Principled Missing Data Treatments

Kyle M. Lang[1] and Todd D. Little[2]

[1] Department of Psychology, University of Kansas

[2] Institute for Measurement, Methodology, Analysis, and Policy at Texas Tech University

November 28, 2014

Author's note.

This study was supported by grant NSF 1053160 (Wei Wu & Todd D. Little, co-PIs) and by the Institute for Measurement, Methodology, Analysis, and Policy (Todd D. Little, Director) at Texas Tech University. Correspondence can be directed to either author: Kyle M. Lang (kylelang@ku.edu) or Todd D. Little (yhat@ttu.edu or yhat@statscamp.org).

Abstract

We review a number of issues regarding missing data treatments for intervention and prevention science researchers. Many of the common practices in prevention science research are, in fact, ill-advised. 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 science research.

Keywords: Missing Data, Multiple Imputation, Full Information Maximum Likelihood, Auxiliary variables, Intent-to-treat, Statistical inference

Principled Missing Data Treatments

Missing data are a common problem for prevention science 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, the task of missing data analysis is a ubiquitous data pre-processing step for which many powerful methods have been developed (e.g., multiple imputation [MI] – Rubin, 1978, 1987; the expectation maximization [EM] algorithm – Dempster, Laird, & Rubin, 1977; full information maximum likelihood [FIML] – Anderson, 1957; and multiple imputation with chained equations [MICE] – Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). When applied correctly, these principled approaches to missing data treatment can help recover the underlying inferential model even in the presence of high rates of nonresponse (Little, Lang, Wu, & Rhemtulla, in press) and maximize a study's validity.

We review current best practice in missing data analysis, with a specific focus on applications in the prevention sciences. Treating missing data correctly is not only necessary to ensure the validity of scientific research but is also an ethical obligation of all research scientists. Improperly handling nonresponse can substantially compromise a study's inferential conclusions, and we suggest that doing so is an egregious form of data misrepresentation. The methods we discuss below are easily implemented and well suited to the types of analysis that are common in prevention science research, and they perform very well in those circumstances (Enders, 2010; Graham, 2012). The techniques we recommend will optimize the veracity of inferences in terms of four important quantities: bias, efficiency, validity, and statistical power.

In the following, we highlight some important characteristics of applied missing data problems and introduce the two flagship methods of modern missing data analysis, namely, explicit model-based multiple imputation (MI) and full information maximum likelihood (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, and nonresponse weighting.

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. In order to ground our recommendations firmly in missing data theory, we will first discuss those problem components that will play the largest role in the design and execution of 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, and easiest to fix, of these patterns is *univariate nonresponse* in which missingness occurs on only one variable. A second nonresponse pattern is the so-called *monotone nonresponse* pattern, which is characterized by a monotonic decrease in the completeness of the variables when moving from one side of a rectangular data set to the other. Such patterns are common in longitudinal research where they arise from *attrition* (i.e., participants permanently dropping out of the study). The final, and most common, nonresponse pattern is *arbitrary nonresponse*, which occurs when cells of the data set go missing in arbitrary, apparently random, arrangements—though this evident randomness is rarely truly random, as we discuss below. In our experience, this pattern is, by far, the most common in real-world missing data problems, where it can be readily treated with modern, principled missing data tools.

There is an alternative, binary, classification of nonresponse that can also help guide the decisions underlying 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 patient in a clinical trial, a designated respondent in a sample survey study) fails to give any data. Thus, unit nonresponse leads to entire rows of rectangular data sets being missing. Item nonresponse, on the other hand, occurs when individual cells in a data set are empty, but each row contains at least one observed data element. Unit nonresponse is actually a degenerate special case of monotone missingness, while item nonresponse subsumes the typical presentations of univariate, monotone, and arbitrary missingness. Item nonresponse is much more common than unit nonresponse in the experimental and quasi-experimental designs that are typical for prevention science research.

Nonresponse Mechanism. One of the most important issues to consider when planning a missing data analysis is the underlying reason that the data are incomplete, that is, the *nonresponse mechanism*. In the real world, nothing happens without a cause, and missing data are no exception. Every missing datum is unobserved for a specific reason, and the nonresponse mechanism is a model of this reason. There are three such mechanisms: *missing at random* (MAR), *missing completely at random* (MCAR), and *missing not at random* (MNAR). Each mechanism is defined according to a specific pattern of predictive relationships between the observed data and the probability of nonresponse.

MAR is the most general (and probably most common) nonresponse mechanism. Missing data that arise via a MAR process are predictable from other variables on the data set. Thus, for MAR missingness, the probability of nonresponse can be modeled by a standard logistic regression. MCAR missingness is actually a special case of MAR missingness that occurs when the nonresponse is a purely random sample of the complete data. In this sense, the probability that an individual cell in the complete data set is missing can be modeled as an independent

Bernoulli trial. The final nonresponse mechanism, MNAR, is the least desirable situation to encounter in practice. MNAR missingness occurs when the probability of nonresponse is predictable only by the incomplete variable itself. This case is nearly impossible to treat well because the missing values have no observed predictors on the data set, so the observed data cannot provide any information about the characteristics of the missing data's distribution. If the researchers plan for the inevitable missing data, however, correlates of the missingness mechanism can be proactively included in a study protocol. These correlates can often bring an otherwise MNAR mechanism back into the realm of a MAR mechanism.

Neither MCAR nor MNAR missing data have any observed predictors on the data set, so it is impossible to distinguish between the two mechanisms analytically. This similarity between the MNAR and MCAR mechanisms motivates a further refinement of the MNAR case. Enders (2010) distinguishes between *direct* and *indirect* MNAR mechanisms. Direct MNAR occurs as described above: the participants' levels of the incomplete variable are directly keeping them from responding. 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 incomplete variable and the propensity to respond, but both of these are related to an unmeasured third variable that induces a spurious association that manifests as an MNAR process.

Although various inferential tests are available and can be easily implemented to examine the missing-data mechanisms, we strongly caution against placing too much trust in them. The very nature of the three different nonresponse mechanisms makes it impossible to isolate a single causal agent for any given missing data problem. One strategy that is often recommended entails using logistic regression to predict each binary nonresponse indicator from all of the observed variables. One might naively conclude that finding any significant predictive relationships would indicate a MAR mechanism while finding none would indicate an MCAR mechanism, but this conclusion would be erroneous. While finding significant predictors does rule out a totally MCAR process, it does not fully support a conclusion of MAR because one can never be sure that all (or even the most important subset) of the predictors of nonresponse have been measured. Therefore, this approach can never fully differentiate between MAR and MNAR. If this test finds no significant predictors, the conclusions are no more certain. Although the MAR case is now ruled out (assuming adequate power for the logistic regression), there is no way to differentiate between MCAR and MNAR. The only way to achieve a definitive test of the nonresponse mechanism would be to know the true distribution of the complete data (both the missing and observed components) which can never occur in practice.

MCAR and MAR are *ignorable* mechanisms because their effects on bias, validity, precision, and power can be mitigated by techniques that do not require an explicit model for the nonresponse process (i.e., properly employed MI or FIML). MNAR, on the other hand, is *nonignorable* because it will lead to biased results unless the missing data analysis incorporates an explicit, and correct, model for the nonresponse process or additional variables are introduced that correlate strongly enough with the MNAR process to transform it into a MAR process. Here, planning for missing data would involve measuring potential correlates of MNAR missingness to approximate the MAR process and thereby reduce bias and increase validity.

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. Percent missing is an important early screening measure that gives a rough idea of the severity of the missing data problem and is critical to approximating the expected *fraction of missing information*. Thus, it is always one of the first quantities to be calculated during a missing data analysis. Yet, percent missing, alone, gives little information on how well the missing data treatment will perform or how the missing data model should be parameterized, because the raw percent missing does not take into account 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 simply gives the proportion of observations that are available to estimate each pairwise relationship. Scrutinizing the covariance coverage is important because low coverage values indicate that the observed data offer little information to help the estimation process. So, relationships with low coverage values will tend to be poorly recovered by most missing data treatments.

The most important measure of nonresponse rate is the fraction of missing information (FMI). The 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 & Rhemtulla, 2012), the convergence rates of missing data algorithms (Schafer, 1997), and the number of imputations required when using multiple imputation (Graham, Olchowski, & Gilreath, 2007).

Although the FMI can be readily estimated as a byproduct of both MI- and FIML-based missing data analyses, it can only be approximated before the analysis is complete. The maximum value for the FMI of univariate quantities (e.g., means and variances of incomplete variables) is the proportion of missing data (i.e., percent missing/100). The FMI will achieve this upper bound only when the missing data follow an MCAR or MNAR mechanism (so that the propensity to respond has no measured predictors) and the incomplete variables themselves have no correlations with other variables on the data set (so that they cannot borrow *proxy information* from these covariates). To the extent that either the response propensity or the incomplete variables themselves are predictable from other measured variables, the FMI will be lower than the proportion of missing data. Likewise, the maximum value for the FMI of bivariate quantities (i.e., covariances between two incomplete variables) is two times the proportion of missing data. Although it is possible for the FMI to be higher than the percent missing, in real-world missing data problems it is common for the FMI to be equal to or less that the percent missing, especially when the missing data follow a well predicted MAR process (Enders, 2010).

There are two commonly reported measures of nonresponse rate that we feel are misapplied so often that they tend to do more harm than good: proportion of complete cases and attrition rate. These metrics are problematic because they encourage researchers and research consumers to view missing data problems (and missing data treatments) from an inappropriate perspective. Of these two metrics, the attrition rate is the more useful measure. The attrition rate only applies to longitudinal data and simply quantifies the proportion of participants who permanently “drop out” of the study at each measurement wave. Reporting attrition rates is a necessary part of conducting transparent science, but the raw attrition rate should play almost no part in the data analytic decisions. Studies suffering from attrition will also tend to have low coverage values and high percent missing and FMI values; these latter quantities should be consulted when designing the missing data treatment.

The proportion of complete cases is an insidious measure that we strongly recommend against using. It simply gives the proportion of observations that contain no missing data. This measure is problematic for a number of reasons, not the least of which is that it makes nearly all missing data problems appear much more severe than necessary. We have encountered a number of real-world missing data problems where the proportions of complete cases were nearly zero, but the final, principled, missing data treatments were entirely successful. Another major reason to avoid this measure is that the information it offers is only applicable to listwise deletion, which should never be used in practice.

Antiquated Missing Data Treatments

In order to highlight the relative strengths of modern missing data treatments, we describe several antiquated ad hoc approaches that, unfortunately, remain common in the literature.

Deletion-Based Techniques. Missing data theorists have long decried deletion-based techniques as some of the worst options for treating missing data (Wilkenson & Task Force on Statistical Inference, 1999). Unfortunately, they still remain common in many scientific studies (Bodner, 2006; Little, Jorgenson, Lang, & Moore, 2013). Deletion-based techniques come in two flavors, listwise deletion (or complete case analysis) and pairwise deletion (or available case analysis). Listwise deletion is the more insidious of the two since it (1) leaves nonresponse bias unaddressed and thus leads to biased statistical inferences unless the data are MCAR (Little & Rubin, 2002), and (2) 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 not necessarily increase bias in inferences, but it can lead to sufficient statistics with inconsistent degrees of freedom. This inconsistency can produce estimated correlations outside of the interval [-1, 1] and sample covariance matrices that are not positive definite (Little & Rubin, 2002). Neither deletion approach attempts to remedy a MAR process.

Single Imputation Techniques. Single imputation techniques will usually lead to very poor results, and they are not candidates for general-purpose missing data treatments. There are three commonly employed types of single imputation: unconditional mean substitution, conditional mean substitution (i.e., deterministic regression imputation), and stochastic 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 can lead to inflated Type I error rates because it does not adequately quantify the uncertainty introduced by the missing data (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).

Last Observation Carried Forward. LOCF is a deterministic single imputation technique that has long been a popular method for treating attrition in longitudinal studies. It simply entails filling in all of an observation’s post drop-out missing values with its last observed value. This method represents disastrously bad practice that can seriously compromise a study’s inferences and lead to highly invalid conclusions (Enders, 2010; Little & Yau, 1996, van Buuren, 2011). The implicit assumption underlying 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 it is trivial to construct circumstances where LOCF will lead to liberal bias. One only needs to consider a treatment for a degenerative condition. If all participants with the condition are expected to demonstrate a monotonic decrease in some outcome measure over the course of the study (e.g., cognitive functioning in people with Alzheimer's), and the effect of the treatment is simply to slow this degeneration, then freezing drop-outs' responses at an early measured level will spuriously inflate the treatment's measured effect. Little and Yau (1996) described a randomized controlled trial of an Alzheimer's drug that demonstrated exactly this pattern. LOCF should be universally rejected as a missing data treatment simply because its underlying premise is so blatantly incongruent with the reality of longitudinal processes.

Nonresponse Weighting Approaches. Weighting techniques were primarily developed to address unit nonresponse (Little & Rubin, 2002). When a sampled unit gives no data (i.e., the whole row of data is missing), weighting can be a viable option. Weighting shares one of the end goals of MI, namely, correcting the unit nonresponse bias. So, it is useful to consider the specific sets of circumstances that would lead one technique to outperform the other.

When can nonresponse weighting outperform MI? First, if there are no auxiliary data available for the missing unit, then MI is intractable. If this unit is missing from an MCAR process, however, then there is no nonresponse bias to correct, and the weights are unnecessary. Yet, if the unit is missing from a MAR process, then weighting does nothing to correct for the nonresponse bias because it also relies on auxiliary variables for this purpose. So, both weighting and MI require a good set of auxiliary data to effectively correct the nonresponse bias. The one circumstance in which weighting *may* outperform MI arises if the auxiliaries are predictive of the nonresponse mechanism but uncorrelated with the incomplete variables (though we question how often such a premise would arise in practice). In such cases, MI can only replace the missingness with simulated values from the sampling distribution of the incomplete variable's mean, and well-constructed weights *might* do a better job of correcting for the nonresponse bias; however, Little and Vartivarian (2005) demonstrate that weights constructed from auxiliary variables of this type (i.e., that are uncorrelated with the incomplete variables) will lead to over-estimated variances for the incomplete items. So, for weighting to perform optimally, it needs a good set of auxiliaries that predict both the nonresponse mechanism and the incomplete items. Yet, with such a set of auxiliary data, MI will also perform optimally.

When can MI outperform nonresponse weighting? There are several common characteristics of real-world missing data problems that will cause MI to outperform nonresponse weighting. First, nonresponse weighting can only be expected to perform equally to or better than MI under unit nonresponse. Unit nonresponse is quite rare outside of randomization-based survey research, and most missing data problems are subject to some degree of arbitrary nonresponse—even when the majority of the nonresponse arises due to attrition or unit nonresponse. Because this type of item nonresponse is exactly what MI was designed to address, MI will be preferred to weighting for prevention science research. Because real-world nonresponse is usually peppered throughout an entire data set, the number of complete cases is often quite small relative to the proportion of observed cells in the data. This final point is important because it highlights a significant limitation of weighting-based approaches, namely, these methods are really just small modifications of listwise deletion.

A good set of auxiliaries can support a weighting-based treatment that corrects for the nonresponse bias and produces reasonable estimates of the variance (Little & Vartivarian, 2005), but such an analysis would still suffer from two fatal flaws. First, nonresponse weighting-based analyses, just like those employing listwise deletion, can lead to a substantial loss of power. If the weights are adjusted to reflect the final number of complete cases (as is common practice), then nonresponse weighting suffers from the same degree of power loss as listwise deletion and can restrict detectable findings to only large effects. Because MI builds the incomplete data back up to the same dimensions as the anticipated sample, it does not suffer from this same loss of power. When MI is used with a large enough number of imputations (we recommend m = 100), power can be maintained at near optimal levels (Graham, et al., 2007).

A second limitation of weighting comes from the implicit restriction on the size and number of the adjustment classes that any weighting scheme can produce. Nonresponse weighting can only correct for the nonresponse bias when the number of adjustment cells estimated makes the probability of nonresponse approximately constant within each cell (Little & Rubin, 2002). In data sets with high nonresponse rates, the small number of complete observations can support only a few adjustment cells. We urge caution when considering the ability of a small number of adjustment cells to adequately satisfy the MAR assumption.

Finally, because missing data patterns are often arbitrary, weighting schemes can get quite complex. Researchers who still wish to (mis-)apply weighting as the primary missing data treatment in these cases may consider ad hoc solutions. For example, a unique set of weights could be estimated for each pattern of missingness, or some of the missingness could be imputed so that the nonresponse pattern is monotone, and the weighting could then be applied more efficiently. These approaches would accommodate the weighting strategy, but we feel that the more parsimonious solution is simply to employ MI directly. After all, arbitrary nonresponse patterns are exactly the sort that MI was designed to address.

Recommended Missing Data Treatments

As intimated throughout, there are two flagship techniques in modern missing data analysis: multiple imputation (MI) and full information maximum likelihood (FIML). These methods provide optimal results in nearly all missing data problems, and we emphatically advocate their use whenever missing data occur in applied research. FIML is very 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. A MI routine can be tailored to address essentially any missing data problem one could encounter.

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 & 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 inference.

Full Information Maximum Likelihood. FIML (Anderson, 1957; also known as Direct

Maximum Likelihood) is a maximum likelihood estimator that is robust to nonresponse. It is a clever extension of ordinary maximum likelihood estimation that modifies the sample loglikelihood 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 & Rhemtulla, 2012). In practice, FIML has been shown to perform very well (Arbuckle, 1996; Enders, 2001a; 2001b; Enders & Bandalos, 2001). Under a MAR response 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 (Enders, 2008; Savalei & Rhemtulla, 2012).

Choosing between FIML and MI. FIML performs very well when its assumptions are met, but there are circumstances where MI is preferred. Because the underlying FIML objective function is derived from the multivariate normal distribution, FIML requires normally distributed variables to operate at full capacity. Although FIML will be robust to moderate violations of normality in the form of skewed or kurtotic data (Enders, 2001a), its standard implementation cannot be used when the data are categorical and modeled as such. The capabilities of FIML estimation can be extended to categorical data if one is able to manually program the likelihood function, but doing so can be very challenging for complicated models. This difficulty limits many researchers' abilities to apply FIML to non-normal data, whereas MI, especially when conducted within the MICE framework, easily accommodates categorical distributions for the missing data (van Buuren, 2012). 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 missingness out of the likelihood function while estimating the analysis model, but it never "fills in" any of the missing cells. Thus, there is no readily apparent way to aggregate the incomplete items (e.g., 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).

Choice of Imputation Model: Normal, Categorical, or Implicit? When using ordinary FIML, the missing data must be modeled by a multivariate normal distribution. MI, on the other hand, is much more flexible in readily available implementations. 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 science research are not continuous or normally distributed (e.g., Likert-type items, gender, patient survival). 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, Enders, and Jia (2013) 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 led to unbiased and efficient parameter estimates and outperformed imputation methods that overtly 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 & 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 will end up re-using many donor observations, and the standard errors of the analysis model parameters will be attenuated (van Buuren, 2012).

Addressing Temporal Dependence among the Missing Data. When the incomplete data are longitudinal in nature, additional care must be taken to preserve the temporal dependence of the imputed values. 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, and this flexibility allows one to impute longitudinal missing data according to a model that incorporates whatever function of time is deemed appropriate. To fully capture this capability, however, requires that the data analyst program their own implementation. Fortunately, there are also several solutions available in common MI software packages.

The R package **Amelia II** (Honaker, King, & Blackwell, 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 its own trend. Honaker and King (2010) demonstrated the effectiveness of this approach for normally distributed missing data.

Because longitudinal data can be viewed as repeated measures nested within individual, another convenient class of imputation model is multilevel regression models (also known as mixed effects models, hierarchical linear models, and growth curve models). Goldstein, Carpenter, and Browne (2014), Goldstein, Carpenter, Kenward, and Levin (2009), Liu, Taylor, and Belin (2000), Yucel (2008), and Schafer and Yucel (2002) all developed multiple imputation methods based on multilevel regression models that can be applied to longitudinal nonresponse. The R packages **mice** (van Buuren & Groothuis-Oudshoorn, 2011) and **pan** (Zhao & Schafer, 2013) as well as the stand-alone package **REALCOM-IMPUTE** (Carpenter, Goldstein, & Kenward, 2011) are all capable of creating multiple imputations from multilevel regression models. Imputing from a multilevel regression model generally produces satisfactory results that are more accurate than imputation ignoring the nested data structure and deletion based techniques (Black, Harel, & McCoach, 2010; van Buuren, 2011; Zhao & Yucel, 2009).

Another approach entails converting the incomplete data into *wide format* (i.e., where rows represent participants and columns represent repeated measures), and simply applying MI or FIML as usual (Allison, 2002). 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 non-normally 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 totally ignore the additional temporal dependence in the data. This disregard will contribute to imputations with inaccurate variance estimates that will, in turn, lead to 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, but Howard, Rhemtulla, and Little (in press) have proposed a very powerful solution to this problem. Once the data set has been augmented by including all the necessary interaction and polynomial terms, its principle components are extracted. Provided the number of components is taken to be large enough to capture the majority of the shared information in the items (empirical evidence suggests that ten component is usually sufficient), the raw auxiliaries can be discarded and the principle components can act as the sole predictors in the imputation model or the sole auxiliary variables in 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 of paramount importance. Fortunately, these two characteristics describe many of the practical missing data problems encountered in prevention science research, and the inclusive PCA auxiliary approach offers a promising solution to a very difficult problem in applied missing data analysis.

Addressing Nonignorable Missingness. When missingness arises from an MNAR mechanism, it is said to be *nonignorable* in the sense of Rubin (1976), because the nonresponse mechanism must be overtly modeled as part of the missing data analysis in order to accurately recover the missing values (i.e., the nonresponse mechanism cannot be *ignored*). When missingness is nonignorable, only a limited set of options are available. If the data analyst possesses reasonable knowledge of the content area to guide their decisions, plausible values can be manual substituted for the MNAR missingness. Alternatively, the nonresponse mechanism can be included as part of the imputation model through *selection modeling* or *pattern mixture modeling*. These methods can be difficult to apply in practice, however, because they are very sensitive to strong and untestable assumptions (Enders, 2010; Little, 1995; Little & Rubin, 2002). Yet, even when the data appear to follow an MNAR mechanism, special imputation schemes may not be necessary. Collins, Schafer, and Kam (2001) showed that a MNAR mechanism can be effectively transformed into a MAR mechanism if good “proxy indicators” of the nonresponse mechanism are included as auxiliary variables during the missing data analysis. This last point again highlights the importance of planning for missing data when designing scientific studies. Researchers can “stack the deck” toward an easily treated missing data problem by considering plausible predictors of the nonresponse 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* 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 are ideally suited to perform and antiquated ad hoc missing data methods (e.g., LOCF) are woefully ill-equipped to address. 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 MAR process in which the attrition is not directly caused by 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 & Kenward, 1994; Little & Yau, 1996).

When faced with *informative dropout* (an MNAR process in which dropout is directly related to the treatment), simply including treatment *as randomized* can bias the final intent-to-treat inferences if the drop-outs end up receiving a different treatment after they leave the study. For example, participants in a drug trial who experience severe side-effects may leave the study and stop taking the drug that they were assigned (or simply reduce the dosage). If the intent-to-treat analysis only includes dosage information as randomized, then the imputed outcome data for these participants will be associated with the wrong dosage group, and the validity of the inferences will be compromised. When such informative dropout presents, the possible change in the treatment levels for the drop-outs must be overtly incorporated into the imputation model. Although explicit models for MNAR missingness (e.g., pattern mixture models) can be applied to informative drop-out problems (Little, 1995), these techniques are very sensitive to violations of their assumptions (Enders, 2010; Little, 1995). Fortunately, the research team will often have expert knowledge to suggest likely values for unobserved predictors (e.g., in a drug trial, it may be possible to bound the dosages that noncompliant participants would maintain). Little & Yau (1996) suggest deterministically introducing this auxiliary information into the imputation model in order to increase the plausibility of the intent-to-treat inferences. The sensitivity of the final result can be elucidated by repeating the analysis with different extrapolated treatment levels.

Multiple Imputation with Outcomes and Mediators. One of the greatest flexibilities of MI is that it is implemented entirely during data pre-processing. This means that the data analyst can specify an imputation model 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. This makes 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 should outcome variables enter the imputation model? 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 regression-based imputation strategies that will tend to inflate linear association 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 spurious 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 pattern of interrelationships in the data, rather than spuriously amplifying the hypothesized associations. Second, because MI quantifies all sources of uncertainty introduced by the missing data, it can be thought of as employing an implicit “self-correction” that effectively mitigates spurious inflation of the linear associations (Allison, 2002).

The current consensus among missing data researchers is to impute incomplete outcome variables (Allison, 2002; Enders, 2010; Little, 1982; von Hippel, 2007), but some work has suggested that participants with imputed outcomes should be subsequently excluded from the analysis models (von Hippel, 2007). This argument is based on two premises. (1) Observations with missing outcomes contribute no information to the estimation of regression coefficients, and (2) including these observations will increase the uncertainty in the final estimates. For congenial imputation and analysis models (i.e., those that contain exactly the same variables) we do not question this reasoning. Yet, we suggest that most applied missing data problems do not conform to these conditions. Real-world missing data problems lend themselves to uncongenial imputation and analysis models wherein the imputation models have the possibility of employing considerably more variables than the analysis models. These situations admit the possibility of *superefficient* imputations in the sense of Rubin (1996). Such *overspecified* imputation models can leverage the *proxy information* contained in the auxiliary data to “recover” some of the missing outcome information, above and beyond what is implied by the congenial predictor set. von Hippel (2007) provides a small simulation to show that the intuition given above does hold. He shows that including just a single strongly predictive auxiliary variable can lead to a situation in which a traditionally implemented MI outperforms the method based on deleting participants with imputed outcome data. Thus, we suggest that incomplete outcomes should always be included in the imputation model and retained for the final analysis.

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.

Conclusion

We have delved deeply into the many issues that surround missing data problems in prevention science research and have emerged with a singular recommendation. Prevention science research will elevate the quality of its evidence base for guiding practice and policy if modern, principled treatments for missing data are routinely utilized. Thus, for the sake of the stake-holders, 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.

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*(278), 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*(1). 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.

Black, A. C., Harel, O., & McCoach, D. B. (2010). Missing data techniques for multilevel data: Implications of model misspecification. *Journal of Applied Statistics, 38,* 1845–1865. doi: 10.1080/02664763.2010.529882

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., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistcal Software, 45*(5). 1-12.

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

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

Enders, C. K. (2001a). The impact of nonnormality on full information maximum likelihood estimation for structural equations models with missing data. *Psychological Methods, 6,* 352–370. doi: 10.1037/1082-989X.6.4.352

Enders, C. K. (2001b). 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. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling, 15*(3), 434 – 448. doi: 10.1080/10705510802154307

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., & 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*(2), 553-564. doi: 10.1111/rssa.12022

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling. 9*(3), 173-197. doi: 10.1177/1471082X0800900301

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

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science, 54*(2), 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. (in press). Using principal components as auxiliary variables in missing data estimation. *Multivariate Behavioral Research*.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association, 77*(378), 237–250. doi: 10.2307/2287227

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., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology, 31*(2), 161-168.

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., Lang, K. M., Wu, W., & Rhemtulla, M. (in press). Missing data. In D. Cicchetti (Ed.), *Developmental Psychopathology* (3rd Ed., pp 000-000). New York: Wiley.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2013). On the Joys of Missing Data. *Journal of Pediatric Psychology,* 1-12, doi: 10.1093/jpepsy/jst048

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

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*(1), 85–96.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592. doi: 10.2307/2335739

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,* 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*(2). 437-457. doi: 10.1198/106186002760180608

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

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

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

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1-67.

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., Enders, C. K., & Jia, F. (2013). A comparison of imputation strategies to ordinal categorical data. *Paper presented at the annual meeting of the Modern Modeling Methods (M^3) Conference.*

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, 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*. 2800-2810.

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