Doidge, J. C. (2016). Responsiveness-informed multiple imputation and inverse probability-weighting in cohort studies with missing data that are non-monotone or not missing at random. *Statistical Methods in Medical Research*, Advanced online publication. http://dx.doi.org/10.1177/0962280216628902

Correspondence can be addressed to James Doidge at James.C.Doidge@gmail.com

This manuscript was produced while based at the Centre for Population Health Research, University of South Australia.

Abstract

Population-based cohort studies are invaluable to health research because of the breadth of data collection over time, and the representativeness of their samples. However, they are especially prone to missing data, which can compromise the validity of analyses when data are not missing at random. Having many waves of data collection presents opportunity for participants' responsiveness to be observed over time, which may be informative about missing data mechanisms and thus useful as an auxiliary variable. Modern approaches to handling missing data such as multiple imputation and maximum likelihood can be difficult to implement with the large numbers of auxiliary variables and large amounts of non-monotone missing data that occur in cohort studies. Inverse probability-weighting can be easier to implement but conventional wisdom has stated that it cannot be applied to non-monotone missing data. This paper describes two methods of applying inverse probability-weighting to non-monotone missing data, and explores the potential value of including measures of responsiveness in either inverse probability-weighting or multiple imputation. Simulation studies are used to compare methods and demonstrate that responsiveness in longitudinal studies can be used to mitigate bias induced by missing data, even when data are not missing at random.

1 Introduction

Missing data are one of the few problems faced by researchers from all disciplines. This problem is a particularly strong feature of longitudinal studies involving humans, where the logistics of following participants over years or decades combine with human error and omission to degrade the representativeness of samples and thus the utility of their data. Cohort studies often collect large quantities of information and exhibit complex patterns of missing data. While the strength of cohort studies lies in the breadth of information that can be collected over long periods of time, large numbers of variables push the computational limits of the statistical methods available for analysing missing data. Modern techniques for handling missing data generally assume that data are *missing at random* given observed variables. The challenge for the analyst then is to include a set of observed variables that is sufficient to maximise the plausibility of the *missing at random* assumption. This set can include auxiliary variables, which are informative about missingness but extraneous to the analytic model.

In large cohort studies, even the list of *prima facie* good candidates for inclusion as auxiliary variables may be long. The motivation for this paper was a program of research focusing on child maltreatment in a prospective birth cohort that had been followed through 15 waves of data collection over 28 years. Child maltreatment was recorded through the retrospective self-report of participants in wave 14 (age 23–24 years), by which time a substantial portion had been lost to follow-up and another group or were non-respondent in that wave. There are good reasons to expect that missingness does not occur at random in variables like child maltreatment, which correlates with disadvantage and social marginalisation.[1] Hundreds of potential risk factors and outcomes of child maltreatment were identified in the available dataset, across all waves of data collection. Nearly all were associated with both child maltreatment and missingness and thus good *prima facie* candidates for inclusion in analysis of missing data. Furthermore, child maltreatment itself was associated with missingness in other waves. Thus, questions arose as to if and how responsiveness could be utilised to maximise the plausibility of the missing at random assumption, and how to best utilise the large set of candidate auxiliary variables.

This paper reviews the available methods for dealing with missing data in large cohort studies and presents some adaptations of inverse probability-weighting and multiple imputation that utilise auxiliary variables, with a particular focus on the utilisation of responsiveness. The methods are applicable primarily to longitudinal studies with at least three waves of follow-up. They are particularly relevant to research questions where strong relationships are suspected between model variables and the likelihood of data being missing or being not missing at random.

The paper is organised as follows: Section 2 provides some background on types of missing data, common approaches to addressing missing data, and the potential value of responsiveness as an auxiliary variable; Section 3 describes a simple approach for applying inverse probability-weighting to non-monotone missing data, which makes some implicit use of responsiveness and has been previously implemented but not fully described or tested; Section 4 presents a novel approach to implementing inverse probability-weighting with non-monotone missing data, which also makes implicit use of responsiveness; Section 5 discusses some of the specific limitations of these approaches and extensions for addressing them; Section 6 presents four simulation studies that compare various approaches to inverse probability-weighting and multiple imputation, with and without inclusion of responsiveness and under different missing data conditions; and Section 7 summarises the main conclusions from these.

2 Background

2.1 Patterns and mechanisms of missing data

Patterns of missing data can be broadly classified as either monotone or non-monotone.[2] In cohort studies, progressive loss of participants, such as from death, withdrawal of consent or loss of contact, results in often-large portions of missing data that are strictly increasing over time ('monotone'). However, this situation is usually complicated by non-response to individual items and to whole waves of the survey (and a range of less common sources of missing data), which produce 'non-monotone' patterns of missing data. Monotone patterns are useful because they can simplify some of the methods for addressing missing data.[3, 4]

‘Mechanisms of missingness’ refer to the probability of missingness with respect to variables of interest and may or may not relate to the true causes of missingness.[5] Missing data can be categorised into three conditions relative to the assumption upon which an analysis is based: *missing completely at random (MCAR)*, *missing at random (MAR)*, and *not missing at random (NMAR).*[2] The MCAR assumption holds if missingness is unrelated to the values or missingness of any variables included in the analysis—generally, a strong assumption but the one underlying, for example, complete case analysis/listwise deletion. The weaker MAR assumption adopted by 'modern' methods holds if missingness of model variables is unrelated to the values or missingness of other model variables, *given the observed values of any variables included in the analysis* (including auxiliary variables). If there are relationships between missingness and model variables that exist after conditioning on other observed variables, then the MAR assumption does not hold and data are considered to be NMAR for the purposes of that analysis.[2] The validity of the MAR assumption cannot truly be tested, as it relates to what are essentially 'unknown unknowns'.[6]

2.2 Multiple imputation, maximum likelihood and inverse probability-weighting

When the level of missing data is non-trivial and there are not good reasons to expect the MCAR assumption to hold (e.g. data were accidentally destroyed), there are three broad approaches to handling missing data that may produce valid inference under less restrictive MAR assumptions: multiple imputation, maximum likelihood and inverse probability-weighting. Imputation involves replacing missing values with those drawn from observed conditional distributions given any other available information. The assumptions underlying imputation are weaker than they may seem, operating at the level of the sample distributions rather than individual participants. Further, *multiple* imputation accounts for uncertainty in the imputed values, as first demonstrated by Rubin[7].

While multiple imputation is perhaps the most commonly utilised of the modern approaches to missing data, imputation of large amounts of missing data can result in computational problems that impede the use of large imputation models that may be required to minimise the strength of the *missing at random* assumption.[4] Additionally, large numbers of variables and non-normally distributed variables can be difficult to handle, and multiple imputation is best conducted in the context of pre-specified analyses as it is important that the imputation models support (are consistent with) the analysis model.[4] All of these limitations present problems for cohort studies, where large numbers of variables are collected over long periods of time with associated high levels of missingness, for the purpose of unspecified future analyses. Multiple imputation can, however, handle non-monotone patterns of missing data, with the most common approaches to these being imputation by chained equations and multivariate normal imputation.[4]

Maximum likelihood-based approaches to missing data, including Heckman selection models, involve explicitly modelling the likelihood of having complete data, simultaneously with implementation of the analysis model.[8, 9] While having some advantages over multiple imputation and inverse probability-weighting with respect to assumptions and efficiency, implementation of maximum likelihood is generally restricted to linear models. Also, the ability to draw on information from auxiliary variables is restricted by the simultaneous implementation of missingness and analysis models, and the availability of software is limited.[10]

Inverse probability weighting is related to maximum likelihood-based approaches in the explicit modelling of an auxiliary 'missingness model'. However, rather than estimate sampling probabilities simultaneous and include them explicitly in the analysis model, the missingness model is estimated first, then each complete case is weighted according to the inverse of their probability of being complete, e.g. a participant who has half the probability of providing complete data carries twice as much weight in a substantive analysis that only includes weighted complete cases.[3, 11] Compared with multiple imputation and maximum likelihood, the implementation of inverse probability-weighting can be relatively straightforward; first the likelihood of having complete data given observed data is estimated using logistic regression, then the inverse of the fitted values are used to weight participants in a weighted complete case analysis. While this is a quite different approach to multiple imputation, the fundamental assumption underlying each is MAR.

While inverse probability-weighting is commonly implemented to weight known-probability samples (in national surveys, etc.), its use in other missing data problems has been restricted. This is largely because of the problems that non-monotone missing data patterns present to its implementation.[3] With monotone missing data, exclusion from the analysis because of incomplete data can be stepped out over time; e.g. in the case of progressive loss to follow-up, partial weights can be derived to account for loss to follow-up at each wave, based on all of the information observed prior to that wave, with the final weights being the product of the partial weights for each wave. With non-monotone missing data, the conventional approach is to construct weights using only variables that are fully observed.[3] In the longitudinal setting, this usually restricts the weighting to variables measured at baseline only, which can mean ignoring a lot of information that may be informative about mechanisms of missingness, particularly missingness that occurs much later in time.

2.3 Auxiliary variables and the potential value of responsiveness with respect to the *missing at random* assumption

A key point to note about each of the above approaches to missing data is that the MAR assumption is defined by the observed information that is fed into the analysis. Auxiliary variables can reduce bias by weakening the MAR assumption in both multiple imputation and inverse probability-weighting.[3, 12, 13] Measures of responsiveness are not usually considered as potential auxiliary variables, perhaps because they are 'metadata' that may need to be derived, and perhaps because information about responsiveness over time is limited or absent in most study designs.[14]

Longitudinal studies with many waves of data collection provide a unique opportunity for observation of responsiveness in waves outside of those containing key variables. Responsiveness can be measured as, for example, the proportion of other waves responded to or the proportion of items missed in each wave. Using data from the 1958 National Child Development Study (NCDS), Hawkes and Plewis[15] modelled non-response over time in the as dependent variables, to identify important missing data mechanisms, but did not evaluate their potential uses as auxiliary independent variables. When imputing missing predictors of treatment assignment before applying inverse probability of *treatment* weighting (propensity scores, used to address confounding rather than missing data), inclusion of missing value indicators has been found to reduce bias under NMAR conditions but induce bias under MAR.[16]

Adding measures of responsiveness to missingness or imputation models may weaken the MAR assumption in the same way that adding other auxiliary variables does, with the weakening occurring if they predict both the values and missingness of variables in the analysis model.[13] One potential difference from other auxiliary variables is that responsiveness may be a very good predictor of missingness. If relationships between model variables and missingness varies over time, then auxiliary measures of missingness may be less informative or even misinformative. The same is true, however, for more tangible auxiliary variables and this reflects the irreducible part of the MAR assumption; that missingness cannot be determined by unobserved variables. When many waves of follow-up are available, the stability of an association between a variable and response in other waves could be observed or even modelled.

Apart from weakening the MAR assumption for the given analysis, inclusion of responsiveness provides opportunity for indirect evaluation of the MAR assumption. Strictly speaking, the MAR assumption cannot be *tested* because it is a function of unknown information.[6, 17, 18] However, observed relationships between model variables and responsiveness at one time point may be informative about the relationships to missingness that cannot be observed at other time points.[14, 19] This is the same basic premise that underlies follow-up studies of non-respondents. For example, Fielding et al.[14] adapt the method proposed by Fairclough[20] to test whether their outcome was significantly associated with response to a follow-up survey, after controlling for all other predictors of response to the follow-up survey. With many waves of follow-up, similar tests could be implemented using more sensitive continuous or count measures of missingness. In either inverse probability-weighting or multiple imputation, insight might also be gained by comparing estimates that include or exclude the explicit measures of missingness.

In Sections 3–5, I will now discuss two techniques for applying inverse probability-weighting to cohort data with non-monotone patterns of missing data. The first is a relatively simple approach which can only be applied in relatively simple non-monotone patterns. It has been previously implemented but not discussed in detail or evaluated. The second is a novel, more complex and versatile approach that extends some of the principles of the first. Both techniques make some implicit use responsiveness and allow for further explicit utilisation of it. In Section 6 they are evaluated using simulated data, alongside approaches to multiple imputation that vary in their utilisation of responsiveness as an auxiliary variable. If your interest lies more in multiple imputation (which is more efficient and, in some cases, more effective), then feel free to skip ahead to Section 6.

3 Stratified inverse probability-weighting

3.1 Overview

Implementation of inverse probability-weighting in non-monotone missing data can be potentiated by stratifying on the pattern of missingness.[3, 21, 22] First, a minimum threshold for completeness must be set, such as response to a certain wave of data collection. Participants meeting this threshold will form the final weighted subsample. The threshold may include observation of all variables required for the analysis, or it may include only some (e.g. the outcome), with an expectation that missingness in the remainder can be handled by some other means (i.e. combining with multiple imputation or simpler, less robust methods that me be adequate for small amounts or certain types of missing data). Then, the sample is stratified according to their pattern of missingness in other waves or variables, and the probability of inclusion in the weighted subsample estimated within each stratum given only variables that are complete in that stratum. This approach was implemented by Seaman and White[3], although little of their report was devoted to this aspect of the analysis. Using data from four waves of the NCDS, Seaman and White stratified their weighting by response in the two intermediate waves, resulting in four strata (respondent in both intermediate waves, non-respondent in wave 2 only, non-respondent in wave 3 only and non-respondent in both wave 2 and wave 3). Stratified inverse probability-weighting has a few key benefits:

1. It is compatible with non-monotone missing data, provided that the number of missingness patterns is small. If the number of missingness patterns is large, a few extensions are available (more on these to follow).

2. Compared with the conventional inverse probability-weighting approach which utilises only complete variables, it includes more potentially-informative variables for those in whom additional variables are observed. Thus, the underlying MAR assumption is weakened (in those strata).

3. By allowing the probability of completeness to vary according to the pattern of missingness (through stratification), prior responsiveness becomes an implicitly modelled predictor of completeness.

3.2 Assumptions

The fundamental assumption characterising stratified inverse probability-weighting is that, within each pattern of response, missingness does not depend on variables in the model of interest, given variables that were observed within that stratum.

Formally, if we let

$$I = \text{inclusion in the analysis versus exclusion because of incomplete data}$$

$$X_1, \ldots, X_k = \text{vectors of variables recorded in waves } 1, \ldots, k$$

$$M_1, \ldots, M_k = \text{missingness pattern in } X_1, \ldots, X_k$$

$$X_M = \text{variables complete within missingness pattern } M$$

then the MAR assumption is that, for each stratum,

$$P(I|M_k, X_1, \dots X_k) = P(I|M_k, X_M) \tag{1}$$

For example, in the analysis implemented by Seaman and White[3], let the four patterns of response in waves two and three be defined as M_{11}, M_{10}, M_{01}, and M_{00}, respectively. Inclusion, I, was defined by response to the fourth wave. The set of assumptions relating to missingness mechanisms were the following:

$$\begin{cases} P(I|M_{11}, X_1, \dots X_4) = P(I|M_{11}, X_1, X_2, X_3) \\ P(I|M_{10}, X_1, \dots X_4) = P(I|M_{10}, X_1, X_2) \\ P(I|M_{01}, X_1, \dots X_4) = P(I|M_{01}, X_1, X_3) \\ P(I|M_{00}, X_1, \dots X_4) = P(I|M_{00}, X_1) \end{cases} \tag{2}$$

4 Stepped inverse probability-weighting

4.1 Overview

One limitations of stratified inverse probability-weighting is that as the number of waves increases, the number of potential patterns of missingness increases exponentially (other limitations will be discussed in Section 5). In such cases, some compromises may be possible through collapsing strata with similar response patterns, and either imputing missing values or replacing the variables from conflicting waves with a set of dummy variables indicating response in those waves. However, this will only get you so far before large quantities of information are ignored or the analysis becomes overly complicated. Stepped inverse probability-weighting is a novel and relatively straightforward approach to handling situations like these.

Stepped inverse probability-weighting combines elements of stratified inverse probability-weighting with the standard implementation of inverse probability-weighting for monotone missing data, by identifying a subset of non-monotone missing data that exhibits monotone missingness. It is particularly relevant to large cohort studies with many waves of data collection. Like the standard approach to monotone missing data patterns, inclusion as a (minimally) complete case is stepped out over time. While patterns of wave response may be non-monotone, inclusion is strictly monotone; participants become progressively excluded from the analysis after the final wave in which they contribute a response. For example, in the above illustration from Seaman and White[3] participants in patterns M_{11} and M_{01} who did not respond to wave 4 both would have become excluded after wave 3, regardless of their response in wave 2.

The next step derives from the insight that, using this stepped definition of 'inclusion', participants who become excluded at any given point in time are by definition a subset of respondents to the previous wave. Therefore, partial weights can be estimated using only these individuals, assigning partial weight = 1 for that step to any individuals who did not respond to the previous wave. The implication of this is that now, instead of relying only on baseline information, stepped inverse probability-weighting can be applied using any variables that were observed at baseline *or in the previous wave*. Further, observations of responsiveness in the intermediate waves (between baseline and the previous wave) will also be completely observed and these can be added to the missingness models.

In stepped inverse probability-weighting, the initial stratification on response to the previous wave can then be extended, depending on within-stratum power, by further stratifying on response to intermediate waves. Theoretically, this can progress up to point at which 'full stratification' is achieved—all missingness patterns are represented within each step. At this point, stepped inverse probability-weighting becomes a somewhat less efficient version of stratified inverse probability-weighting (less efficient because of the increased variability in weights that would result from the assignment of partial weights = 1 to the non-respondents to the previous wave at each step). In such cases, it would be preferable to use stratified inverse probability-weighting, without stepping inclusion out over time. In its simplest form—stratified only on response to the previous wave—stepped inverse probability-weighting involves a single regression for each wave of follow-up but has the potential to include information that may be highly informative about missingness mechanisms: variables observed immediately prior to loss to follow-up, and indicators of prior responsiveness itself.

4.2 Assumptions

Stepped inverse probability-weighting is in some respects a collapsed form of stratified inverse probability-weighting; the starting point is a 'fully collapsed' stratum within each step. Thus, the basic assumption underlying stepped inverse probability-weighting is usually stronger than for stratified inverse probability-weighting. In the fully collapsed scenario, the assumption is that exclusion (read: loss to follow-up) at each wave is unrelated to variables in the model of interest, *given variables that are complete amongst respondents to the previous wave*. This usually includes at least baseline variables, variables from the previous wave and intermediate responsiveness.

Formally, in addition to the terms defined above, if we let

$$I_2, \dots, I_k \ = \ \text{Inclusion in the analysis, stepped over waves } 2 \dots k$$

$$R_{k-1} = \text{response to the previous wave}$$

$$X_{R_{k-1}=1} = \text{variables complete amongst respondents to the previous wave}$$

$$M_{R_{k-1}=1} = \text{measures of prior responsiveness in respondents to the previous wave}$$

then, for each wave of follow-up, the basic assumptions is that

$$P(I_k|R_{k-1}, X_1, \dots X_k) = P(I_k|R_{k-1}, X_{R_{k-1}=1}) \qquad (3)$$

which can usually be further specified to

$$P(I_k|R_{k-1}, X_1, \dots X_k) = P(I_k|R_{k-1}, X_1, X_{k-1}, M_{R_{k-1}=1}) \qquad (4)$$

For example, if stepped inverse probability-weighting had been applied to the example provided by Seaman and White[3] in its most simple (least stratified) form, then the set of assumptions would have been

$$\begin{cases} P(I_2|X_1, \dots X_k) = P(I_2|X_1) \\ P(I_3|R_2 = 1, X_1, \dots X_k) = P(I_3|R_2 = 1, X_1, X_2) \\ P(I_4|R_3 = 1, X_1, \dots X_k) = P(I_4|R_3 = 1, X_1, X_3, R_2) \end{cases} \qquad (5)$$

Of course, such an analysis should not have been implemented because the number of strata was low so it was possible to implement stratified inverse probability-weighting.

Note that the first two assumptions in (5) are almost equivalent to the assumptions that would be adopted in a standard application of inverse probability-weighting should the data have been monotone missing, while the final assumption excludes X_2 (which would have been observed if data were monotone missing) and includes a reference to response in the second wave, R_2 (which corresponds to $M_{R_{k-1}=1}$ in (4)). Including such explicit measures of responsiveness is optional.

5 Limitations and extensions of stepped/stratified inverse probability-weighting

5.1 Item non-response in missingness models

The examples and assumptions above consider only wave non-response, ignoring the possibility of item non-response, which usually occurs to some extent and was of course also encountered by Seaman and White[3]. As there are usually far more data items than waves of data collection, item non-response has the potential to increase the number of missingness patterns by orders of magnitude. Apart from further stratification, at least a couple of alternatives are available: impute the missing values prior to weighting, or incorporate missing value indicators into the weighting procedure. While incorporating missing value indicators is not generally a good idea in substantive analysis models, it may be defended in missingness models on the grounds that the missingness itself can be informative too—potentially even more informative than the underlying value. The usual coefficient biases and problems with interpretation that occur when using missing value indicators in analysis models do not apply as clearly to the missingness model, because inferences derived from the missingness model do not need to be generalised beyond the study sample; missingness does not even exist in the population. Seaman and White[3] used single imputation for missing values with a prevalence of <2% and missing indicators for items with >2% missing values.

5.2 Item and wave non-response in analysis models

In multivariate analyses, especially those incorporating many variables, weighting only complete cases may erode too much of the sample, diminishing statistical power and strengthening the MAR assumption (the smaller the fraction of complete cases, the less likely it is that weighting can make them representative of the whole sample). In such cases, it may be preferable to define a threshold of 'minimal completeness' and to combine inverse probability-weighting for handling the bulk of missing data (usually loss to follow-up) with another method for handling the remainder. Building on their previous analysis, Seaman et al.[21] demonstrate how inverse probability-weighting can be combined with multiple imputation; first weighting respondents to wave 4, then multiply imputing any remaining missingness due to item non-response or prior wave non-response. This approach is equally applicable to stepped/stratified inverse probability-weighting. Combining other simpler but less robust methods (single imputation, last observation carried forward, etc.) may also be acceptable depending on the variable concerned and extent of missingness within the minimally-complete cases.

5.3 Many, rare, and strong predictors of missingness

In large cohort studies in particular, each wave of data collection may include hundreds of variables, reaching into the thousands once level-indicators of categorical variable are considered. This can present obvious problems for the fitting of missingness models. Seaman and White[3] propose using forwards stepwise selection in such cases. However, if an analysis model is pre-specified, then one should also consider the relationship between the potential predictors of missingness and the variables of interest (those included in the substantive model); including variables that predict missingness *but not the variables of interest* will only reduce efficiency of the weights without improving bias.[3] For categorical predictors of missingness, it is also worth considering collapsing categories that exhibit similar associations with missingness, thereby improving overall model power and potentially allowing for a wider range of (more coarsely measured) variables to be included. This could be done prior to performing stepwise variable selection, and may reduce the need to select variables.

Another time to consider collapsing levels of categorical missingness predictors is when some of those levels are rare. Rare variables may produce instability in weights or lead to perfect prediction. Usually, this perfect prediction will be of *inclusion*, as exclusion, at least when stepped over time, will usually be less common. This makes collapsing rare categories less problematic. When categories strongly predict *exclusion*, then there is likely to be a mechanism violating the MAR assumption that must be acknowledged and the implications for the analysis considered.

It should be noted that collapsing variables for the purposes of the missingness model does not necessarily inhibit their use in other forms (e.g. uncollapsed) in the analysis model. Provided that the conditional associations with missingness are in fact similar, then the MAR assumption does not greatly change. However, collapsing categories that are *dissimilar* with respect to missingness could substantially alter the MAR assumption and reduce the effectiveness of weighting.

5.4 Power requirements

For stepped/stratified inverse probability-weighting to be implemented, each stratum must contain a sufficient number of observations (participants) proportional to the prevalence of the missingness being modelled, to ensure that regression coefficients, and thus weights, are reasonably stable. When retention is near-perfect within a stratum or at a step, either a smaller set of weighting variables must be selected (which may not matter given the small potential for bias) or the stratum can be combined with another exhibiting a similar pattern of missingness.

5.5 Structural support for inverse probability weights

For weighting to be effective, there must be good overlap of the distributions of weights between the included minimally complete cases and those excluded because of missing data. Some insights about this may be able to be gleaned from the literature on inverse probability of treatment (propensity score) analysis. For example, Rubin[23] proposed three guidelines for comparing the distributions of propensity scores between treatment and control groups: (1) that the difference in mean propensity score should be less than half a standard deviation, (2) that the ratio of the variances of the propensity scores be close to one and certainly between 0.5 and 2.0, and (3) that the ratio of the variances of the residuals of covariates after adjusting for the propensity score must be close to 1.0 and certainly between 0.5 and 2.0. When these guidelines are exceeded, the capacity of the propensity scores to reduce confounding becomes limited, and likewise the capacity of inverse probability-weighting to reduce bias from missing data may also be reduced. Such guidelines would, however, effectively restrict the application of inverse probability-weighting to situations where only low levels of selection bias would be present in complete case data. It may be that the use of inverse probability-weighting in missing data requires different thresholds to the ones proposed by Rubin for use with confounding. In any case but especially when there is poor overlap of the distributions, close attention to the characteristics of participants with very low predicted probabilities of completeness, and to the factors most strongly associated with non-response, may help to understand limits of representation of the weighted subsample.

5.6 Plausibility of the stratified/stepped *missing at random* assumption

Stepped/stratified inverse probability-weighting allows the missingness model to vary by step and/or stratum, only including predictors of missingness in those for whom they are observed. That a variable would be only be related to missingness if it is observed might seem unrealistic but becomes more intuitive when you consider missingness occurring over time. For example, variables observed in wave 2 may well be unrelated to loss to follow-up *at* wave 4 *given* variables observed in wave 3, while still being associated with loss to follow-up at wave 3. In cohort studies that collect a broad amount of information at each wave, and in trials where the information collected during follow-up are repeated measures of the same underlying variable, this assumption may be more defensible. In their multivariate analysis of the NCDS cohort, Hawkes and Plewis[15] found that previous-wave variables were the strongest predictors of non-response.

6 Simulations

6.1 Objectives and methods

Below are presented a series of simulations exploring (1) the value of stepped or stratified inverse probability-weighting in reducing bias compared with other available methods, (2) the potential value of including measures of responsiveness in inverse probability-weighting or multiple imputation, and (3) the sensitivity of each method to violations of the missing at random assumption. Eleven approaches were compared: (i) complete case analysis (to benchmark the degree of bias reduction in other methods), (ii) inverse probability-weighting using baseline variables only, (iii) inverse probability-weighting using baseline measures and responsiveness, (iv) stratified inverse probability-weighting, (v) stepped inverse probability-weighting without responsiveness, (vi) stepped inverse probability-weighting with responsiveness, (vii) multiple imputation using baseline variables only, (viii) multiple imputation using baseline variables and responsiveness, (ix) multiple imputation using responsiveness only (for comparison purposes only; this would not usually be recommended); (x) multiple imputation by chained equations without responsiveness, and (xi) multiple imputation by chained equations with responsiveness. Maximum likelihood-based approaches were not used because of the focus on categorical and auxiliary variables. Another approach not considered here is multivariate normal multiple imputation, as all variables were categorical. However, methods for its application with categorical variables have been a topic of recent research[4, 24, 25] and it may warrant consideration in similar situations.

The simulations focus on wave non-response in a longitudinal setting with non-monotone patterns of missingness. The simulations were designed to mimic the motivating example: estimation of the prevalence of child maltreatment, retrospectively recorded in a population-based birth cohort and associated with non-response at all points in time. For simplicity and comparability, item non-response is not simulated, but could be handled by some of the approaches in ways described in section 5.

Each simulation included one outcome variable, to be estimated from observation in the final wave, and 20 weak correlates (c = 0.1 with outcome, otherwise independent of each other) observed in each other wave (i.e. distinct rather than time-varying correlates). In keeping with the categorical nature of most epidemiological data and to demonstrate handling of such data, variables were all created as binaries and assigned prevalence = 50%. Each simulation included 1000 participants who exhibited a probability of response of 50% at each wave after the first, with a portion of non-respondents becoming permanently lost to follow-up at each wave (a high level of missing data was simulated to better demonstrate the of the value of the approaches to missing data). Each simulation was repeated 20 times, and the mean estimated outcome prevalence and the mean of its standard error were used to compare the different approaches to missing data.

The first simulation included only four waves of data collection and thus relatively simple patterns of missingness (four patterns based on response to waves 2 and 3, mimicking the example discussed in Section 3). Responsiveness was allowed to depend on baseline variables and variables measured in the previous wave (OR = 0.67 with respect to each correlate), thus satisfying the *missing at random* assumption in any of the approaches that included these variables in missingness or imputation models. The second simulation extended the first by adding 6 additional waves of follow-up, thus complicating the patterns of missing data and increasing the number of relevant variables. The final two simulations compared the robustness of methods to *not missing at random* conditions, where response was directly related to the outcome variable. In simulation 3, responsiveness was allowed to depend on the outcome variable and variables measured in the previous wave. In simulation 4, responsiveness was determined only by the outcome measure. Thus, simulation 3 represents a weaker violation of the missing at random assumption, with only part of the relationship between the outcome and missingness being determined by observed variables, and simulation 4 represents a strong violation where the only thing truly driving missingness is the outcome variable.

Simulations were conducted using Stata 12 (Statacorp, Texas) and input code is provided in the Supplementary Appendix. Correlates were drawn from binomial distributions with the probability of success being a function of the outcome variable (p = 0.55 if outcome = positive; p = 0.45 if outcome = negative). Response variables were drawn from binomial distributions with the probability of success being a function of the outcome variable and correlates, depending on the simulation, set to approximately maintain selection odds with respect to the outcome across simulations. Approaches based on inverse probability-weighting used fitted logistic regressions without modifications (weight stabilisation, etc.). Approaches based on multiple imputation used logit imputation models with 20 imputations.

6.2 Results

Results of each simulation are presented in Table 1. In simulation 1, as would be expected, the methods that ignored information from the intermediate waves (and thus for which the missingness assumption was violated) exhibited greater bias. Stratified and stepped inverse probability-weighting performed comparably to multiple imputation by chained equations in terms of mean bias reduction but with about half the precision. Inclusion of responsiveness (indicators of response to waves 2 and 3) in weighting or imputation models made little difference to the estimates of prevalence or their standard error, improving them only slightly in each case. While multiple imputation by chained equations was able to be implemented, it required the specification of 41 imputation models (one for each variable from waves 2–4).

The increased complexity of patterns of missingness in simulation 2 precluded stratified inverse probability-weighting from being implemented. The increased number of variables that predict both missingness and the outcome (160) further complicated the implementation of multiple imputation by chained equations, leading to computation problems that precluded it being fully implemented. While it may have been feasible to implement some variable selection method to identify reasonably good imputation models given the available power, this was beyond the scope of the study. The effectiveness of stepped inverse probability-weighting diminished slightly, but it was still better than multiple imputation or inverse probability-weighting using baseline measures alone. However, the apparent benefit conveyed by including measures of missingness in multiple imputation or inverse probability-weighting became much more pronounced. Inclusion of responsiveness in either of these methods—even the inclusion of responsiveness alone—was sufficient to remove most bias.

Results of simulation 3 were similar to simulation 2 in terms of the comparative effectiveness of each method and the inability to implement stratified inverse probability-weighting or multiple imputation by chained equations. The importance of including responsiveness, however, became more pronounced. Any method that ignored responsiveness had only limited effect in terms of reducing bias, while methods that included responsiveness exhibited comparable effectiveness to simulation 2, despite the introduction of a direct relationship between the outcome and the probability of response. The effectiveness of stepped inverse probability-weighting diminished in this scenario, even with inclusion of responsiveness.

In simulation 4, with only the outcome truly driving responsiveness, stepped inverse probability-weighting and any method that ignored responsiveness proved essentially ineffective, while other methods that included responsiveness retained excellent levels of bias-reduction.

7 Discussion and conclusions

While multiple imputation is generally preferable to inverse probability-weighting, it may not be well-suited to some scenarios in which large numbers of variables predict both the missing variables of interest and missingness itself. In such cases, stepped or stratified inverse probability-weighting may provide a simple alternative, requiring less model specification and allowing for automated variable selection methods to be incorporated. Simplification always comes at a cost, though; in this case, strengthening of the MAR assumption and sacrificing statistical power. There is a clear trade-off that must be weighed by the analyst within the context of the study.

The value of stepped or stratified inverse probability-weighting is mostly restricted to situations where auxiliary variables from intermediate waves are likely to be particularly valuable. If variables observed at baseline are fairly comprehensive and only limited additional information about mechanisms of missingness is likely to be provided by intermediate or subsequent waves, then focusing on that baseline information—or baseline information plus responsiveness—may be sufficient. Both inverse probability-weighting or multiple imputation with only complete auxiliary variables (including responsiveness) are relatively straightforward procedures in modern computing software. As novel procedures, stepped and stratified inverse probability-weighting require additional coding but potentially less than the adequate specification and testing of multiple imputation by chained equations if a large number of imputation models are required.

What was really highlighted in the results of the simulation studies was the potential value of prior responsiveness in weakening the MAR assumption and improving performance of any method under NMAR conditions. The fact that methods which included responsiveness were still able to address most bias in simulations 3 and 4 implies that that inclusion made the missing at random assumption plausible, despite the omission of variables that were truly related to both missingness and the outcome. Under MAR conditions (when the auxiliary variables matched the simulated determinants), inclusion of responsiveness made little difference to estimates, but as the conditions became increasingly NMAR, responsiveness became increasingly important and some methods that utilised it even performed better than under MAR. The increase in bias that Seaman and White[16] observed with respect to use of missingness indicators in propensity score analysis did not appear to extend to the utilisation of responsiveness in missingness or imputation models.

The simulation study presented here focused on the utilisation of responsiveness and the use of techniques in specific contexts (cohort studies with many waves and many variables). It is not a comprehensive comparison of the techniques under all conditions. Several potentially-relevant study parameters were not varied, including the number of auxiliary variables and the correlations between the auxiliary variables and outcome variable. Some research on the influence of such factors on the performance of techniques has been previously published but it is a relatively unexplored field.[12] These factors are unlikely to influence findings with respect to the utilisation of responsiveness, apart from making it more or less important relative to the value contributed by other auxiliary variables.

How responsiveness should be measured and incorporated to optimise performance of the techniques has also not been explored. One potential pitfall in the utilisation of responsiveness that was not revealed by the simulation study relates to the assumptions that are made about participants lost to follow-up. Loss to follow-up may not indicate a low level of responsiveness and almost certainly does not to the degree that would be indicated by tallying the number of non-responses long after a participant was lost. It is therefore advisable that responsiveness only be measured during the windows in which the participant had fair opportunity to respond (i.e. up until the point at which they become lost to follow-up). Measures of responsiveness may therefore also have some level of missing data, which should be accounted for just like missing data in other auxiliary variables (such as by specifying an imputation model for responsiveness).

In conclusion, cohort studies and trials that involve many waves of follow-up provide a unique opportunity for observing responsiveness of participants over time. This information should not be disregarded in analysis of missing data and its use should be prioritised whenever there are concerns that data are not missing at random.

Acknowledgements

The author would like to thank Dr Shaun Seaman for insightful discussion of the assumptions presented above, and Profs Leonie Segal and Paul Delfabbro for their detailed comments on a draft of this paper.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Drake B and Jonson-Reid M. Poverty and Child Maltreatment. In: Korbin JE and Krugman RD, (eds.). *Handbook of Child Maltreatment*. Springer Netherlands, 2014, p. 131-48.

2. Little RJA and Rubin DB. *Statistical Analysis with Missing Data, Second Edition*. Hoboken, NJ, USA: John Wiley & Sons, 2002.

3. Seaman SR and White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*. 2013; 22: 278-95.

4. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009; 338.

5. Schafer JL and Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002; 7: 147-77.

6. Jaeger M. On Testing the Missing at Random Assumption. In: Fürnkranz J, Scheffer T and Spiliopoulou M, (eds.). *Machine Learning: ECML 2006*. Springer Berlin Heidelberg, 2006, p. 671-8.

7. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, 1987.

8. Muthén B, Kaplan D and Hollis M. On structural equation modeling with data that are not missing completely at random. *Psychometrika*. 1987; 52: 431-62.

9. Allison PD. Estimation of linear models with incomplete data. In: Clogg CC, (ed.). *Sociological methodology*. San Fancisco: Jossey-Bass, 1987, p. 71-103.

10. Cole JC. How to Deal With Missing Data: Conceptual Overview and Details for Implementing Two Modern Methods. In: Osborne JW, (ed.). *Best Practices in Quantitative Methods*. Thousand Oaks, CA: SAGE, 2008, p. 214-39.

11. Wooldridge JM. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*. 2007; 141: 1281-301.

12. Mustillo S and Kwon S. Auxiliary Variables in Multiple Imputation When Data Are Missing Not at Random. *The Journal of Mathematical Sociology*. 2015; 39: 73-91.

13. White IR, Royston P and Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011; 30: 377-99.

14. Fielding S, Fayers PM and Ramsay CR. Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health Qual Life Outcomes*. 2009; 7: 57.

15. Hawkes D and Plewis I. Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2006; 169: 479-91.

16. Seaman S and White I. Inverse Probability Weighting with Missing Predictors of Treatment Assignment or Missingness. *Communications in Statistics - Theory and Methods*. 2014; 43: 3499-515.

17. Rubin DB. Inference and Missing Data. *Biometrika*. 1976; 63: 581-92.

18. Molenberghs G, Beunckens C, Sotto C and Kenward MG. Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2008; 70: 371-88.

19. Listing J and Schlittgen R. Tests If Dropouts Are Missed at Random. *Biometrical Journal*. 1998; 40: 929-35.

20. Fairclough DL. *Design and analysis of quality of life studies in clinical trials*. Boca Raton: CRC Press, 2002.

21. Seaman SR, White IR, Copas AJ and Li L. Combining Multiple Imputation and Inverse-Probability Weighting. *Biometrics*. 2012; 68: 129-37.

22. Thomas C, Hypponen E and Power C. Prenatal exposures and glucose metabolism in adulthood: are effects mediated through birth weight and adiposity? *Diabetes Care*. 2007; 30: 918-24.

23. Rubin DB. Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology*. 2001; 2: 169-88.

24. Bernaards CA, Belin TR and Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat Med*. 2007; 26: 1368-82.

25. Lee KJ and Carlin JB. Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. *Am J Epidemiol*. 2010; 171: 624-32.

Table 1 Simulation results comparing methods for handling non-monotone categorical missing data under *missing at random* and *not missing at random* conditions

Analysis	**Simulation 1[a]**		**Simulation 2[b]**		**Simulation 3[c]**		**Simulation 4[d]**	
	Valid[e]	$\widehat{Pr}$ **(SE)**	**Valid[e]**	$\widehat{Pr}$ **(SE)**	**Valid[e]**	$\widehat{Pr}$ **(SE)**	**Valid[e]**	$\widehat{Pr}$ **(SE)**
Whole sample		50.0		50.0		50.0		50.0
Complete case analysis	no	33.3 (2.3)	no	32.0 (2.3)	no	34.6 (2.4)	no	34.0 (2.4)
IPW using baseline variables only	no	41.4 (3.0)	no	40.3 (3.0)	no	37.1 (2.5)	no	36.4 (2.5)
IPW using baseline variables and responsiveness	no	42.7 (3.2)	no	45.2 (3.2)	no	45.2 (2.8)	no	46.1 (2.8)
Stratified IPW	no	46.5 (4.3)	no	UTC	no	UTC	no	UTC
Stepped IPW without responsiveness	yes	45.7 (4.5)	yes	44.3 (5.1)	no	39.5 (3.6)	no	31.4 (2.8)
Stepped IPW with responsiveness	yes	45.8 (4.6)	yes	44.1 (5.0)	no	41.0 (3.7)	no	34.6 (2.9)
MI using baseline variables only	no	41.3 (2.6)	no	40.6 (2.6)	no	37.5 (2.3)	no	36.8 (2.4)
MI using baseline variables and responsiveness	no	42.8 (2.6)	no	46.4 (2.5)	no	46.9 (2.2)	no	47.8 (2.0)
MI using responsiveness only[f]	no	39.0 (2.7)	no	46.7 (2.6)	no	47.1 (2.2)	no	47.2 (2.3)
MI by chained equations excluding responsiveness	yes	45.8 (2.5)	yes	UTC	no	UTC	no	UTC
MI by chained equations including responsiveness	yes	46.1 (2.6)	yes	UTC	no	UTC	no	UTC

IPW: inverse probability-weighting; MI: multiple imputation; $\widehat{Pr}$: mean estimated prevalence; SE: mean standard error; UTC: unable to compute without substantial modification

[a] Four waves of follow-up, missingness determined by baseline and prior-wave variables.

[b] Ten waves of follow-up, missingness determined by baseline and prior-wave variables.

[c] Ten waves of follow-up, missingness determined by prior-wave and outcome variables.

[d] Ten waves of follow-up, missingness determined by outcome variable alone.

[e] Does the method incorporate all variables that are simulated determinants of missingness (is MCAR/MAR theoretically supported)?

[f] MI using responsiveness only would not usually be considered an appropriate approach, despite its good performance in some of these simulations. It is presented here to illustrate the potential value of responsiveness relative to other auxiliary variables.