverage values tend to be poorly recovered by most missing data treatments.

The most important measure of nonresponse rate is the fraction of missing information (FMI). FMI quantifies the amount of a parameter's information that is lost to nonresponse. Because information and variance are inversely proportional quantities, the FMI also quantifies the increase in a parameter's sampling variability due to the missing data (Rubin 1987). In this sense, the FMI can be viewed as analogous to an R^2 statistic for the missing data (Enders 2010). The FMI underlies many important components of a missing data analysis, including statistical power lost to nonresponse (Savalei and Rhemtulla 2012), the convergence rates of missing data algorithms (Schafer 1997), and the number of imputations required when using multiple imputation (Graham et al. 2007). The FMI can be readily estimated as a byproduct of both MI- and FIML-based missing data analyses (Savalei and Rhemtulla 2012). When reporting a missing data analysis, the FMI of important parameter estimates should be presented to facilitate the reader's ability to judge the missing data's impact on the inferences presented. As mentioned, the FMI was not reported in any of the articles reviewed for this paper.

Nonresponse Mechanism

Each study variable (e.g., gender, body mass index) can be augmented with a binary random variable coding nonresponse. These nonresponse indicators represent the *missingness* of the respective variables. Some of the most crucial assumptions underlying modern missing data analysis pertain to the way that the missingness is related to the observed and missing components of the study variables. These *nonresponse mechanisms* are a set of probability statements that describe the interrelations of the missingness (i.e., the binary nonresponse indicators) and the study variables. There are three such mechanisms: *missing at random* (MAR), *missing completely at random* (MCAR), and *missing not at random* (MNAR).

MAR missingness can be predicted by the observed components of other variables on the data set, but, after controlling for these observed predictors, MAR missingness is not predictable by the missing components of any study variables.

Note that a variable's nonresponse indicator will be constant for the nonmissing elements of that variable, so MAR missingness must, by definition, be independent of the study variable whose missing data it encodes, after controlling for other predictors on the data set. MCAR missingness is actually a special case of MAR missingness that occurs when the nonresponse is a purely random sample of the complete data. Thus, MCAR missingness is independent of both the observed and missing components of all study variables. MNAR missingness occurs when the missingness remains predictable by the missing components of some study variables (possibly the variable whose missing data it encodes), even after controlling for the observed portions of all variables on the data set. MNAR missingness is nearly impossible to treat well because the missing values are not sufficiently predicted by the observed portions of the data set, so the observed data cannot provide enough information to adequately approximate the missing data's distribution.

The residual dependence that characterizes the MNAR mechanism can be induced through multiple processes; consequently, Enders (2010) distinguishes between *direct* and *indirect* MNAR mechanisms. Direct MNAR occurs as described above: the participants' latent levels of the missing components of the study variables are directly associated with their propensity to respond. The indirect MNAR case, on the other hand, is actually a corrupted MCAR mechanism that arises from the proverbial *third variable problem*. Under indirect MNAR, there is no true relationship between the study variables' missing components and the missingness, but both of these variables are related to an *unmeasured* third variable that induces a spurious association that manifests as an MNAR mechanism.

By way of example, consider a hypothetical study of an intervention to prevent teen pregnancy by promoting condom usage among high school students. If some religious students dropped out of the study because they were offended by the subject matter, their attrition would be MAR. As long as the study data included some measure of religiosity, controlling for religiosity would account for any differences between completers and dropouts in the post-drop out levels of the study variables. If some students dropped out of the study because they moved out of the intervention area, then their attrition would be MCAR. As long as these moves were not associated with some aspect of the intervention, these students' hypothetical post-drop out values would be stochastically equivalent to the rest of the students' values. If some students dropped out of the study because they became pregnant, then their attrition would be MNAR. This missingness would remain associated with an unmeasured outcome (i.e., the students' pregnancies) even after accounting for all measured variables. If the study did not measure religiosity, then the MAR example given above would degrade into an indirect MNAR situation. Students who dropped out because of religious objections may also have low rates of condom usage and correspondingly high rates of teen pregnancy. Without controlling for religiosity, these students' attrition would remain associated with unmeasured variables (i.e., condom usage and pregnancy), after accounting for all measured variables. Optimizing the chance that the inevitable missing data will be MAR (as opposed to MNAR) is the primary reason to proactively plan for missing data. Including likely correlates of the missingness in the data collection minimizes the chances of encountering indirect MNAR in scientific studies.

Ignorability of the Nonresponse Mechanism

MAR and MCAR are *ignorable* mechanisms because their effects on bias, validity, precision, and power can be mitigated without explicitly modeling the nonresponse mechanism (i.e., by using MI or FIML). Thus, MCAR and MAR are *ignorable nonresponse mechanisms* in the same way that simple random sampling is an *ignorable sampling mechanism*. MNAR, on the other hand, is *nonignorable* (in the same way that stratified random sampling is a *nonignorable sample mechanism*) because it will lead to biased results unless the missing data analysis incorporates an explicit, and correct, model for the nonresponse mechanism or additional variables are introduced that correlate strongly enough with the missingness to induce a MAR mechanism. By planning for missing data and proactively measuring potential correlates of the missingness, researchers can approximate a MAR mechanism and thereby reduce bias and increase validity (Enders 2010).

Antiquated Missing Data Treatments

To highlight the strengths of the modern missing data treatments that we discuss below, we will first describe several antiquated ad hoc approaches that remain common in the literature. Due to space limitations, our discussion of these techniques is limited. Readers are encouraged to consult Enders (2010; Chapter 2), the articles cited therein, and Little and Rubin (2002; Chapters 3 & 4), for thorough discussions of the deficiencies inherent in antiquated missing data methods.

Deletion-Based Techniques

Missing data theorists have long decried deletion-based techniques as some of the worst options for treating missing data (Wilkinson and Task Force on Statistical Inference 1999). Unfortunately, they still remain common in many scientific studies (Bodner 2006; Little et al. 2014; Peugh and Enders 2004). Deletion-based techniques come in two flavors, listwise deletion (or complete case analysis) in which any incomplete row is deleted and pairwise deletion (or available

case analysis) in which sufficient statistics are computed using only those rows for which every constituent variable has been observed (e.g., the correlation between X and Y is computed using only rows with no missing on either X or Y). Listwise deletion has two major problems: (1) it leaves nonresponse bias unaddressed and thus leads to biased statistical inferences unless the data are truly MCAR (Little and Rubin 2002) and (2) it can lead to a substantial loss of power since a large proportion of the sampled units will tend to be discarded (Enders 2010). Pairwise deletion will maintain higher power than listwise deletion will, but, in addition to requiring MCAR to produce unbiased estimates, pairwise deletion can lead to sufficient statistics with inconsistent degrees of freedom (since each statistic can be derived from a different set of observations). This inconsistency can produce estimated correlations outside of the interval [−1, 1], sample covariance matrices that are not positive definite, and biased standard errors that lead to incorrect inferences, even when the missing data are MCAR (Little and Rubin 2002). Despite these well-known deficiencies, deletion-based techniques were the most common missing data treatment in the articles reviewed for this paper.

Single Imputation Techniques

Single imputation techniques are not suitable candidates for general-purpose missing data treatments. There are three common types of single imputation: *unconditional mean substitution* in which each variable's missing entries are replaced with the mean of that variable's observed portion, *deterministic regression imputation* (i.e., *conditional mean substitution*) where each variable's missing entries are replaced with predicted values from a regression equation in which the incomplete variable acts as the dependent variable, and *stochastic regression imputation* which adds an additional random error term to the predicted values imputed by deterministic regression imputation. Unconditional mean substitution can introduce high levels of bias in the final parameter estimates by pulling the distribution of the imputed data toward the mean of the observed data (van Buuren 2012). Deterministic regression imputation will underestimate the variance of the imputed items and inflate linear associations involving imputed variables because the imputed values fall directly on the regression surface (Enders 2010). Finally, stochastic regression imputation will produce unbiased point estimates of model parameters, but it can lead to inflated type I error rates because it does not adequately quantify the uncertainty introduced by the missing data and, thereby, attenuates standard errors for model parameters (Rubin 1987). Stochastic regression imputation does incorporate random error into the imputed values themselves, but it treats the imputation model as fixed. To achieve *proper* imputations in the sense of Rubin (1987), the uncertainty in the imputation model itself must also be modeled, either via Bayesian simulation or bootstrapping (Allison 2002; van Buuren 2012), which is the defining characteristic of MI (see below). Methods are available to correct the standard errors in stochastic regression imputation (see Little and Rubin 2002; Chapter 5), but they are complicated solutions to a problem that MI addresses automatically.

Last Observation Carried Forward

LOCF is a deterministic single imputation technique that simply entails replacing all of a longitudinal observation's post-drop out missing values with its last observed value. This method can seriously compromise a study's inferences and lead to highly invalid conclusions (Enders 2010; Little and Yau 1996; van Buuren 2011). An implicit assumption of LOCF is that participants who drop out of a study would have maintained their last observed levels on all variables. This limitation is often acknowledged and cited as leading to conservative conclusions, but LOCF can just as easily lead to liberal bias. Any intervention that is designed to decrease the rate of some behavior will be liberally biased by LOCF. If all participants subject to the intervention are expected to demonstrate a monotonic increase or decrease in some outcome measure over the course of the study (e.g., increasing frequency of risky sexual behavior or decreasing school attendance), and the effect of the intervention is primarily to slow this progression, then freezing dropouts' responses at an early measured level will spuriously inflate the intervention's estimated effect. The goal of statistical inference is not to be conservative or liberal but rather as unbiased as possible.

Nonresponse Weighting Approaches

Weighting techniques are not necessarily antiquated or ad hoc, but they were developed to address unit nonresponse (Little and Rubin 2002), so their applicability is limited in prevention research. Nonresponse weighting involves constructing and applying columns of weights in an effort to remove nonresponse bias from a study's final inferences. Nonresponse weighting should not be applied to item nonresponse; on the other hand, MI and FIML are intractable when facing true unit nonresponse in which no data are available for some units. This relatively rare circumstance leaves nonresponse weighting as one of the only principled missing data methods available for such problems. Yet, even when a unit gives no data, there will sometimes be information on that unit available from an alternative source (e.g., many intervention studies can access baseline and demographic data collected as part of the experimental sites' normal operating procedures). When such supplementary data are available, they can be incorporated into the missing data analysis to turn much (or all) of the unit nonresponse into monotone nonresponse. In these cases, we recommend using MI or FIML to treat the

missing units at the same time as any item nonresponse. Missing units that are unobserved due to a failure to consent or refusal to participate in the study should be flagged during the missing data treatment (e.g., with a dummy coded variable) to ensure that any differences between the groups are represented in the imputations or the FIML estimates.

Recommended Missing Data Treatments

As intimated throughout, there are two flagship techniques in modern missing data analysis: MI and FIML. These methods provide optimal results in the majority of missing data problems, and we clearly advocate their proper use whenever missing data occur in applied research. FIML is easily implemented and is particularly well suited to latent variable modeling. MI is slightly more labor intensive than FIML, but this additional effort is paid back with extreme flexibility.

Multiple Imputation

MI was originally introduced by Rubin (1978) and later refined by Rubin (1987). It is an incredibly powerful missing data tool that originates from the Bayesian analysis of large-scale sample surveys (e.g., national censes). This pedigree is one of MI's greatest strengths. Because it was developed from a Bayesian perspective for use within a randomization-based framework, the conclusions drawn from a well-implemented MI analysis are valid from both Bayesian and Frequentist perspectives and lead to valid model-based or randomization-based inferences (Little and Rubin 2002).

MI analyses can be broken into three steps: the imputation phase, the analysis phase, and the pooling phase. The imputation phase entails create $m > 1$ replacements for the missing data by taking m random draws from their posterior predictive distribution. These m replacements are then used to fill in the missing data to create m imputed data sets. The analysis phase consists of fitting m replicates of the analysis model to these m imputed data sets. Finally, the pooling phase employs *Rubin's Rules* (Rubin 1987, pp. 76–77) to aggregate the m sets of estimates into the final pooled point estimates and standard errors that are used for optimally accurate inference.

Full Information Maximum Likelihood

FIML (Anderson 1957; also known as direct maximum likelihood) is a maximum likelihood estimator that is robust to ignorable item nonresponse. It is a clever extension of ordinary maximum likelihood estimation that modifies the sample log-likelihood function to consider only the observed elements of the data matrix. In this way, FIML can leverage all of the available information when fitting a statistical model (Savalei and Rhemtulla 2012). In practice, FIML has been shown to perform very well (Arbuckle 1996; Enders 2001; Enders and Bandalos 2001). Under a MAR nonresponse mechanism, *when a good set of auxiliary variables are included in the model* (e.g., via the saturated correlates technique, Graham 2003), FIML will produce optimal estimates that are asymptotically equivalent to those derived from MI (Savalei and Rhemtulla 2012). If, however, auxiliary variables are not employed, the MAR assumption will only hold when all predictors of the missingness are included in the inferential model. Whenever this is not the case, the FIML estimates will be biased (Enders 2010; Graham 2003). Only 3 out of 46 FIML-based analyses reviewed above reported using some form of auxiliary variables. This finding suggests that the high rates of FIML adoption among prevention scientists may be undercut by incorrect applications of the FIML technique that leave inferences compromised. Yet, these deficiencies can be easily addressed by planning for the inevitable missing data and proactively including good auxiliary variables into the study design.

Suggested Resources

Due to space limitations, we do not provide detailed guidance on implementing MI or FIML. We strongly encourage readers to consult Enders (2010) and Graham (2012) for exhaustive, yet very approachable, introductions to missing data analysis. More details on implementing MI can be found in Carpenter and Kenward (2013) and van Buuren (2012). Finally, Little and Rubin (1987, 2002), and Schafer (1997) represent definitive resources for the technical underpinnings of modern missing data analysis.

Practical Guidance for MI and FIML

MI and FIML are very powerful and versatile missing data treatments, but there are several practical issues that can arise when implementing these methods. We now delineate several pieces of practical advice for researchers who are using MI or FIML in their own work.

Choosing between FIML and MI

FIML performs very well when its assumptions are met and, when using modern statistical analysis packages, FIML is often simpler to implement than MI, but there are several common circumstances where MI is preferred. Although there is no mathematical reason that FIML cannot be applied to categorical data, this capability remains unavailable in most statistical software packages outside of the IRT context. At this time, limited software implementation impedes many researchers' abilities to apply FIML to nonnormal data, whereas

many MI software packages, especially those that employ the MICE framework, readily accommodate categorical distributions for the missing data. FIML is also limited when the raw, incomplete, data must be aggregated into composite items (e.g., scale scores, parcels) before the analysis. FIML simply partitions the missing data out of the likelihood function while estimating the analysis model, but it never "fills in" any of the missing cells. Thus, there is no obvious way to aggregate the incomplete items (how does one compute a sum when a subset of the summands does not exist?). When employing MI, however, the data can simply be imputed at their lowest level of granularity and pooled to whatever level of abstraction is convenient for the final data analysis. Finally, because FIML is a maximum likelihood procedure, it cannot be applied to any modeling enterprise where maximum likelihood estimation is inapplicable (e.g., ordinary least squares regression, decision tree modeling, back-propagated neural networks). In these situations, MI is the preferred missing data method.

Choosing the Imputation Model

Some of the earliest MI approaches employed the multivariate normal distribution (Rubin 1987). Creating imputations under the multivariate normal model is the most computationally expedient approach due to the convenient mathematical properties of the normal distribution. Unfortunately, much of the incomplete data in prevention research are not continuous or normally distributed (e.g., Likert-type questionnaire items, counts of substance use, indicators of school dropout). Normal-theory imputation can still be employed in many of these circumstances, but one must be cognizant of the violated assumptions and actively scrutinize the appropriateness of a normal-theory approach.

There is ample evidence for imputing under the normal model when the discrete measurement level of the items is not meaningful or when the final analysis model will treat the items as continuous, anyway. Enders (2010), Honaker and King (2010), and Schafer (1997) all suggested that imputing under the multivariate normal model can lead to accurate statistical inference when the final analysis model is naïve to the true (discrete) measurement level of the incomplete values. Wu et al. (2015) conducted a study examining how different imputation models affected the performance of MI for ordinal items that were aggregated to mean scores for analysis. They found that imputing under the multivariate normal model can lead to unbiased and efficient parameter estimates that outperform imputation methods that employed discrete distributions for the missing data (e.g., multinomial logistic regression).

When the categorical measurement level of the nonresponse must be preserved (e.g., when the imputed variable will be the outcome in a logistic regression model), the MICE framework can be tailored to use different distributions for the missing data on a variable-by-variable basis. By employing an appropriate generalized linear model as the elementary imputation method within the MICE framework, very good, principled imputations of categorical items can be created (van Buuren 2012; van Buuren et al. 2006).

Implicit, donor-based imputation methods (e.g., hotdeck imputation, predictive mean matching, *K*-nearest neighbors imputation) are intuitively appealing, but we advise against relying on donor-based methods as general missing data treatments. Donor-based methods can only perform at their optimum when they have a reasonable pool of donor cases from which they can sample to create the imputations (Andridge and Little 2010). In many missing data problems, such a representative pool is not possible because the nonresponse can shrink the observed sample size considerably—thereby producing a donor pool that is too homogenous. In such circumstances, donor-based methods need to re-use too many donor observations and the standard errors of the analysis model parameters will be attenuated (van Buuren 2012).

Addressing Temporal Dependence

When the incomplete data are longitudinal in nature, additional care must be taken to preserve the temporal dependence of the imputed values (here, we focus on the complications of MI because FIML merely requires specifying an adequate longitudinal analysis model). The most principled approach to this problem entails explicitly modeling time as part of the imputation model. The MI framework can employ essentially any predictive model to create imputations of the missing data. This flexibility allows one to impute longitudinal missing data according to a model that incorporates whatever function of time is deemed appropriate. Several common MI software packages offer such capabilities.

The R package *Amelia II* (Honaker et al. 2011) implements a rather general approach by offering the ability to include a polynomial or spline function of time into the imputation model. Cross-sectional grouping variables can also be interacted with this temporal component, so the imputations are created according to a model that allows each group to have its own trend. Honaker and King (2010) demonstrated the effectiveness of this approach for normally distributed missing data. Amelia II assumes multivariate normality for all imputed data, which does limit its applicability when the incomplete variables are categorical.

Longitudinal data can also be viewed as repeated measures nested within individual, so another convenient class of imputation model is multilevel regression models (also known as mixed effects models, hierarchical linear models, and growth curve models). Goldstein et al. (2009, 2014), Liu et al. (2000), Yucel (2008), and Schafer and Yucel (2002) have all developed MI methods based on multilevel models that can be applied to longitudinal nonresponse, and the R packages *mice* (van Buuren and Groothuis-Oudshoorn 2011) and *pan* (Zhao

and Schafer 2013) can create multiple imputations from multilevel models. Imputing from a multilevel regression model generally produces satisfactory results that are more accurate than those derived from imputation ignoring the nested data structure or deletion (van Buuren 2011; Zhao and Yucel 2009).

A more straightforward approach, suggested by Allison (2002), entails simply applying MI or FIML as usual to the *wide formatted* dataset (i.e., a dataset in which rows represent participants and columns represent repeated measures). This method implicitly models time by imposing a *panel* structure on the data. Thus, imputations derived from this approach can be considered to arise from a *cross-lagged panel model*. In most applications, this approach will produce unbiased imputations because the wide formatting of the data allows the imputation model to leverage past and future information when filling-in the missing data. Because this approach can employ any available MI scheme, it also easily accommodates nonnormally distributed missing data (e.g., by treating the wide formatted data with MICE). Naively imputing data in the *tall format* (i.e., where rows represent participant by time intersections) is not generally appropriate because such disaggregated models ignore the additional temporal dependence in the data. This disregard will contribute to imputations with inaccurate variance estimates that will induce bias in the standard errors of the analysis model (van Buuren 2011).

The Inclusive PCA Auxiliary Approach

Both MI and FIML can struggle when there are a high number of variables relative to the number of observations. This problem is made worse by the fact that missing data analyses are only optimal when all important interaction and polynomial terms are included in the missing data model (Graham 2012; von Hippel 2009). If the number of variables is already relatively large, expanding the data set to include important nonlinearities can lead to an unmanageable number of variables. Howard et al. (2015) have proposed a powerful solution to this problem. First, the data set is extended to include all the necessary interaction and polynomial terms, then the missing data are roughly filled-in using a single, stochastic regression imputation; finally, a set of principle component scores are extracted. Provided that the number of component scores retained is large enough to capture the majority of the shared information in the original items, the raw auxiliaries can be discarded and the principle component scores can act as the sole auxiliary variables in the imputation model or the FIML model. This approach can be particularly effective when the high dimensionality of the data is induced by (1) scales with many highly correlated items that can be mostly described by a small number of principle components or (2) a large pool of potential auxiliary variables for which capturing all of the information is not important. Fortunately, these two characteristics describe many practical missing data problems, and the inclusive PCA auxiliary approach offers a promising solution to a difficult problem in missing data analysis.

Addressing Nonignorable Missingness

When missingness is nonignorable, only a limited set of options are available. If there is reasonable knowledge of the content area to guide decisions, plausible values can be manually substituted for the MNAR data. Alternatively, the additional information contributed by the nonresponse indicators can be explicitly included into the missing data treatment by applying *selection modeling* (Heckman 1976, 1979) or *pattern mixture modeling* (Little 1993). These methods can be difficult to utilize in practice, however, because they are very sensitive to strong and untestable assumptions (Enders 2010; Little 1995; Little and Rubin 2002). Yet, even when the data appear to follow an MNAR mechanism, special modeling schemes may not be necessary. Collins et al. (2001) showed that a MNAR mechanism can be effectively transformed into a MAR mechanism if good "proxy indicators" of the missingness (i.e., variables that contain similar information to the MNAR variable) are used as auxiliary variables. This last point again highlights the importance of planning for missing data when designing scientific studies. Researchers can create an easily treated missing data problem by considering plausible predictors of the missingness during the research design phase and proactively including these variables in the data collection.

Multiple Imputation for Intent-to-Treat Analysis

In addition to salvaging inferences when faced with arbitrary nonresponse, MI can also be used to facilitate valid *intent-to-treat* (ITT) analyses. To implement such an analysis, the outcome data for those participants who dropped out of the study must be approximated or implied. This task is one that modern missing data treatments (especially MI) are ideally suited to perform. The simplest way to conduct MI-based intent-to-treat analyses is to impute the additional missing data that arise from attrition along with the arbitrary nonresponse that occurs elsewhere on the data set. In studies affected by *random drop out* (i.e., a type of MAR nonresponse in which the attrition is not directly associated with the treatment or complications thereof), employing a principled missing data method and incorporating correlates of the attrition into the imputation model will ensure optimal intent-to-treat inferences (Diggle and Kenward 1994; Little and Yau 1996).

When faced with *informative drop out* (a type of MNAR nonresponse in which drop out is directly related to the treatment), simply including treatment *as randomized* can bias the final intent-to-treat inferences if the dropouts end up receiving a different treatment after they leave the study. Consider the

hypothetical teen pregnancy prevention study presented above. If intervention group students who drop out of the study are left in the treatment group (i.e., treatment as randomized) for the ITT analysis, then their imputed outcome data will contaminate the estimated treatment effect. These students probably did not maintain exposure to the intervention after leaving the study, so the possible change in their treatment levels should be overtly incorporated into the imputation model. Explicit models for MNAR missingness (e.g., pattern mixture models) can be applied to informative drop out problems (Little 1995), but they may not be necessary. The research team will often have expert knowledge to suggest likely values for unobserved predictors. For example, in this hypothetical intervention, it might be reasonable to assume that, after leaving the intervention, students received the same treatment as the control group. Little and Yau (1996) suggest deterministically introducing this auxiliary information into the imputation model (e.g., by moving the dropouts into the control group for the ITT analysis) in order to increase the plausibility of ITT inferences. The result's sensitivity can be elucidated by repeating the analysis with different extrapolated treatment levels.

Multiple Imputation with Outcomes and Mediators

The fact that MI is implemented entirely during data preprocessing is one of its greatest strengths. At the data preprocessing stage, an imputation model can be specified that is much more complicated than the final inferential model. Imputing under a complex model allows the complete data sufficient statistics to be reproduced as faithfully as possible, independent of the choice of analysis model (Rubin 1996). Also, as Honaker and King (2010) discuss, MI is based on systems of predictive, rather than causal, equations. Predictive equations make MI agnostic with regard to whether a variable is a predictor, mediator, or outcome, so it is perfectly acceptable to impute variables that will enter the final analysis model as outcome variables or as mediators.

How to Treat Outcome Variables?

It is worth discussing the apparently "unfair" advantage that may be induced by imputing outcome variables as linear combinations of their hypothesized predictors. This concern is valid with single, deterministic regression imputation that will tend to inflate linear associations between the imputed variables and those that were used as predictors in the imputation model (Enders 2010; van Buuren 2012). For well-implemented MI, however, this inflation is not a problem. First, by including a large pool of auxiliary variables, the imputed values will reflect, as generally as possible, the true patterns in the data, rather than spuriously amplifying the hypothesized associations. Second, because MI quantifies all sources of uncertainty introduced by the missing data, it employs a type of implicit "self-correction" that mitigates spurious inflation of the linear associations (Allison 2002). Consequently, the current consensus among missing data researchers is to impute incomplete outcome variables (Allison 2002; Enders 2010; von Hippel 2007).

Limitations of Modern Missing Data Methods

The primary limitation of modern missing data methods is computational effort. Because MI is a highly iterative algorithm, it will be more demanding than alternative approaches that require minimal iteration. This limitation, however, does not outweigh the overwhelming benefits that come with modern, principled missing data methods. Moreover, FIML estimation does not entail substantially more computation than other ML-based analyses, so this limitation does not apply with FIML. MI and FIML also require ignorable missing data in order to perform optimally, but we reiterate that many otherwise nonignorable missing data problems can be made ignorable by including appropriate auxiliary variables in the missing data analysis. Yet, this possibility will only exist with careful planning for the inevitable missing data. We strongly recommend planning for missing data in the initial design phase and continually considering strategies to minimize the impact of nonresponse and optimize the final missing data analysis. The quality of the missing data analysis can impact the veracity of a study's conclusions as much as any other aspect of the research design, so it should certainly be considered just as carefully.

Conclusion

We have considered many issues that surround missing data problems in prevention research and have emerged with a singular recommendation. Prevention research will elevate the quality of its evidence base for guiding practice and policy if missing data are proactively anticipated and modern and principled treatments for missing data are routinely and appropriately utilized. Thus, for the sake of the stakeholders, we recommend that all future publications in journals such as *Prevention Science* should be required to implement one of the principled approaches we have outlined herein, and we implore all prevention scientists to actively plan their missing data analyses with the same care that they devote to planning substantive analyses.

Acknowledgments The authors wish to acknowledge the diligent assistance of Jacob Curtis, Brooke Bell, Naomi Norwid, Virginia Stokes, and Jacquelyn Wall in preparing the systematic literature review presented in this article.

Compliance with Ethical Standards

Conflict of Interest Todd D. Little owns and receives remuneration from Yhat Enterprises (yhatenterprises.com), which runs educational workshops such as Stats Camp (statscamp.org), and processes his royalties and his fees for consulting on statistics and methods with life science researchers.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Funding This study was supported by grant NSF 1053160 (Wei Wu and Todd D. Little, co-PIs) and by the Institute for Measurement, Methodology, Analysis, and Policy (Todd D. Little, Director) at Texas Tech University.

References

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage Publications.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association, 52*, 200–203. doi:10.1080/01621459.1957.10501379.

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review, 78*, 40–64. doi:10.1111/j.1751-5823.2010.00103.x.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bodner, T. E. (2006). Missing data: prevalence and reporting practices. *Psychological Reports, 99*, 675–680. doi:10.2466/PR0.99.7.675-680.

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. Chichester, West Sussex: Wiley.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*, 330–351. doi:10.1037//1082-989X.6.4.330.

Diggle, P., & Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics, 43*, 49–94.

Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement, 61*, 713–740. doi:10.1177/00131640121971482.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*, 430–457. doi:10.1207/S15328007SEM0803_5.

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling., 9*, 173–197. doi:10.1177/1471082X0800900301.

Goldstein, H., Carpenter, J., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 177*, 553–564. doi:10.1111/rssa.12022.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling, 10*, 80–100. doi:10.1207/S15328007SEM1001_4.

Graham, J. (2012). *Missing data: analysis and design*. New York: Springer.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*, 206–213. doi:10.1007/s11121-007-0070-9.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement, 5*, 475–492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47*, 153–161. doi:10.2307/1912352.

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science, 54*, 561–581. doi:10.1111/j.1540-5907.2010.00447.x.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: a program for missing data. *Journal of Statistical Software, 45*, 1–47.

Howard, W., Rhemtulla, M., & Little, T. D. (2015). Using principal components as auxiliary variables in missing data estimation. *Multivariate Behavioral Research, 50*, 285–299. doi:10.1080/00273171.2014.999267.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association, 88*, 125–134. doi:10.2307/2290705.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90*, 1112–1121. doi:10.1080/01621459.1995.10476615.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.

Little, R. J. A., & Yau, L. (1996). Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics, 52*, 1324–1333. doi:10.2307/2532847.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology, 39*, 151–162. doi:10.1093/jpepsy/jst048.

Little, T. D., Lang, K. M., Wu, W., & Rhemtulla, M. (2016). Missing data. In D. Cicchetti (Ed.), *Developmental Psychopathology: Vol. 1. Theory and method* (3rd ed., pp. 760–796). New York: Wiley.

Liu, M., Taylor, J. M. G., & Belin, T. R. (2000). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics, 56*, 1157–1163. doi:10.1111/j.0006-341X.2000.01157.x.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: a review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525–556. doi:10.3102/00346543074004525.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*, 85–96.

Rubin, D. B. (1978). *Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse* (Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 30–34).

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473–489. doi:10.2307/2291635.

Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling, 19*, 477–494. doi:10.1080/10705511.2012.687669.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman Hall.

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics., 11*, 437–457. doi:10.1198/106186002760180608.

van Buuren, S. (2011). Multiple imputation of multilevel data. In J. Hox & J. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173–196). Milton Park, UK: Routledge.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1–67.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*, 1049–1064. doi:10.1080/10629360600810434.

von Hippel, P. T. (2007). Regression with missing Ys: an improved strategy for analyzing multiply imputed data. *Sociological Methodology, 37*, 83–117. doi:10.1111/j.1467-9531.2007.00180.x.

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology, 39*, 265–291. doi:10.1111/j.1467-9531.2009.01215.x.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist, 54*, 594–604. doi:10.1037//0003-066X.54.8.594.

Wu, W., Jia, F., & Enders, C. K. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research, 50*, 484–503. doi:10.1080/00273171.2015.1022644.

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A, 366*, 2389–2403. doi:10.1098/rsta.2008.0038.

Zhao, J. H., & Schafer, J. L. (2013). *pan: multiple imputation for multivariate panel or clustered data (Version 0.9) [R Package]*.

Zhao, E., & Yucel, R. M. (2009). Performance of sequential imputation method in multilevel applications. In *the Proceedings of the American Statistical Association Survey Research Methods Section* (pp. 2800–2810).