Original Article

Sociological Methods & Research
1–31

DOI: 10.1177/00491241221113873
journals.sagepub.com/home/smr
SAGE

A Comparison of Three Popular Methods for Handling Missing Data: Complete-Case Analysis, Inverse Probability Weighting, and Multiple Imputation

Roderick J. Little [1], James R. Carpenter[2,3], and Katherine J. Lee[4], on behalf of the STRATOS initiative

Abstract

Missing data are a pervasive problem in data analysis. Three common methods for addressing the problem are (a) complete-case analysis, where only units that are complete on the variables in an analysis are included; (b) weighting, where the complete cases are weighted by the inverse of an estimate of the probability of being complete; and (c) multiple imputation (MI), where missing values of the variables in the analysis are imputed as draws from their predictive distribution under an implicit or explicit statistical model, the imputation process is repeated to create multiple filled-in data

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
[2]Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK
[3]MRC Clinical Trials Unit, UCL, UK
[4]Murdoch Children's Research Institute, Royal Children's Hospital, Melbourne, Australia

Corresponding Author:
Roderick J. Little, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor MI 48209, USA.
Email: rlittle@umich.edu

sets, and analysis is carried out using simple MI combining rules. This article provides a non-technical discussion of the strengths and weakness of these approaches, and when each of the methods might be adopted over the others. The methods are illustrated on data from the Youth Cohort (Time) Series (YCS) for England, Wales and Scotland, 1984–2002.

Keywords

incomplete data, imputation, missing data, weighting

Preliminaries

Missing data are a pervasive problem in statistical analysis. The topic has an extensive literature – textbooks on the topic include Little and Rubin (2019), van Buuren (2018), Raghunathan (2015), Carpenter and Kenward (2013), and Schafer (1997). We consider and compare three common approaches to the analysis of data with missing values, namely complete-case analysis (henceforth CC), inverse probability weighting (henceforth IPW), and multiple imputation (henceforth MI). In CC — or complete-record analysis (e.g. Carpenter and Kenward 2013, chapter 1) to avoid confusion with the terminology of cases and controls in medical studies — only units that are complete on the variables in an analysis are included; in IPW, the complete cases are weighted by the inverse of an estimate of the probability of being complete; and in MI, missing values of the variables in the analysis are imputed as draws from their predictive distribution under an implicit or explicit statistical model; the imputation process is repeated to create multiple filled-in data sets, and analysis is carried out using simple MI combining rules (Rubin 1987).

All methods for handling missing data make unverifiable assumptions; perhaps the closest to an assumption-free method is that in Horowitz and Manski (1998), which presents bounds on parameter inferences based on best and worst-case values of the missing variables. This method is (as they acknowledge) very conservative, and is essentially limited to missing variables that have known finite support.

Our focus here is principally on inference for regression coefficients and sample means. We restrict attention to models that assume the missingness mechanism is missing at random (MAR), as discussed in the next section, although it is important to recognize that missing not at random (MNAR) MI methods are possible as well. CC, IPW and MI are all quite general, in that (given sufficient information) they can be used to handle missing data in any

statistical analysis an analyst might wish to perform on the data without missing values.

Other valid approaches exist for handling missing data (by "valid" we mean that when its assumptions hold, a method yields consistent estimates of target parameters, confidence intervals with close to nominal coverage, and tests close to the stated size). Specifically, likelihoods can be defined for nonrectangular data sets with missing values, and hence methods based on these likelihoods can be implemented. In particular, maximum likelihood (ML) estimates can be computed, with standard errors based on the information matrix or sample re-use methods like the bootstrap; or a prior distribution can be added to the specification and inferences based on the Bayesian posterior distribution. Indeed, ML methods for missing data are quite widely used in the social sciences – often implicitly, as incorporated into structural equation modeling software like Mplus (Muthen and Muthen 2017). ML is asymptotically equivalent to MI under the same model for the data, so it shares some of the properties of MI discussed here. However, MI is more flexible than ML in some settings, because it allows variables not included in the final analysis model to be included in the imputation model and readily extends to settings where data may be MNAR. Augmented inverse-probability weighted estimating equations (Robins and Rotnitsky 1995; Robins, Rotnitsky and Zhao 1995) employ estimating equations that include model predictions of missing values and weighted residual terms, which provide some protection against model misspecification.

We focus on the three methods described above because they are used extremely widely. In particular, CC is the default method in much statistical software, is intuitive and is simple to implement. IPW is the standard approach to handling unit nonresponse in surveys, and is also relatively simple to carry out. MI methods are more varied and complex, but increasingly common because of extensive availability in computer software packages. For example, MICE and other R packages (Van Buuren and Groothuis-Oudshoorn 2011, Su et al. 2011), IVEware (Raghunathan et al. 2001), PROC MI in SAS (2015), and Stata (see https://www.stata.com/features/multiple-imputation/).

Additional modeling assumptions are unavoidable when analyzing data with missing values, so the most important step in dealing with missing data is to limit the extent of missing values, by careful design and data collection (e.g. National Research Council 2010). Because some data are likely to be missing despite these efforts, it is important to try to collect covariates that are predictive of the missing values, so that an adequate adjustment can be made. In addition, the processes that lead to missing values should be assessed during the collection of data if possible (e.g. Little 1995), because

this information plays a role in the choice of missing data adjustment method, as discussed further below.

A basic assumption in all missing-data methods is that missingness of a particular value hides a true underlying value that is meaningful for analysis. Deciding whether a value is meaningful is not always as simple as it seems. For example, consider a longitudinal analysis of measures of quality of life; for subjects who leave the study because they move to a different location, it makes sense to consider quality of life as missing, whereas for subjects who die during the course of the study, it is not reasonable to consider quality of life after time of death as missing. Rather it is preferable to restrict the analysis of quality of life to individuals while they are alive. More complex missing data problems arise when individuals leave a study for unknown reasons, which may include relocation or death. Another example is nonresponse to opinion polls, where the target population consists of individuals who will vote – nonresponse for people who do not vote is arguably not missing data, since an imputed value is not meaningful for estimating the proportion of votes cast for each candidate.

Despite the fact that CC, IPW and MI are common in practice, we believe that the principles underlying the choice between these methods are not as well understood as they might be. Therefore, this article provides a relatively nontechnical discussion of the strengths and weakness of CC, IPW and MI, and guidelines for when each of the methods might be favored over the others. For those who believe that the material is well known, here are four preliminary facts that may surprise some readers:

1. We use the term *auxiliary* variables to mean fully-observed variables used for imputation or weighting but not included in the substantive model of interest. IPW based on auxiliary variables is widely viewed as reducing bias in CC estimates. However, in many realistic survey settings where the auxiliary variables are strongly related to the propensity to respond and weakly related to the survey variable of interest, IPW actually leads to worse inferences than CC (see the subsection "Inference for the Mean of an Incomplete Variable Y" for details).
2. In the statistics literature, CC is widely criticized and seen as inferior to methods like MI that use all available data. However, for some regression problems CC is optimal, and MI is actually less, not more, efficient (see the subsection "Missing data in Regression" and Hughes et al. (2019)).
3. CC is often described as biased unless the data are missing completely at random, as defined in the next section, and IPW is widely viewed as

for reducing the bias of CC analysis. However, for some problems, CC analysis is actually less biased than IPW.

4. In some settings, a hybrid combination of CC and MI is less biased than CC, IPW or MI.

We expand upon points 2–4 in the subsection "Missing data in Regression". We illustrate the methods by analyzing data from a UK youth cohort study, and conclude by summarizing our recommendations concerning the methods.

Pattern and Mechanism of Missing Data

The *pattern* and *mechanism* of missing data are important features of the problem that play an important role in choosing between CC, IPW and MI. The *pattern* refers to which values in the data set are observed and which are missing. Specifically, let $Y = (y_{ij})$ denote an $(n \times p)$ rectangular dataset without missing values, with ith row $y_i = (y_{i1}, \ldots, y_{ip})$ where y_{ij} is the value of variable Y_j for subject i. With missing values, the pattern of missing data is defined by the *response indicator matrix* $R = (r_{ij})$, such that $r_{ij} = 1$ if y_{ij} is observed and $r_{ij} = 0$ if y_{ij} is missing; equivalently, $1 - r_{ij}$ is the *missing-data indicator* for y_{ij}.

Some methods for handling missing data apply to any pattern of missing data, whereas other methods assume a special pattern. A simple special pattern is *univariate* nonresponse, where missingness is confined to a single variable. Another example is *monotone* missing data, where the variables can be ordered so that $Y_{j+1}, \ldots, Y_p$ are missing for all subjects where Y_j is missing, for all $j = 1,\ldots, p$-1. Looking at the matrix R, the result is a "staircase" pattern, where all variables and units to the left of the broken line forming the generally irregular staircase are observed, and all to the right are missing. This pattern arises in longitudinal data subject to attrition, where once a person drops out, no more data are observed for that person.

The missingness *mechanism* addresses the reasons why values are missing, and whether these reasons relate to values in the data set. For example, subjects involved in a longitudinal intervention may be more likely to drop out of a study because they feel a treatment was ineffective, which might be related to a poor value of an outcome measure. Rubin (1976) treated R as a random matrix, and characterized the missingness mechanism by the conditional distribution of R given Y, say $\Pr(R|Y, \phi)$, where ϕ denotes unknown parameters. When missingness does not depend on the values of the data Y, missing or observed, that is,

$$\Pr(R|Y, \phi) = \Pr(R|\phi) \text{ for all } Y, \phi,$$

the missingness is called missing completely at random (MCAR). An MCAR mechanism is plausible in some planned missing-data designs, but is a strong and often unrealistic assumption, especially when missing data do not occur by design, because missingness often does depend on values of variables.

Using a slightly informal notation, let $Y_{(1)}$ denote the observed components of Y and let $Y_{(0)}$ denote the missing components of Y. A less restrictive assumption is that missingness depends only on values $Y_{(1)}$ that are observed, and given these not on values $Y_{(0)}$ that are missing. That is:

$$\Pr(R|Y_{(1)}, Y_{(0)}, \phi) = \Pr(R|Y_{(1)}, \phi) \text{ for all } Y_{(0)}, \phi. \quad (1)$$

The missing data are then called missing at random (MAR) at the observed values of R and $Y_{(1)}$. If (1) does not hold, the data are missing not at random (MNAR). Concerning our three compared methods, CC is always valid under MCAR, and in particular circumstances to be described, it may be valid under weaker assumptions about the missingness mechanism. The implementations of IPW and MI in widely-available software are valid under MAR (and hence under MCAR), and we restrict attention to these versions here; it is possible to develop versions of IPW and MI for MNAR mechanisms, but these lie outside the scope of this article.

Methods

Complete-Case (CC) Analysis

CC for a set of variables simply discards units where any of these variables are missing. It has the advantage of simplicity, and it is the default analysis in most statistical software packages. It has two main drawbacks. Firstly, the complete cases are not a random subsample of the original sample unless the data are MCAR. This is usually an unrealistic assumption, because cases with missing values often differ from complete cases in terms of the variables of interest. If the complete cases are not a random subsample, CC will give biased answers for simple summary measures (such as mean, sd) and may yield biased answers for regression models, although not in all situations, as discussed below. Secondly, CC discards information in the incomplete cases, which has typically cost non-trivial resources to collect, and which will often contain information for reducing bias or increasing the efficiency of CC estimates. A key question is thus how much information is contained in the incomplete cases – if nearly all the information is contained in the complete cases, CC might be a reasonable approach. Unfortunately, the answer to this question is not straightforward, because it depends on the

fraction of complete cases, the distribution of variables (observed and missing) in the incomplete cases, the missingness mechanism, and the nature of the specific analysis of interest. Some specific examples are provided below.

Inverse Probability Weighting (IPW)

A modification of CC, commonly used to handle unit nonresponse in surveys, is inverse probability weighting (IPW), which weights complete units by the inverse of an estimate of the probability of response (see e.g. Seaman and White 2011). In particular, when estimating a population mean, the sample mean is replaced by the weighted mean. IPW can also be applied to estimators other than means, such as regression coefficients, or more generally, estimators for generalized estimating equations (weighted GEE).

A simple approach for creating weights is to form adjustment cells (also called subclasses) based on background variables measured for both respondents and nonrespondents; for unit nonresponse adjustment, these are often based on geographical areas or groupings of similar areas based on aggregate socioeconomic data. All nonrespondents are assigned zero weight and the nonresponse weight for all respondents in an adjustment cell is then the inverse of the estimated response rate in that cell. For more details see Little and Rubin (2019, Example 3.6).

With more extensive background information, a generalization of adjustment cell weighting is *response propensity* stratification, where (a) the indicator for unit nonresponse is regressed on the background variables, using the combined data for respondents and nonrespondents, using a method such as logistic regression appropriate for a binary outcome; (b) a predicted response probability is computed for each respondent based on the regression in (a); and (c) adjustment cells are formed based on a categorized version of the predicted response probability. The creation of adjustment cells can be useful to reduce extreme weights, which otherwise can inflate the variance of weighted estimates. Theory (Rosenbaum and Rubin 1983) suggests that this is an effective method for removing nonresponse bias attributable to the background variables when unit nonresponse is MAR. Adjustment cell weighting is a special case of this method when the adjustment cell variables are indicators of the cells. In both methods, weights are rescaled so that they sum to the number of respondents.

Although IPW can be useful for reducing nonresponse bias, it does have serious limitations. First, information in the incomplete cases is only used to determine the weights (i.e. the weight model uses variables that are fully

observed on both respondents and non-respondents), and partially observed cases are still discarded in the weighted analysis. This fact means that the method is generally inefficient when, as will often be the case, there is substantial information in these partially observed cases. Therefore, weighted estimates can have unacceptably high variance, especially when extreme values of a variable are given large weights. For modifications of the inverse probability weights to increase efficiency, see for example Cao, Tsiatis and Davidian (2009).

Variance estimation for weighted estimates will ideally take into account uncertainty in the estimated weights, otherwise standard errors will be overestimated so inferences will be conservative. Approaches include Taylor series expansion (Robins, Rotnitsky and Zhao 1995) or computing bootstrap standard errors, with weights recalculated for each bootstrap sample (Little and Rubin 2019, Chapter 5).

Multiple Imputation (MI)

Methods that impute or fill in the missing values have the advantage that, unlike CC or IPW, observed values in the incomplete cases are retained to make full use of them in the analysis. In fact, the goal of MI is to preserve the information in the observed values for inference, not to get the best predictions of the missing values.

Because we can never recover the actual missing value, a single imputation for each missing value cannot reflect the imputation uncertainty, and as a result standard errors of estimates based on analysis of a single filled-in data tend to be underestimated. Large-sample results show that for simple situations with 30% of the information missing, single imputation under the correct model results in nominal 90% confidence intervals having actual coverages below 80% (Rubin and Schenker 1986). The inaccuracy of nominal levels is even more extreme in multiparameter testing problems (Rubin 1987, Chapter 4).

Multiple (as opposed to single) imputation fixes this problem (Rubin 1987, 1996, Schafer 1998). The basic steps of MI are to (a) estimate a predictive distribution for the missing values $Y_{(0)}$ given the observed values $Y_{(1)}$ in the data set— approaches to this are described below; (b) fill in, or impute, the missing values with draws from this predictive distribution (note that the imputations are draws, that is random selections from the predictive distribution, not means of the predictive distribution); (c) repeat step (b) $M > 1$ times (where, say, $M = 10$ or 20) to create M datasets, each containing different sets of draws of the missing values. For any particular analysis of the filled-in

data, we then apply the standard complete-data analysis to each of the M datasets, yielding M estimates, say $(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(M)})$ of parameters θ; (d) combine the parameter estimates to create an overall estimate of θ — a method for doing this is called a MI combining rule. In particular for scalar estimands, the MI estimate is the average $\hat{\theta}_{\mathrm{MI}} = \sum_{m=1}^{M} \hat{\theta}^{(m)} / M$ of the estimates from the M datasets, and the sampling variance of the estimate is estimated as $\hat{V} = \hat{W} + (1 + 1/M)\hat{B}$, where $\hat{W}$ is the average of the estimated sampling variances from the M datasets, and $\hat{B}$ is the sample variance of the estimates across the M datasets; the factor $1 + 1/M$ is a small-M correction. The quantity $(1 + 1/M)\hat{B}$ is crucial, because it estimates the increase in the variance from imputation uncertainty, which is omitted (i.e., set to zero) by single imputation methods. Other combining rules provide refinements of this basic method, and include combining rules for test statistics and p-values. See, for example, Little and Rubin (2019, Section 10.2).

The imputation of draws from the predictive distribution creates the variability in the estimates over the MI data sets, allowing the appropriate assessment of imputation uncertainty. Imputing draws is inefficient, but the fact that $\hat{\theta}_{\mathrm{MI}}$ is averaged over datasets reduces this inefficiency, roughly by a factor of M. In fact, MI under a well-specified model is essentially fully efficient from a statistical perspective, providing M is sufficiently large. The appropriate choice of M depends on the fraction of missing information, which is estimated for each parameter θ by $(1 + 1/M)\hat{B}/\hat{V}$. Larger fractions of missing information require larger values of M to yield good estimates of the imputation uncertainty.

Once the data are imputed, the remaining steps of MI are not much more difficult than doing a single imputation. The additional computing from repeating an analysis M times is not a major burden and MI combining rules for standard errors are standard in MI programs. Modern MI programs yield imputed data sets that lead to proper inferences, in the sense that they appropriately incorporate uncertainty in the parameter estimates in the imputation models; they also can be applied to a general missing data pattern. The imputation models generally assume the missing data are MAR, although MNAR mechanisms can also be incorporated (e.g. Tompsett et al. 2018, Giusti and Little 2011). Most of the work is in generating good predictive distributions for the missing values.

There are three primary approaches to creating the predictive distributions for multiple imputation of the missing data. (1) *Joint modeling*, where predictive distributions are derived from an explicit parametric joint model $f(Y|\theta)$ for the variables in the data set, indexed by parameters θ. Examples of models include the multivariate normal model for continuous variables, loglinear

models for categorical variables, and the general location model for mixture of continuous and categorical variables (see Little and Rubin 2019, Chapters 11–14). (2) *Sequential regression imputation* (Raghunathan et al. 2001), also called *chained-equation* imputation (White, Royston and Wood 2011, van Buuren and Oodshoorn 2011), or *full conditional specification*, where a model is specified for the conditional distribution $f_j(Y_j|Y_{(j)})$ of each variable Y_j with missing values, given the other variables $Y_{(j)} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$, for $j = 1,\ldots,p$. These methods are iterative, and impute the missing values of each variable as draws from their conditional distribution, given the observed or most recently imputed values of the other variables. (3) *Hot deck* imputation, which matches each incomplete case (which we call the recipient) to a complete case (which we call the donor) based on some closeness metric. The values of missing variables for the recipient are then imputed with the corresponding values of those variables for the donor. A variety of metrics are used, but a common and principled choice is the distance between the predicted means from a regression of the missing variables on the observed variables (predictive mean matching, see Little 1988). Hot deck methods were originally defined for single imputation – for a review of these methods, see Andridge and Little (2010); they can be extended to MI by defining a set of close donors for each recipient, and randomly picking a donor for each MI data set (Little 1988).

Joint modeling is well-motivated theoretically – the underlying theory is Bayesian, which creates imputations that take into account uncertainty in model parameters. The approach is well suited to a monotone pattern, where the joint distribution of Y can be factored as $f(Y_1, \ldots, Y_p) = f_1(Y_1)$ $f_2(Y_2|Y_1)\ldots f(Y_p|Y_1, \ldots, Y_{p-1})$, with variables arranged from most observed (Y_1) to least observed (Y_p). The distributions in this product can then be modeled using regressions appropriate for the outcome variable type – for example, normal linear regression for continuous outcomes, logistic regression for binary outcomes, and so on. These regressions are quite flexible, in that they can include polynomial terms and interactions as covariates. Imputations are created sequentially, first filling in missing values of Y_1 as draws from $f_1(Y_1)$, then filling in missing values of Y_2 as draws from $f_2(Y_2|Y_1)$, conditioning on observed and previously imputed values of Y_1, and so on. The data analyst creating the MIs needs to provide appropriate specifications of these regressions, using subject-matter knowledge and regression diagnostic tools applied to the set of cases that are observed on the relevant set of variables.

For non-monotone patterns, imputation algorithms are iterative and involve an application of Markov Chain Monte Carlo methods. This means

that methods are needed to monitor convergence of the chain, and the methods can be computationally intensive if the data matrix is large. A challenge for the joint modeling approach is the limited availability of models for joint distribution of Y. For example, the popular multivariate normal model implies that the normal regression models for the imputations that are linear and additive in the covariates, with a constant residual variance. This limitation can be eased by strategies such as transformation of the variables, or more generally by using a latent normal model for binary and unordered categorical data, a flexible approach which has recently been shown to perform well (Quartagno and Carpenter 2019).

The chained equation approach sidesteps this limitation of joint modeling for non-monotone patterns by not requiring that the set of conditional distributions $\{f_j(Y_j|Y_{(j)})\}$ for each j corresponds to a coherent joint distribution for $(Y_1, \ldots, Y_p)$. This allows much more flexibility in the choice of imputation model for each variable, at the expense of some theoretical coherence. In practice, simulations studies suggest that the approach does well, provide careful attention is given to specifying the set of imputation models so they are mutually consistent.

Finally, hot deck approaches avoid the need to formally specify imputation models, and are potentially less vulnerable to model misspecification (although they still rely on the MAR assumption). These methods tend to perform well with large data sets, where potential donors that are close matches to recipients are plentiful. They are less useful (and results may have relatively high variance) in smaller datasets, where good matches are less plentiful. In such setting, the joint modeling or sequential regression approaches tend to be superior.

Methods Compared on Some Common Applications

Inference for the Mean of an Incomplete Variable Y

For inference about a mean (or other location parameter like the median), CC is vulnerable to bias unless the complete cases can be viewed as akin to a random sample of the original data, as when the missing data are MCAR. The bias of CC depends on the fraction of incomplete cases and the extent to which complete and incomplete cases differ on the variable of interest. Specifically, suppose a variable Y has missing values, and partition the population into strata consisting of respondents and nonrespondents to Y. Let μ_{CC} and μ_{IC} denote the population means of Y in these strata, that is of the complete and incomplete cases, respectively. The overall mean is

$\mu = \pi_{CC}\mu_{CC} + (1 - \pi_{CC})\mu_{IC}$, where π_{CC} is the expected fraction of complete cases. Assuming a form of CC that yields an unbiased estimate of μ_{CC}, the bias of CC is:

$$\mu_{CC} - \mu = (1 - \pi_{CC})(\mu_{CC} - \mu_{IC}),$$

the expected fraction of incomplete cases multiplied by the difference in the means for complete and incomplete cases. If the mechanism is MCAR then $\mu_{CC} = \mu_{IC}$ and the bias is zero. For a given value of $(\mu_{CC} - \mu_{IC})$, the bias clearly increases with the expected fraction $(1 - \pi_{CC})$ of the incomplete cases; this is one reason why the fraction of incomplete cases is considered a useful indicator for the potential seriousness of the problem of missing data. However, the bias also depends on $(\mu_{CC} - \mu_{IC})$, a quantity that we typically know little about.

Suppose now we have a set of fully-observed auxiliary variables $X = (X_1, \ldots, X_p)$ as well as the data on Y. The data are then $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, r$ and $(x_{i1}, \ldots, x_{ip})$ for $i = r+1, \ldots, n$ where here r is the number of complete cases. CC inference for the mean of Y discards the units with $X_1, \ldots, X_p$ observed and Y missing. If the data are not MCAR, the distribution of $X_1, \ldots, X_p$ differs for the complete and incomplete cases, and comparisons of these distributions, such as t tests comparing the means, provide tests of MCAR.

Both IPW and MI exploit the information about $X_1, \ldots, X_p$ to potentially reduce the bias of CC. Specifically, in IPW the complete units are weighted by the inverse of the estimated probability that Y is observed, computed based on the response rate within adjustment cells or a regression of R on $X_1, \ldots, X_p$. In MI, the missing values of Y are multiply-imputed as draws from the predictive distribution of Y given $X_1, \ldots, X_p$.

Table 1, an extension of a table in Little and Vartivarian (2005), compares the bias and variance of estimates of the mean of Y from MI and IPW relative to CC. It summarizes theoretical properties of the methods; for example, for an X variable to reduce bias it needs to be related to both Y and R, and to reduce variance it needs to be related to Y. Eight cells are displayed, based on strength of association between the auxiliary variables X and nonresponse R and outcome Y. For characterizing the association between X and Y, it is helpful to split X into the propensity, that is the best predictor of R in the regression of R on X, and components of X orthogonal to the propensity, say Z. Two types of association between X and Y are distinguished, the strength of association between the propensity to respond and Y, and the strength of association between Z and Y. With a single X, the propensity is a function of X and Z is a null set. Within each

cell, the absolute bias and variance of IPW and MI estimates relative to CC are tabulated. See the footnote to the table for details.

When X is weakly associated with both R and Y (Cell LLL), CC, IPW, and MI are similar, and CC may be preferred on grounds of simplicity. For either IPW or MI to reduce absolute bias of CC, the propensity to respond needs to be related to both R and Y, as in the cells HHL and HHH in Table 1. When the propensity is strongly associated with R but weakly associated with Y (Cells HLL and HLH), IPW actually makes things worse than CC in terms of variance, because the variability of the sample weights increases the sampling variance of the weighted mean without a compensating reduction in absolute bias. When the propensity is strongly related to Y (cells LHL, HHL, LHH and HHH), IPW can have lower variance than CC. MI does not have the increased variance of IPW in Cell HLL, and otherwise is more efficient than CC and IPW when there are auxiliary variables other than the propensity that

Table 1. Bias and Variance of MI and IPW Relative to CC for Estimating a Mean, by Strength of Association of the Auxiliary Variables $X = (X_1, \ldots, X_p)$ with Response (***R***) and Outcome (***Y***).

Association of X with Response R (ie. strength of propensity to respond)	Association of outcome Y with (i) propensity to respond and (ii) Z (as defined in the text)			
	Propensity: Low Z: Low	Propensity: Low Z: High	Propensity: High Z: Low	Propensity: High Z: High
Low	Cell LLL IPW MI	Cell LLH IPW MI	Cell LHL IPW MI	Cell LHH IPW MI
	Bias: --- ---	Bias --- ---	Bias: --- ---	Bias: --- ---
	Var: --- ---	Var --- ↓	Var: ↓ ↓	Var: ↓ ↓↓
High	Cell HLL IPW MI	Cell HLH IPW MI	Cell HHL IPW MI	Cell HHH IPW MI
	Bias: --- ---	Bias: --- ---	Bias: ↓ ↓	Bias: ↓ ↓
	Var: ↑ ---	Var: ↑ ↓	Var: ↓ ↓	Var: ↓ ↓↓

Notes:.
"---" for Bias (or Var) within a cell indicates that the estimate for the method has similar bias (or variance) to the estimate for CC.
" ↓ " for Bias (or Var) within a cell indicates that the estimate for the method has less absolute bias (or variance) than the estimate for CC.
" ↑ " for Bias (or Var) within a cell indicates that the estimate for the method has greater absolute bias (or variance) than the estimate for CC.
In summary, " ↓ " indicates that a method is better than CC, " ↑ " indicates that a method is worse than CC, and "---" indicates that a method is similar to CC.

predict Y, namely cells LLH, HLH, LHH and HHH). See Belin (2009) for an application where MI is more efficient than CC, and Collins, Schafer and Kam (2001) for more on the utility of auxiliary variables for enhancing the precision of MI inferences.

Note that MI is seen to be superior to IPW in Table 1, a result of the fact that both methods can reduce bias, but MI can reduce both bias and variance (Little 1986). However, this property relies on the assumption that the imputation model is well specified, for example, nonlinear terms and interactions among the X's are included as predictors if they are needed. If the imputation model is misspecified but the model for R on X for IPW is well specified, then IPW may be superior to MI. Also, IPW based on a single regression of R on X can be applied to a set of variables Y with the same pattern of missing values, whereas MI requires a different imputation model for each Y - variable in the set.

In summary, absolute bias is reduced by IPW and MI when there are auxiliary variables that are predictive of both response and Y. Sampling variance is reduced by IPW when the response propensity is predictive of Y, and MI is generally more efficient than IPW, particularly when auxiliary variables, Z, orthogonal to the propensity to respond are predictive of Y. These comments generally apply for subgroup means, with X interpreted as the set of auxiliary variables other than the variable used to form the subgroups.

What if $X_1, \ldots, X_p$ also have missing values? Because MI can be applied to a general pattern of missing values, it can still be used to recover the information about the mean of Y in the observed auxiliary variables for units where Y is missing, although additional modeling is needed to develop imputation models for the missing values of $X_1, \ldots, X_p$. The weights in IPW can only condition on the subset of $X_1, \ldots, X_p$ that are completely observed, and therefore may be less effective in reducing absolute bias or variance than MI. This is one reason why imputation is generally favored over weighting for item nonresponse, which often gives an unstructured "swiss-cheese" appearance to the matrix R of response indicators.

Missing Data in Regression

The available information in the incomplete cases is different if interest concerns the coefficients of the regression of Y on $X = (X_1, \ldots, X_p)$ rather than

the mean of Y. With X fully observed and missing values confined to Y, and assuming MAR, the likelihood has the form

$$L(\theta, \phi|\text{data}) = \prod_{i=1}^{r} f_{Y|X}(y_i|x_{i1}, \ldots, x_{ip}, \theta) \times \prod_{i=1}^{n} f_X(x_{i1}, \ldots, x_{ip}, \phi)$$

where $f_{Y|X}(y_i|x_{i1}, \ldots, x_{ip}, \theta)$ is the density of the conditional distribution of Y given X and $f_X(x_{i1}, \ldots, x_{ip}, \phi)$ is the density of the marginal distribution of X. It follows that provided θ and ϕ are distinct parameters, the complete cases carry all the information for the parameters θ of the regression of Y on X, so CC is in fact fully efficient. CC is also unbiased under MAR, not requiring the stronger MCAR assumption. Under MAR, MI is not necessary in this situation. IPW is sometimes advocated over CC because it is potentially more robust than CC when the regression of Y on X is misspecified but the model for the propensity to respond is correctly specified. From a robustness perspective, comparing the results from CC and IPW is sensible, if only as a specification check for the regression of Y on X.

The incomplete cases have more information when there are missing values in the covariates X rather than missing values in the outcome Y. Suppose, for simplicity, that values of one of the covariates, say X_1, are missing, and $X_2, \ldots, X_p$, Y are fully observed. The incomplete cases with X_1 missing then have considerable information for the intercept and coefficients of $X_2, \ldots, X_p$, but very limited information for the coefficient of X_1 (Little 1992). The incomplete cases are thus of limited value if the primary interest is in the coefficient of X_1, but are of considerably more value if the primary interest is in other coefficients; in particular, if X_1 is weakly associated with Y, the incomplete cases have about as much information as the complete cases for these other regression coefficients. For MI of covariates, it is important to include the outcome variable Y in the imputation model so as not to bias the estimated regression coefficients from the fitted data (Little 1992).

CC also has the (perhaps unexpected) property of being unbiased for regression coefficients when the probability that a case is complete depends on the covariates but — given these — not the outcome, under a well-specified model (Hughes et al. 2019, Little and Rubin 2019, Example 3.3). For the simple missingness pattern discussed above, we consider three cases:

1. If missingness of X_1 depends on $(X_2, \ldots, X_p)$ but not Y or X_1, then data are MAR, and CC, IPW and MI are all consistent;

2. If missingness of X_1 depends on $(X_2, \ldots, X_p)$ and Y but not on X_1, then data are again MAR; IPW and MI are consistent but CC is generally biased;
3. If missingness of X_1 depends on $(X_1, X_2, \ldots, X_p)$ but not Y, then data are MNAR; MI (based on an MAR model) and IPW are generally biased, but CC is consistent.

Thus CC is biased when missingness depends on Y (given $(X_1, X_2, \ldots, X_p)$), whereas IPW and MI methods are biased when missingness depends on X_1 (given $(X_2, \ldots, X_p)$ and Y). MI is, however, more efficient than CC and IPW when it comes to the standard error of the regression coefficients, so it may be preferred from a mean square error perspective even if a moderate MNAR mechanism is suspected.

If that p is large, and missing data are scattered over the covariates $X_1, \ldots, X_p$ in a haphazard way, such that the fraction of complete cases is relatively small. Then there could be substantial payoff in terms of increased precision in using MI to include the incomplete cases in the analysis. As before, IPW has more limited potential because weighting the complete cases does not exploit the information available in incomplete data.

A hybrid of CC and MI called *subset MI* (SMI, Little and Zhang 2011) has potential value in situations where something is known about the missingness mechanism. Partition the covariates into two sets, $X = (U, V)$, and suppose it is suspected that the probability that U is complete depends on the covariates U and V but not Y, and missing values of Y and V are MAR in the set of units where U is fully observed. Then cases that have missing values for any of the variables in U are discarded, and in the remaining cases, MI is applied to fill in any missing values in V. The resulting data set contains complete cases in U and some cases with imputed values in V. Little and Zhang (2011) show that this method is unbiased, whereas as CC, IPW and MI (applied to the full dataset) are all subject to bias. Thus, for example, if *income* is an incomplete covariate, and missingness of *income* is suspected to depend on the underlying *income* value, then one might discard units with *income* missing and apply MI to the resulting dataset. Another application of this idea arises when the outcome variable Y has missing values: apply MI to the whole data set, but then drop units with Y missing when estimating the regression of Y on X, because after MI these units carry no additional information for the regression (Von Hippel 2007). Here, dropping the incomplete cases avoids simulation error from multiply-imputing values of Y.

Bias and Precision of Complete-Case Inferences for an Odds Ratio

Suppose the data consist of two binary variables Y_1 and Y_2, and the complete cases have both Y_1 and Y_2 observed, and the incomplete cases have either Y_1 or Y_2 missing, The parameter of interest is the odds ratio in the 2×2 table of counts classified by Y_1 and Y_2, The CC estimate is then the sample odds ratio based on the complete cases. This analysis is not subject to selection bias if the probability of response depends on Y_1 alone, or Y_2 alone, or more generally if the logarithm of the probability of response is an additive function of Y_1 and Y_2. This result underpins the validity of case-control studies for estimating odds ratios from observational studies. In terms of precision, supplemental margins on Y_1 and Y_2 provide very little information for the odds ratio, but can reduce bias and increase precision for estimating the marginal distributions of these variables, which may be of substantive interest (Little and Rubin 2019, Example 3.4) For further discussion see Bartlett, Harel and Carpenter (2015).

Similar comments apply to the coefficient of Y_1 in the logistic regression of Y_2 on Y_1 and X, when the incomplete cases have either Y_1 or Y_2 missing and X are additional fully observed covariates. The exponent of the coefficient of Y_1 in that case represents a conditional odds ratio, given X.

Missing Data in Repeated Measures

Longitudinal data are often subject to missing data because individuals drop out before the study ends. This form of missingness is often not MCAR, because the distribution of the study outcomes is different for those who do and do not drop out. A common analysis approach is CC, an approach whose validity again depends on the analysis of interest, and how much information the incomplete cases carry for that analysis.

For example, consider a simple design with fully observed baseline measure on a study variable Y_0 and a single follow-up measure of the study variable, say Y_1, which is missing for individuals who drop out. How much information is available in the dropouts with Y_0 observed but Y_1 missing? If Y_0 is highly predictive of Y_1 but weakly predictive of the change variable $Y_1 - Y_0$, the incomplete cases are informative for the mean of Y_1 but relatively uninformative for the mean of $Y_1 - Y_0$, and (assuming MAR), as we have already seen, have no information at all for the regression of Y_1 on Y_0. So CC is justified for the latter two analyses but less justified for first analysis.

As a more complex example, suppose the study involves fully-observed baseline measures X and K repeated measures $Y = (Y_1, Y_2, \ldots, Y_K)$ on a

study variable, and individuals dropping out between times k and $k+1$ have $Y_1, \ldots, Y_k$ observed and $Y_{k+1}, \ldots, Y_K$ missing. The incomplete cases often have substantial information for the regression of Y_K on X, and a repeated-measures model should be used to model the longitudinal distribution of Y given X. This is particularly so if the intermediate values of Y measured prior to drop-out are predictive of the missing values of Y after dropout. Specifically, use a repeated measures model (fully efficient if correctly specified) for all the observed data, with carefully chosen covariance structure and a parameterization of the mean model chosen to answer the scientific question; for examples, see Carpenter & Kenward (2008) chapter 3. Because ML for the repeated-measures model is fully efficient, MI or IPW is not needed.

Other Analyses

We have focused attention on analysis of means and regression parameters here, because these analyses are common in social science data sets. CC, IPW and MI can all be applied to other types of analysis, such as loglinear models for contingency tables, time series modeling, or analyses that involve latent variables like factor analysis or latent structure analysis. To keep this paper a manageable length, we preclude a detailed discussion of these other kinds of analyses, but offer some general comments for completeness.

1. CC analysis simply applies the analysis of interest to the complete cases, and is often a default option in computer packages, provided missing data codes created to represent missing values are recognized by the package – beware of having missing-data codes like −9999 treated as if they are real values!
2. IPW can be applied in any computer software provided the software allows for weighting the cases.
3. MI can be applied to fill in the missing values, and then the analysis method of interest applied to each of the filled-in data sets. Parameter estimates from the analysis of each data set, and associated standard errors, can be combined using standard MI combining rules. This approach should work well provided the predictive distributions for filling in each of the incomplete variables are reasonable – a MI approach like chained equations (van Buuren et al. (2011); Raghunathan et al. (2001)) is recommended, as these methods handle a general pattern of missing data, and allow for flexible

modeling of the conditional distribution of each of the incomplete variables given the other variables in the data set.

4. An alternative is to apply ML for the analysis of interest. Examples are provided in Little and Rubin (2019). This may be more efficient than chained equation MI because it reflects features of the analysis of interest in the joint distribution of the variables. On the other hand, as noted above, chained equation MI is more flexible, and can included auxiliary variables not included in the final analysis.
5. As discussed in above when comparing the methods on some common applications, the relative gain of MI over CC or IPW depends on how much information is contained for the analysis of interest in the incomplete cases, and this can vary quite dramatically depending on the context. For model-based analysis, information is measured by the observed or expected information matrix, which is the second derivative of the loglikelihood with respect to the parameters. The relative size of this measure of information for a particular parameter in complete and incomplete cases then determines the loss of efficiency of CC, and can be calculated for particular models and parameters, under assumptions about the missing data mechanism. For more details, See Little and Rubin (2019, Section 8.4.3).

Illustrative Example

Introduction

We illustrate some of the points from previous sections with an analysis of data from the Youth Cohort (Time) Series (YCS) for England, Wales and Scotland, 1984 – 2002 (Shapira, Iannelli and Croxford 2007). The raw data are freely available from the U.K. data archive, http://data-archive.ac.uk, study number SN 5765. The data come from two UK representative government-funded cohort studies set up to examine the effects of social, economic and policy change on young people's experiences of education and transitions to the labor market. For our analyses we use a subset of the data from school children attending all school types in England and Wales from five YCS cohorts, who reached the end of Year 11 (ie age 16 + years) in years 1990, 1993, 1995, 1997 and 1999). All our analyses use Stata 15.1.

We compare estimates from CC, IPW and MI of the distribution of parental occupation, and the regression of Year 11 educational achievement (in the General Certificate of Secondary Education qualifications), on the covariates cohort, boy, ethnicity and a three-level classification of parental occupation

derived from information provided by the school children. A description of these variables is given in Table 2.

The dataset contains information on 76,891 children. Data on the covariates boy and cohort are complete; however, the other variables have some missing values, as shown in Table 3. We see that the principal missing data pattern is missing parental occupation (which was derived from a series of questions asked to the pupils).

Estimation of Weights for IPW Analyses

The weights for IPW analyses are the inverse of estimates of the probability of being complete, computed by a logistic regression with outcome 1 if a unit is complete and 0 otherwise. If we include all the 76, 891 units in this regression, we are restricted to variables that are fully observed, namely boy and cohort. To include other predictors in this regression that are relatively complete, we fit the logistic regression using the 74,488 units that have boy, cohort, occupation, ethnicity and GCSE score observed. Because of the size of the data set, all the coefficients in this model are highly significant and are kept in the final model. Table 4 shows the resulting distribution of inverse probability weights.

Creation of Multiply-Imputed Data Sets for MI Analyses

MI is more computationally demanding, but still quite straightforward. We apply MI by chained equations, using the software program Stata. As there

Table 2. Description of Youth Cohort Series Variables.

Variable name	Description
Educational Achievement Score	GCSE points score. Each pupil sits up to 15 GCSE exams. The results for each are converted into a score from 7 (highest grade) to 0 (fail). These are summed across a pupil's exams and capped at 84 (equivalent to 12 GCSEs at the top grade).
Cohort	year of data collection: 1990, 93, 95, 97, 99
Boy	indicator variable for boys
Occupation	parental occupation, categorized as managerial, intermediate or working
Ethnicity	Categorized as Bangladeshi, Black, Indian, other Asian, Other, Pakistani or White

Table 3. Principal Missing Data Patterns in the YCS.

Pattern	GCSE score	Occupation	Ethnicity	n	% of total
1	√	√	√	66965	87%
2	√	?	√	7523	10%
3	?	√	√	760	<1%
4	√	?	?	651	<1%
5		Other patterns		892	<1%

Table 4. Distribution of Inverse Probability Weights from Logistic Model for the Probability That a Unit is Complete.

Percentile of weight distribution								
2.5	25	50	75	97.5	99	99.5	99.9	100
1.02	1.05	1.07	1.13	1.39	1.72	2.06	3.10	6.19

are three variables with missing data, we have three chained regression imputation equations which the imputation algorithm cycles through:

linear regression of GCSE score on: ethnicity, parental occupation, sex, cohort
multinomial regression of ethnicity on: GCSE score, parental occupation, sex, cohort
multinomial regression of parental occupation on: GCSE score, ethnicity, sex, cohort

Each of these properly imputes the missing data in the dependent variable, and then takes these imputed values through to the next model. We complete 10 cycles before imputing each data set and a further 10 cycles between each of our 20 imputations.

A complication in this imputation is that the ethnicity variable has a number of relatively sparse categories, leading to quasi-complete separation. Unless this is corrected for, this can cause coefficients in the multinomial regression model to become large in magnitude, and corresponding SEs to be large too, leading to poor imputations. A relatively simple fix, which we use here, is to temporarily augment the data set with small number of observations at each point when this occurs (White, Daniel and Royston 2010).

We now illustrate and compare the results of two analyses using CC, IPW and MI.

Estimated Marginal Distribution of Ethnicity

Because Ethnicity is the variable with the highest proportion of missing values, we compare estimates of the distribution of this variable from the different methods in Table 5. This analysis is similar to an analysis of means as discussed in above, because the proportion of cases in a particular category can be viewed as the mean of a binary variable indicating belonging to that category.

Because of the relatively small proportion of values with parental occupation missing, overall the marginal distribution of the variable is similar for CC, IPW and ML analysis, both in terms of point estimates and standard errors. However, the proportion of missing parental occupation values is much higher in some ethnic groups, and highest in the Bangladeshi group. For the Bangladeshi group, in particular, point estimates for CC appear biased relative to those which assume MAR (IPW and MI) and we see that MI has notably narrower confidence intervals. As preliminary analysis suggests we are in cells HHL or HHH of Table 1, this is in line with what we expect.

Estimated Regression of GCSE Score on Covariates

The regression analysis results CC, IPW and MI are summarized in Table 6. For simplicity, we treat the weights as fixed rather than accounting for their sampling error, a strategy that tends to yield conservative standard errors (which is of no concern in a data set of this size). We focus here on the coefficients of ethnicity, because the coefficients of the other variables are similar for the three methods.

The estimates of the ethnicity coefficients from CC differ markedly from the estimates from IPW and MI (which are similar). In particular, we see that the estimated coefficient for Bangladeshi ethnicity, which is not statistically significant in the CC analysis, is now statistically significant and similar to the coefficient for the Pakistani group; both coefficients are substantially more negative. Also, the coefficient for Black is slightly more negative, and that for Other Asian more positive. The key reason why the IPW/MI results are relatively unbiased compared to CC is that the probability of a CC (equivalent, for most individuals, to the probability that *occupation* is observed) is strongly influenced by the outcome (GCSE score) and *ethnicity*. In this case, theory (see above) suggests that ethnicity coefficients are likely to be biased. For CC to be less biased than IPW/MI, a

Table 5. Distribution of Parental Occupation, Estimated from CC, IPW and MI for (i) Whole Data Set (top) and (ii) Bangladeshi Ethnic Group (Bottom).

	Parental Occupation (estimate, 95% CI)		
Estimated using:	managerial and professional	intermediate	working
CC (n = 66,965)	43.3% (42.9%–43.6%)	33.5% (33.2%–33.9%)	23.2% (22.8%–23.5%)
IPW (n = 66,965)	42.2% (41.9%–42.6%)	33.8% (33.4%–34.1%)	24.0% (23.7%–24.3%)
MI (n = 76,791)	41.8% (41.4%–42.1%)	33.8% (33.5%–34.2%)	24.4% (24.1%–24.7%)
Bangladeshi ethnicity only:			
CC (n = 246)	14.2% (10.4%–19.2%)	41.5% (35.4%–47.8%)	44.3% (38.2%–50.6%)
IPW (n = 246)	12.5% (8.8%–17.4%)	41.0% (34.5%–47.8%)	46.5% (39.8%–53.4%)
MI (n = 593–605)	11.0% (7.7%–13.6%)	40.4% (34.7%–46.1%)	48.9% (43.2%–54.6%)

Note: As Bangladeshi ethnicity also has missing values (which are imputed) the number of cases of Bangladeshi ethnicity varies across the 20 imputations from 593–605 cases.

Table 6. Estimated Effects of Ethnicity on GCSE Score (Estimate, SE), Adjusted for Cohort, sex and Parental Occupation; (i) Left, CC Analysis (ii) Centre, IPW; (iii) Right, MI.

Ethnic group (reference group: white)	CC analysis	IPW analysis	MI analysis (20 imputations)
Black (1.8%)	−5.77 (0.540)	−7.62 (0.593)	−7.57 (0.488)
Indian (2.9%)	4.27 (0.392)	3.71 (0.412)	3.79 (0.380)
Pakistani (2.0%)	−2.10 (0.557)	−5.25 (0.657)	−4.05 (0.452)
Bangladeshi (0.9%)	0.61 (1.007)	−4.37 (1.262)	−3.91 (0.710)
Other Asian (1.3%)	6.20 (0.586)	5.07 (0.653)	5.32 (0.548)
Other (1.2%)	−0.20 (0.630)	−1.72 (0.748)	−1.26 (0.590)

strong MNAR mechanism would need to be operating. However, Carpenter and Kenward (2013:240) analyzed a different version of these data and showed that inferences are robust to departures from MAR.

While coefficient estimates from IPW and MI are similar, Table 6 shows standard errors are considerably reduced with MI (especially in the Bangladeshi group) compared to the CC analysis. This is typical, and consistent with the discussion above: here MI is mostly bringing back into the analysis individuals with missing *occupation*, but observed data on other variables. Therefore, we would expect coefficients for ethnicity categories, to have greatest reduction in their standard error. This is further emphasized here for the Bangladeshi category, because this group is one of the smallest, yet has the one of the highest proportions of individuals with missing *occupation.*

By contrast SEs for IPW are larger than for the CC analysis. This is again typical; an intuitive explanation is that IPW only reweights cases with no missing data. Cases with one or more missing values are therefore discarded by IPW, whereas all the information is included in MI. Indeed, provide the imputation model is appropriately specified, theory and experience suggest it makes best use of the available data.

Finally, note that both IPW and MI assume that the data are MAR. As discussed earlier, this is typically the natural assumption for a primary analysis. However, as it is untestable from the data at hand, it is often useful to perform sensitivity analysis. This can also be readily carried out using MI; see Carpenter and Kenward (2013:240) for an analysis of different version

of these data, which shows that inferences are robust to the MAR assumption.

Conclusions

We have presented a non-technical discussion of three widely used approaches for handling missing data, namely CC, IPW and MI. In applications, we always begin by tabulating and graphing the data, and exploring the associations using complete case analyses. As we move to the definitive analysis, Table 7 summarizes how we choose between the approaches in our work.

In particular, when data are plausibly MAR, MI and IPW can improve efficiency and reduce bias over a CC analysis. The relative gain of MI over CC or IPW depends on how much information is contained in the incomplete cases for the scientific analysis. Further, IPW and MI can both exploit information in auxiliary variables (which are not included in the scientific model) to (i) increase the plausibility of the MAR assumption, and hence reduce bias and (ii) further improve efficiency – especially with MI, when the auxiliary variables are good predictors of the variables with missing values in the scientific model. A further advantage of MI is that it can be used when these auxiliary variables themselves have missing values.

However, the advantages of MI are contingent on the imputation model being well specified, in terms of assumed relationships between the missing and observed variables. In scientific models where there are, for example, non-linear effects, interactions, hierarchical (multilevel) structure and time-to-event outcomes, considerable thought needs to go into the specification of the imputation model. Analysts should also check that the distribution of imputed data is plausible in the scientific context (e.g. graphically). Carpenter and Smuk (2021) discuss a number of examples in detail, and provide further references. One robust MI approach is Penalized Spline of Propensity Prediction, which imputes missing variables based on a model that includes a penalized spline of the estimated response propensity and other predictive covariates (Zhang and Little 2009, 2011).

When MI is not indicated, and we are choosing between CC and IPW, we reiterate the following. First, for inferences about means, use IPW when auxiliary variables are available that are strongly related to both response and the variable with missing values. Second, for inferences about regression with all (or most) missing values in the outcome alone, CC is valid if the regression model is correctly specified. However, it is prudent to compare results from IPW and CC as a specification check. If the estimated regression coefficients from the two analyses are very different, (and a careful check of the IPW

Table 7. Summary of Recommendations.

Analysis method	When to use	When to avoid
CC	• Unbiased (though not fully efficient) for estimating regression coefficients when the probability that a case is complete depends on the covariates but, given these, not the outcome. • Can be used to estimate an OR if the probability of a complete case depends the exposure or the outcome (but not both)	• when estimating the mean of an incomplete outcome if data are not MCAR (because likely to be biased) • when there are auxiliary variables that can be used to recover missing data (because MI or IPW will be more efficient)
IPW	• More efficient than a CC analysis if there are useful auxiliary variables • Valid for estimating a regression coefficient if the missingness mechanism is MAR	• When MI is also valid, IPW is generally less efficient because IPW (i) only reweights complete cases and (ii) cannot use incomplete auxiliary variables
MI	• More efficient than a CC analysis if there are useful auxiliary variables • Valid for estimating a regression coefficient if the missingness mechanism is MAR • When IPW and MI are valid, and correctly implemented, MI is typically more efficient.	• When we are not confident the imputation model is (i) consistent with the scientific model and (ii) well specified (because results at increased risk of bias)

model does not highlight any concerns) the specification of the regression model needs to be checked for errors (for example, assumptions about linearity of absence of interactions may be invalid.) Third, for inferences about regression with missing values in the covariates, IPW is preferred if the missingness mechanism is MAR, CC is preferred if the missingness mechanism plausibly depends on the covariates but not (or only weakly on) the outcome.

In our work, we typically compare the CC analysis with either MI or IPW (or both) and seek to understand and explain in our reporting why they differ, because such explanations typically give additional insights and hence improve confidence in the scientific findings.

In practice the mechanism behind the missing data will not be known, and requires making an assumption of the most plausible mechanism. It is therefore important to conduct a sensitivity analysis to alternative plausible assumptions regarding the missingness mechanism. One approach to clarifying the assumptions regarding the missingness mechanism in the primary and sensitivity analysis is to use causal diagrams (Lee et al., 2021).

Finally, if the data are suspected to be MNAR, then it is important to consider an MNAR model, at least as a sensitivity analysis. A discussion of MNAR models is beyond the scope of this manuscript, but MI provides a practical vehicle, as described by Carpenter and Smuk (2021) and references therein.

Acknowledgments

James Carpenter is supported by UK Medical Research Council Grant MC_UU_00004/07. The manuscript is submitted on behalf of the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative (http://stratos-initiative.org), which aims to provide accessible and accurate guidance documents for relevant topics in the design and analysis of observational studies. The authors thank the the editor, associate editor, three referees and two reviewers on the STRATOS publication panel for their helpful comments on the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the UK Medical Research Council, grant number MC_UU_00004/07.

ORCID iD

Roderick J. Little https://orcid.org/0000-0001-9878-6977

References

Andridge, Rebecca H. and Roderick J. Little. 2010. "A Review of Hot Deck Imputation for Survey Nonresponse." *International Statistical Review* 78(1):40-64.

Bartlett, Jonathan W., Ofer Harel, and James R. Carpenter. 2015. "Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression." *American Journal of Epidemiology* 182(8):730-6.

Belin, Tom R. 2009. "Missing Data: What a Little can do, and What Researchers can do in Response." *American Journal of Ophthalmology* 148(6):820-2.

Cao, Weihua, Anastasios A. Tsiatis, and Marie Davidian. 2009. "Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data." *Biometrika* 96:723-34.

Carpenter, James R. and Michael G. Kenward. 2008. "Missing Data in Clinical Trials – a Practical Guide." National Health Service Co-ordinating Centre for Research Methodology, url = https://researchonline.lshtm.ac.uk/id/eprint/4018500/.

Collins, Linda M., Joseph L. Schafer, and Chi-Ming Kam. 2001. "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures." *Psychological Methods* 6(4):330-51.

Carpenter, James R. and Melanie Smuk. 2021. "Missing Data: A Statistical Framework for Practice." *Biometrical Journal* 63:915-47. https://doi.org/10.1002/bimj.202000196

Carpenter, James R. and Michael G. Kenward. 2013. *Multiple Imputation and Its Application*. New York: Wiley.

Giusti, Caterina and Roderick J. Little. 2011. "A Sensitivity Analysis of Nonignorable Nonresponse to Income in a Survey with a Rotating Panel Design." *Journal of Official Statistics* 27(2):211-29.

Horowitz, Joel L. and Charles F Manski. 1998. "Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations." *Journal of Econometrics* 84:37-58.

Hughes, Rachael A., Jon Heron, Jonathan A.C. Sterne, and Kate Tilling. 2019. "Accounting for Missing Data in Statistical Analyses: Multiple Imputation is not Always the Answer." *International Journal of Epidemiology* 48(4): 1294-304.

Lee, Katherine J., Kate M. Tilling, Rosie P. Cornish, Roderick J. Little, Melanie M. Bell, Els Goetghebeur, Joseph W. Hogan, and James R. Carpenter. 2021. "Framework for the Treatment and Reporting of Missing Data in Observational Studies: The Treatment And Reporting of Missing Data in Observational Studies Framework." *Journal of Clinical Epidemiology*, 134: 79-88.

Little, Roderick J. 1986. "Survey Nonresponse Adjustments." *International Statistical Review*, 54, 139-157

Little, Roderick J. 1988. "Missing Data in Large Surveys (with Discussion)." *Journal of Business and Economic Statistics* 6:287-301.

Little, Roderick J. 1992. "Regression with Missing X's: A Review." *Journal of the American Statistical Association* 87:1227-37.

Little, Roderick J. 1995. "Modeling the Drop-Out Mechanism in Longitudinal Studies." *Journal of the American Statistical Association* 90:1112-21.

Little, Roderick J. and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*,

Little, Roderick J. and Sonya Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31:161-8.

Little, Roderick J. and Nanhua Zhang. 2011. "Subsample Ignorable Likelihood for Regression Analysis with Missing Data." *Journal of the Royal Statistical Society: Series C: Applied Statistics* 60(4):591-605.

Muthen, Linda K. and Bengt Muthen. 2017. *Mplus Version 8 User's Guide*. Muthen and Muthen.

National Research Council. 2010. *The Prevention and Treatment of Missing Data in Clinical Trials*. United States National Research Council.

Quartagno, Matteo and James R. Carpenter. 2019. "Multiple Imputation for Discrete Data: Evaluation of the Joint Latent Normal Model." *Biometrical Journal* 61(4):1003-19.

Raghunathan, Trivellore E. 2015. *Missing Data Analysis in Practice*. New York: Chapman and Hall / CRC.

Raghunathan, Trivellore E., James Lepkowski, John Van Hoewyk, and Peter W. Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology* 27(1):85-95. For associated IVEWARE software see. http://www.isr.umich.edu/src/smp/ive/

Robins, James M. and Andrea Rotnitzky. 1995. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association* 90:122-9.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1995. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data." *Journal of the American Statistical Association* 90:106-21.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41-55.

Rubin, Donald B. 1976. "Inference and Missing Data (with Discussion)." *Biometrika* 63:581-92.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, Donald B. 1996. "Multiple Imputation After 18 + Years (with Discussion)." *Journal of the American Statistical Association* 91:473-89.

Rubin, Donald B. and Nathaniel Schenker. 1986. "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association* 81:366-74.

SAS. 2015. The MI Procedure. SAS/STAT 14.1 User's Guide, SAS Institute Inc., Cary, NC, USA.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. New York: CRC Press.

Schafer, Joseph L. 1998. "Multiple Imputation: A Primer." *Statistical Methods in Medical Research* 8:3-15.

Seaman, Shaun R. and Ian R. White. 2011. "Review of Inverse Probability Weighting for Dealing with Missing Data." *Statistical Methods in Medical Research* 22:278-95.

Shapira, Marina, Cristina Iannelli, and Linda Croxford, 2007. Youth Cohort Time Series for England, Wales and Scotland, 1984–2002. [Data Collection]. Scottish Centre for Social Research, University of Edinburgh, Centre for Educational Sociology, National Centre for Social Research, [original data producer(s)]. Scottish Centre for Social Research. SN: 5765, https://doi.org/10.5255/UKDA-SN-5765-1

Su, Yu-Sung, Andrew Gelman, Jennifer Hill, and Masanao Yajima. 2011. "Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box." *Journal of Statistical Software* 45(2):1-31.

Tompsett, Daniel M., Finbarr Leacy, Margaritat Moreno-Betancur, Jon Heron, and Ian R. White. 2018. "On the use of the not-at-Random Fully Conditional Specification (NARFCS) Procedure in Practice." *Statistics in Medicine* 37(15):2338-53.

van Buuren, Stef. 2018. *Flexible Imputation of Missing Data.* 2nd Edition. FL: Boca-Raton: CRC/Chapman and Hall.

van Buuren, Stef and Karen Groothuis-Oudshoorn. 2011. "Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45(4):1-67. For associated software see http://www.multiple-imputation.com.

Von Hippel, Paul T. (2007). "Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data." *Sociological Methodology* 37(1): 83-117.

White, Ian R., Rhian Daniel, and Patrick Royston. 2010. "Avoiding Bias Due to Perfect Prediction in Multiple Imputation of Incomplete Categorical Variables." *Computational Statistics and Data Analysis* 54(10):2267-75.

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30(4):377-99.

Zhang, Guangyu and Roderick J. Little. 2009. "Extensions of the Penalized Spline of Propensity Prediction Method of Imputation." *Biometrics* 65(3):911-8.

Zhang, Guangyu and Roderick J. Little. 2011. "A Comparative Study of Doubly-Robust Estimators of the Mean with Missing Data." *Journal of Statistical Computation and Simulation* 81(12):2039-58.

Author Biographies

Roderick J. Little is Richard D. Remington Distinguished University Professor of Biostatistics at the University of Michigan, where he also holds appointments in the Institute for Social Research and the Department of Statistics. His research focuses on methods for the analysis of data with missing values and model-based survey inference, and the application of statistics to diverse scientific areas, including medicine, demography, economics, psychiatry, aging and the environment.

James R. Carpenter is Professor of Medical Statistics at the London School of Hygiene & Tropical Medicine, and Methodology Programme Leader at the MRC Clinical Trials unit at UCL, London UK. His research interests include methodology for multilevel modelling and missing data, with applications to observational data and clinical trials. He co-authored Multiple Imputation and its Application (Wiley, 2013) with Mike Kenward.

Katherine J. Lee is Professor of Biostatistics at the Murdoch Children's Research Institute, Melbourne, and the University of Melbourne, Australia. Her research interests are in the method of multiple imputation for missing data and adaptive clinical trial designs.