Published in final edited form as:
Stat Methods Med Res. 2013 February ; 22(1): 14–30. doi:10.1177/0962280211403597.

On weighting approaches for missing data

Lingling Li,
Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, US

Changyu Shen,
Division of Biostatistics, Indiana University School of Medicine, US

Xiaochun Li, and
Division of Biostatistics, Indiana University School of Medicine, US

James M. Robins
Departments of Biostatistics and Epidemiology, Harvard School of Public Health, US

Abstract

We review the class of inverse probability weighting (IPW) approaches for the analysis of missing data under various missing data patterns and mechanisms. The IPW methods rely on the intuitive idea of creating a pseudo-population of weighted copies of the complete cases to remove selection bias introduced by the missing data. However, different weighting approaches are required depending on the missing data pattern and mechanism. We begin with a uniform missing data pattern (i.e., a scalar missing indicator indicating whether or not the full data is observed) to motivate the approach. We then generalize to more complex settings. Our goal is to provide a conceptual overview of existing IPW approaches and illustrate the connections and differences among these approaches.

Keywords

missing data; inverse probability weighting; missing at random; missing not at random; monotone missing; non-monotone missing

1. Introduction

Interest in the use of secondary healthcare databases (e.g., administrative claims, electronic health records, EHR, cancer registries) for medical research is increasing, partially because these data are readily available, relatively inexpensive to access, and cover large representative populations. However, these databases are collected for non-research purposes. For example, administrative and medical claims databases are assembled for the purposes of administering, billing, and reimbursing healthcare services. Moreover, patients in clinical practice settings are not monitored as closely as those in clinical trials. In consequence, a substantial fraction of the needed data is missing for some subjects. These data issues pose analytic challenges and raise validity concerns.

By design, each of these secondary databases may contain only a subset of the variables of interest. For example, administrative claims data contain information on healthcare insurance membership, drug coverage, healthcare utilizations (i.e., diagnosis and procedure codes), and medication dispensing records. But more detailed clinical information (e.g., BMI, vital signs, laboratory tests results) are recorded in EHR. For cancer patients, the cancer stage and histology information are recorded in cancer registries. As a consequence, systematic missing data occurs for some study participants for whom the data in certain databases are unavailable. Even for those with linked databases, missing data may still occur for reasons such as missed office visits, loss to follow-up, switch of healthcare systems, and coding errors. Thus, failure to appropriately handle missing data may lead to inefficient or even invalid use of available data sources.

The simplest and most commonly used method to deal with missing data is the complete case approach in which standard analyses are applied to subjects with complete data on relevant variables. However, this analysis is biased unless the complete cases are representative of the study population (i.e., the data is missing complete at random, MCAR). This MCAR assumption rarely holds in medical applications.[1]

More advanced statistical methods have been developed in the past decades to deal with missing data under less restrictive missing data mechanisms [2], i.e., missing at random (MAR) and missing not at random (MNAR). MAR means the probability of missingness does not depend on unobserved elements conditional on observed data.[3] MNAR indicates settings in which neither MCAR nor MAR holds. In this paper, we review a class of approaches for missing data - the inverse probability weighting (IPW) approaches. The intuitive idea is to create weighted copies of the complete cases to remove selection bias introduced by missing data processes. The weighting idea originates in the survey sampling literature.[4] It has been further generalized by Robins, Rotnitzky, and others to address a variety of important issues such as confounding bias in observational studies and bias due to missing data.[5–8] Alternatives to IPW include parametric likelihood inference [9–11], parametric Bayesian inference [12–14], and parametric multiple imputation [15–17] inference.

We introduce and illustrate the class of IPW approaches for three missing data patterns, uniform missingness, monotone missingness, and non-monotone missingness. For each pattern, we consider both MAR and MNAR mechanisms. We begin with relatively simple scenarios, and then generalize to more complex settings. Due to space limitations, we do not dwell on mathematical detail but refer the interested readers to the original journal articles or to the books by Tsiatis or van der Laan and Robins.[18,19]

The paper is organized as follows. In Section 2, we introduce the notation and models needed to formalize the missing data patterns and mechanisms we consider. We also introduce four motivating examples. In Section 3, we motivate the weighting approaches by demonstrating the bias in the complete case approach when MCAR does not hold. In Sections 4, 5, 6, we introduce weighting approaches for our three missing data patterns. We conclude with a discussion.

2. Models and notations

We let $\left\{\mathbf{L}_i = \left(\mathbf{W}_i^T, \mathbf{V}_i^T\right)^T, \ i = 1, \ldots, n\right\}$ denote the full data on the n study subjects, where the p-dimensional vector $\mathbf{W}_i = (W_{1,i},\ldots,W_{p,i})^T$ denotes the variables that are always observed for each subject i and the q-dimensional vector $\mathbf{V}_i = (V_{1,i},\ldots,V_{q,i})^T$ denotes the variables that are subject to missingness. We let $\mathbf{R}_i = (R_{1,i},\ldots,R_{q,i})^T$ denote the vector of missing indicators for subject i where the sth element $R_{s,i}$ $(1 \le s \le q)$ equals 1 if $V_{s,i}$ is observed, and 0 otherwise. Let $\mathbf{V}_{(\mathbf{R}_i),i}$ denote the observed components of $\mathbf{V}_i$. Let $\mathbf{O}_i = \left(\mathbf{R}_i, \mathbf{L}_{obs,i} = \left(\mathbf{W}_i^T, \mathbf{V}_{(\mathbf{R}_i),i}^T\right)^T\right)$ denote the observed data for subject i and let $\mathbf{L}_{mis,i} = \mathbf{V}_{(\mathbf{1}-\mathbf{R}_i),i}$ denote the unobserved components of $\mathbf{V}_i$. Here **1** denotes a vector of 1's. This notation can be used to represent a wide class of missing data patterns. For example, in a missing outcome model, **W** represents a vector of covariates and **V** is the outcome of interest Y. The parameter of interest might be the marginal outcome mean $E[Y]$ or the coefficients $\boldsymbol{\beta}$ in an outcome regression model $E[Y \mid \mathbf{W}; \boldsymbol{\beta}]$. In missing data models with missing outcome and covariates, **W** would represent the covariates that are always observed and **V** would include both the outcome of interest and the covariates that are subject to missingness.

Throughout we assume that $(\mathbf{W}_i, \mathbf{V}_i, \mathbf{R}_i)$, $i = 1,\ldots,n$ are independent and identically distributed random vectors. We assume the parameter of interest $\boldsymbol{\beta}^*$ is the unique solution to the equation $E[\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta}^*)] = 0$, where $\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta})$ is a known m-dimensional function of the full data $(\mathbf{W}_i, \mathbf{V}_i)$ and a parameter $\boldsymbol{\beta}$, $\boldsymbol{\beta}^*$ is the true value of $\boldsymbol{\beta}$, and the expectation is under the distribution of $(\mathbf{W}_i, \mathbf{V}_i)$. Thus $\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta})$ is an unbiased estimating function for $\boldsymbol{\beta}^*$. Here $\boldsymbol{\beta}^*$ is a functional of the distribution of the full data $(\mathbf{W}_i, \mathbf{V}_i)$.

We consider the following three missing data patterns: uniform missingness, monotone missingness, and non-monotone missingness. The weighting approach applies equally to all. However, its implementation is much more complicated for non-monotone missing data patterns. We will start with a simple uniform missing pattern to illustrate and motivate the basic idea.

Missing pattern 1: uniform missing data, i.e., $R_1 = \cdots = R_q = R$. Under uniform missingness, either the entire vector $\mathbf{V}_i$ is observed for subject i or it is completely missing. This pattern often occurs when information is extracted from multiple data sources. For example, administrative claims data contain information on basic demographics (age, gender), healthcare utilizations, and medication dispensing records. However, more detailed clinical information such as vital signs and lab test results would be available only for a subset of the study participants with linked EHR data.

Motivating example 1: Consider a hypothetical study evaluating the 1-year incidence rate of heart disease among new users of non-steroidal anti-inflammatory drugs. Data are extracted from a health insurance administrative claims database which contains information on medication dispensing records and disease diagnosis history. The indicator variable $V = Y$ indicates whether heart disease occurred during the 1-year follow-up period after drug initiation. Let $\beta^* = E[Y]$. Then $M(Y\,;\beta)$ is $Y - \beta$. The outcome will be missing in participants who dis-enroll from the insurance plan during the follow-up period. The vector of covariates $\mathbf{W}$ includes demographics (age, gender), geographic region, geographically derived socioeconomic status, and comorbidity conditions.

Missing pattern 2: monotone missing data. Under monotone missingness, if the sth element ($R_s = 0$) of $\mathbf{V}_i$ is missing then all subsequent elements are missing ($R_t = 0$ for any $s < t \leq q$). This pattern occurs frequently in longitudinal studies with repeated measurements in which subjects who drop out of the study never re-enter. Then V_s might denote the data that were to be collected at the sth planned clinic visit. Even if some subjects return after missing one or more visits, one can choose to make the data “monotone” for purposes of data analysis by choosing to ignore in the analysis any data recorded subsequent to a missing visit. Note uniform missing data is actually a special case of monotone missing data.

Motivating example 2: Consider an observational study to compare the effects of two anti-hypertensive agents (e.g., angiotensin-converting enzyme inhibitors and beta-blockers) on reducing blood pressure (BP) level among incident users. The study participants were identified using claims and EHR data. Then $\mathbf{W}$ contains the treatment indicator and some baseline covariates (e.g., age, sex, and comorbidity conditions). The vector $\mathbf{V}$ contains two elements; V_1 records the baseline BP and $V_2 = Y$ records the BP at the end of a 12-month follow-up period. The baseline BP V_1 is incomplete as some patients do not have EHR data available or did not have their BP measured during the baseline period. Similarly, some subjects have $V_2 = Y$ missing. We decide to make the data “monotone” by ignoring the data on V_2 for subjects missing V_1. Suppose we are interested in the coefficient $\boldsymbol{\beta}$ in the regression model $E[Y \mid \mathbf{W}, V_1; \boldsymbol{\beta}] = b(\mathbf{W}, V_1; \boldsymbol{\beta}) = (\mathbf{W}^T, V_1)\boldsymbol{\beta}$. We would take

$$\mathbf{M}(\mathbf{W},\mathbf{V};\beta)=[Y-b(\mathbf{W},V_1;\boldsymbol{\beta})]\begin{pmatrix}\mathbf{W}\\ V_1\end{pmatrix}.$$

Missing pattern 3: non-monotone missing data; non-monotone missingness refers to any missing data pattern that is not monotone. Thus we may have $R_t = 1$ but $R_s = 0$ for some subjects and $R_t = 0$ but $R_s = 1$ for others. This is the most complicated missing data pattern. We consider two motivating examples for this pattern.

Motivating example 3: Consider a regression analysis with missing covariates. Suppose we are interested in identifying predictors of episodes of exacerbation for children with persistent asthma. The study cohort of children with persistent asthma was identified using healthcare claims data. The vector $\mathbf{W}$, ascertained from claims data, includes data on demographic characteristics and a binary outcome encoding 2 or more ER visits for asthma during a 12-month study period. Surveys were mailed to parents to obtain data on a baseline asthma severity score (V_1), household income (V_2), and a measure of the parents’ expectation on child functioning with asthma (V_3). Parents may answer none, one, two, or three of the three questions. This missing pattern is non-monotone. We are interested in the regression parameter $\boldsymbol{\beta}^*$ in a logistic regression model regressing the outcome on potential predictors. The estimating equation $\mathbf{M}(\mathbf{W}, \mathbf{V}; \boldsymbol{\beta})$ is the score function for $\boldsymbol{\beta}$.

Motivating example 4: Consider a longitudinal follow-up study with repeated measurements of BP at three time points, s =1,2,3. As before, **W** contains the treatment indicator and baseline covariates (e.g., age, sex). Let $V_{s,i}$ indicate the BP measured at the sth time point and $\mathbf{V}_i = (V_{1,i}, V_{2,i}, V_{3,i})^T$. Unlike in example 2, we do not ignore subsequent data on subjects missing V_1 or V_2. Thus this missing pattern is non-monotone. We are interested in the mean of $\mathbf{V}_i$, $\boldsymbol{\beta}^* = E[\mathbf{V}_i]$. Thus $\mathbf{M}(\mathbf{W}, \mathbf{V}; \boldsymbol{\beta}) = \mathbf{V} - \boldsymbol{\beta}$.

For each missing pattern, we consider both MAR and MNAR data generating processes.[3] Data are said to be MAR if the conditional missing probabilities given the full data do not depend on the unobserved components of **V**, i.e.,

$$P\left(\mathbf{R}_i{=}\mathbf{r}|\mathbf{W}_i,\mathbf{V}_i\right){=}P\left(\mathbf{R}_i{=}\mathbf{r}|\mathbf{L}_{obs,i}{=}\left(\mathbf{W}_i,\mathbf{V}_{(\mathbf{r}),i}\right)\right) \quad (1)$$

In the special case of MCAR, $P(\mathbf{R}_i = \mathbf{r} \mid \mathbf{W}_i, \mathbf{V}_i)$ is constant. Let $\boldsymbol{\gamma}$ denote the parameters governing the missing data process and $\boldsymbol{\theta}$ denote the parameters governing the distribution of the full data $\mathbf{L} = (\mathbf{W}, \mathbf{V})$, and assume they are variation independent. Then under MAR, the likelihood $f(\mathbf{O}_i, \boldsymbol{\gamma}, \boldsymbol{\theta})$ of the observed data factors into a component $\Pr(\mathbf{R}_i = \mathbf{r} \mid \mathbf{L}_{obs,i}; \boldsymbol{\gamma})$ depending on $\boldsymbol{\gamma}$ alone and a component $f(\mathbf{L}_{obs,i}; \boldsymbol{\theta})$ depending on $\boldsymbol{\theta}$ alone. Thus MAR is referred to as ignorable missingness because the missing data process can be “ignored” in likelihood-based inference on a parameter $\boldsymbol{\beta}^*$ that are functions of the parameters $\boldsymbol{\theta}$ governing the marginal distribution of the full data **L**. The IPW approach takes a different perspective than likelihood-based approaches by using estimates of the missing data process to derive valid inferences on the parameter of interest $\boldsymbol{\beta}^*$.

When MAR fails to hold, the missing data mechanism is said to be MNAR or nonignorable, i.e., the missing probabilities depend on unobserved components of **V** conditional on observed data. In this setting, the parameter of interest is typically unidentifiable unless additional assumptions on the missing data process are imposed. These assumptions usually are investigator specified and cannot be empirically tested when the full data model is nonparametric. Therefore, it is a common practice to conduct a sensitivity analysis in which we vary these additional assumptions over a plausible range and examine how inferences on $\boldsymbol{\beta}^*$ change. As we will show next, weighting approaches in MAR settings can be naturally extended to MNAR settings by specifying a selection bias function to quantify the residual association of the missing probabilities and unobserved components of **V** after adjusting for observed data. Sensitivity analysis can then be conducted by varying the parameters in the selection bias function and/or the functional form.

We let $\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{r})$ denote the conditional missing probability $P(\mathbf{R}_i = \mathbf{r} \mid \mathbf{W}_i, \mathbf{V}_i)$. Throughout we assume that $P(\mathbf{R}_i = \mathbf{1} \mid \mathbf{W}_i, \mathbf{V}_i) > 0$ with probability 1.

3. Why the complete case approach may be biased?

We first illustrate why the complete case approach may be biased when MCAR does not hold.[10] If the full data were observed, $\boldsymbol{\beta}^*$ could be estimated by solving

$$\sum_{i=1}^{n}\mathbf{M}\left(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta}\right)=0, \quad (2)$$

the empirical version of $E[\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta})]$. Unfortunately, when missing data exist, the solution to eq. (2) depends on unobserved components of $\mathbf{V}$. Suppose $E[\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta}^*)] = 0$, but $E[\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta}^*) \mid \mathbf{R}_i = 1] \neq 0$, then if we use complete cases only and estimate $\boldsymbol{\beta}^*$ by solving the estimating equation $\sum_{i=1}^{n} I\left(\mathbf{R}_i=\mathbf{1}\right)\mathbf{M}\left(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta}\right)=0$, it is obvious that the solution to the equation above, $\tilde{\boldsymbol{\beta}}_{cc}$, may be biased unless $E[P(\mathbf{R}_i = \mathbf{1} \mid \mathbf{W}_i, \mathbf{V}_i)\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta}^*)] = 0$, e.g., $P(\mathbf{R}_i = \mathbf{1} \mid \mathbf{W}_i, \mathbf{V}_i)$ is constant.

Heuristically, when MCAR fails to hold, the complete cases are a selected, non-random subsample of the study population. Thus inference obtained by applying standard approaches to the complete cases may be biased for $\boldsymbol{\beta}^*$. The IPW approach restores unbiasedness by creating a pseudo-population in which selection bias due to the missing data is removed. We next introduce the IPW methods for the three missing data patterns respectively.

4. Uniform missing pattern

A uniform missing data pattern is a pattern in which the missing indicator vector $\mathbf{R}$ takes only two possible values $\mathbf{1} = (1,1,\ldots,1,\ldots1)^T$ or $\mathbf{0} = (0,0,\ldots0,\ldots0)^T$. Noted above, unless MCAR holds, the complete case approach is likely biased. To remove selection bias due to missing data, the IPW approach weights each subject i with complete data ($\mathbf{R}_i = \mathbf{1}$) by the inverse of the conditional probability of observing the full data $\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})$. For illustration, we temporarily assume $\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})$ is a known function of $(\mathbf{W}_i, \mathbf{V}_i)$ as is the case in studies with missingness by design (e.g., studies with two-stage sampling). Then, the simple IPW estimator $\hat{\boldsymbol{\beta}}_0$ solves the following estimating equation[20]

$$\sum_{i=1}^{n}\frac{I\left(\mathbf{R}_i=\mathbf{1}\right)}{\pi_i\left(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1}\right)}\mathbf{M}\left(\mathbf{W}_i,\mathbf{V}_i;\hat{\boldsymbol{\beta}}_0\right)=0 \quad (3)$$

Under regularity conditions, $\hat{\boldsymbol{\beta}}_0$ is a consistent estimator of $\boldsymbol{\beta}^*$ since

$$\begin{aligned}E\left[\tfrac{I(\mathbf{R}_i=\mathbf{1})}{\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1})}\mathbf{M}\left(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta}^*\right)\right]&=E\left[E\left(\tfrac{I(\mathbf{R}_i=\mathbf{1})}{\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1})}|\mathbf{W}_i,\mathbf{V}_i\right)\mathbf{M}\left(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta}^*\right)\right]\\&=E\left[\mathbf{M}\left(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta}^*\right)\right]=0\end{aligned}$$

Note that the above equalities hold regardless of whether or not the missingness is ignorable (i.e., MAR or MNAR). In addition, a fully parametric model for the full data is not required. Under mild conditions, the solution to eq. (3) is a consistent and asymptotically normal (CAN) estimator of $\boldsymbol{\beta}^*$.[20]

This IPW estimator $\hat{\boldsymbol{\beta}}_0$ demonstrates the fundamental principle of the weighting approach; weighted copies of complete cases remove the selection bias introduced by the missing data process. However, note eq. (3) depends only on data from complete cases. Then $\hat{\boldsymbol{\beta}}_0$ is not fully efficient. To increase efficiency, we can add to the estimating equation augmentation terms. These terms depend on data from both complete and incomplete cases.

From the definition of $\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})$, it is clear that an augmentation term $\mathbf{A}_i(\boldsymbol{\varphi})$ that takes the form $(I(\mathbf{R}_i = \mathbf{1})\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})^{-1} - 1)\boldsymbol{\varphi}(\mathbf{W}_i)$ has mean zero, where $\boldsymbol{\varphi}(\mathbf{W}_i)$ is an m-dimensional vector of arbitrary functions of the always observed variables $\mathbf{W}_i$. Let $\mathbf{D}_i(\boldsymbol{\beta}, \boldsymbol{\varphi})$ be $\frac{I(\mathbf{R}_i=\mathbf{1})}{\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1})}\mathbf{M}(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta})-\mathbf{A}_i(\varphi)$. Then $\mathbf{D}_i(\boldsymbol{\beta}, \boldsymbol{\varphi})$ is mean zero at $\boldsymbol{\beta}^*$ and the solution $\hat{\boldsymbol{\beta}}_{\boldsymbol{\varphi}}$ to $\sum_{i=1}^{n}\mathbf{D}_i(\boldsymbol{\beta},\boldsymbol{\varphi})=0$ is a consistent estimator of $\boldsymbol{\beta}^*$ under regularity conditions.[21] Moreover, the asymptotic variance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\varphi}}$ equals $\boldsymbol{\Gamma}^{-1}$ var$[\mathbf{D}_i(\boldsymbol{\beta}^*, \boldsymbol{\varphi})]\boldsymbol{\Gamma}^{-1,T}$ where $\boldsymbol{\Gamma}\equiv E\left[\frac{\partial\mathbf{M}(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right]$. This implies that the choice of $\boldsymbol{\varphi}$ affects the efficiency of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\varphi}}$ only through the term var$[\mathbf{D}_i(\boldsymbol{\beta}, \boldsymbol{\varphi})]$. By simple algebra, one can easily show that

$$\mathbf{D}_i(\boldsymbol{\beta},\boldsymbol{\varphi})=\mathbf{M}(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta})+\left(\frac{I(\mathbf{R}_i=\mathbf{1})}{\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1})}-1\right)(\mathbf{M}(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta})-\boldsymbol{\varphi}_i)$$

and

$$\text{var}[\mathbf{D}_i(\boldsymbol{\beta},\boldsymbol{\varphi})]=\text{var}[\mathbf{M}(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta})]+\text{var}\left[\left(\frac{I(\mathbf{R}_i=\mathbf{1})}{\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1})}-1\right)(\mathbf{M}(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta})-\boldsymbol{\varphi}_i)\right]$$

as the two terms in the above representation of $\mathbf{D}_i(\boldsymbol{\beta}, \boldsymbol{\varphi})$ are uncorrelated. We want to select $\boldsymbol{\varphi}$ so that var$[\mathbf{D}_i(\boldsymbol{\beta}, \boldsymbol{\varphi})] \le$ var$[\mathbf{D}_i(\boldsymbol{\beta}, \boldsymbol{\varphi} = \mathbf{0})]$ for any $\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta})$. Since the first term in var$[\mathbf{D}_i(\boldsymbol{\beta}, \boldsymbol{\varphi})]$ does not depend on $\boldsymbol{\varphi}$, we need to select $\boldsymbol{\varphi}$ such that var$[(I(\mathbf{R}_i = \mathbf{1})\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})^{-1} - 1)\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta}) - \mathbf{A}_i(\boldsymbol{\varphi}_i)] \le$ var$[(I(\mathbf{R}_i = \mathbf{1})\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})^{-1} - 1)\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta})]$. The inequality above is satisfied when $\mathbf{A}_i(\boldsymbol{\varphi}) = (I(\mathbf{R}_i = \mathbf{1})\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})^{-1} - 1)\boldsymbol{\varphi}(\mathbf{W}_i)$ is the projection of $(I(\mathbf{R}_i = \mathbf{1})\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})^{-1} - 1)\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta})$ onto a subspace Λ_{sub} of $\Lambda_1 \equiv \{(I(\mathbf{R}_i = \mathbf{1})\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})^{-1} - 1)h(\mathbf{W}_i) : h \in L_2(f(\mathbf{W}))\}$, as the norm of the residual from a projection is smaller than or equal to the norm of the original vector. For a given $\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta})$, the most efficient augmentation term, $\boldsymbol{\varphi}_{eff}$, is obtained by projecting $[I(\mathbf{R}_i = \mathbf{1})\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})^{-1} - 1]\mathbf{M}(\mathbf{W}_i, \mathbf{V}_i; \boldsymbol{\beta})$ onto the entire space Λ_1. With uniform missing patterns, when MAR holds, $\boldsymbol{\varphi}_{eff}$ equals $E[\mathbf{M} \mid \mathbf{R} = \mathbf{1}, \mathbf{W}]$. For example, in our motivating example 1, $M = Y - \beta$ and thus $\varphi_{eff} = E[Y \mid R = 1, \mathbf{W}] - \beta$. See references for technical details.[6,19–30]

So far we have assumed that $\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})$ is known, i.e., missingness by design, which occurs infrequently in medical applications. Therefore, we need to estimate $\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})$ using the observed data. We next discuss strategies to obtain estimated missing probabilities $\hat{\pi}_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})$ under MAR and MNAR mechanisms respectively.

4.1. MAR

Under MAR, by eq. (1), $\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{0})$ depends on $\mathbf{W}_i$ only since $\mathbf{V}_{(\mathbf{0}),i}$ is an empty set. Thus $\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})$ also depends only on $\mathbf{W}_i$ since $\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1}) = 1 - \pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{0})$. In other words, for $\mathbf{r} \in \{\mathbf{1},\mathbf{0}\}$, $P(\mathbf{R}_i = \mathbf{r} \mid \mathbf{W}_i, \mathbf{V}_i) = P(\mathbf{R}_i = \mathbf{r} \mid \mathbf{W}_i)$. Since $(\mathbf{R}_i, \mathbf{W}_i)$ is observed for each subject i, then the estimated conditional missing probability $\hat{\pi}_i(\mathbf{W}_i, \mathbf{r})$ can be obtained by regressing the missing indicator $\mathbf{R}_i$ on the always observed covariates $\mathbf{W}_i$ via either a parametric regression model (e.g., logistic regression) or nonparametric, data-adaptive algorithms (e.g., tree-based methods).[31–35]

In many studies that obtain data from electronic medical databases, the number of covariates that need to be adjusted for to make the MAR assumption plausible is quite large.[36] Then it will be difficult to impose a correct parametric model for $P(\mathbf{R}_i = \mathbf{1} \mid \mathbf{W}_i)$ due to the curse of dimensionality. A misspecified parametric model may result in significantly biased results. Data-adaptive, tree-based methods provide promising alternatives.[32,33,35] They are designed to minimize the mean squared prediction error, no matter how many covariates need to be adjusted for. The methods are easy to implement with minimum analyst input. Trees have many advantages including being robust to outliers, insensitive to covariate transformation, and the ability to capture complex interactions and highly correlated variables. See Hastie, Tibshrani, and Friedman[35] and Therneau & Atkinsoon[37] for a comprehensive review of the method and software programs.

After $\{\hat{\pi}_i(\mathbf{W}_i,\mathbf{1}), i = 1,\ldots,n\}$ are obtained, the IPW estimator $\hat{\boldsymbol{\beta}}_0$ is obtained by solving eq. (3), with $\hat{\pi}_i(\mathbf{W}_i,\mathbf{1})$ substituted for $\pi_i(\mathbf{W}_i,\mathbf{1})$. To obtain the efficient augmented IPW estimators $\hat{\boldsymbol{\beta}}_{\varphi_{eff}}$, additional modeling and estimation are needed since $\boldsymbol{\varphi}_{eff}$ depends on the unknown outcome regression function $E[\mathbf{M} \mid \mathbf{R} = \mathbf{1}, \mathbf{W}]$. In example 1, $\varphi_{eff} = E[Y \mid \mathbf{R} = \mathbf{1}, \mathbf{W}] - \beta$. We use the complete cases to estimate $E[Y \mid \mathbf{R} = \mathbf{1}, \mathbf{W}]$. As before, we can use either a parametric working model $E[Y \mid \mathbf{R} = \mathbf{1}, \mathbf{W}; \xi]$ or data-adaptive, tree-based regression techniques. After all the unknown functions and parameters are estimated, the augmented estimator $\hat{\boldsymbol{\beta}}_{\varphi_{eff}}$ is obtained by solving the augmented estimating equation $\sum_{i=1}^{n} \mathbf{D}_i \left(\boldsymbol{\beta}, \hat{\pi}_i, \hat{\boldsymbol{\varphi}}_{eff}\right) = 0$. In this example 1,

$$\hat{\beta}_{\varphi_{eff}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{I(\mathbf{R}_i = \mathbf{1})}{\hat{\pi}_i(\mathbf{W}_i, \mathbf{1})} Y_i - \left(\frac{I(\mathbf{R}_i = \mathbf{1})}{\hat{\pi}_i(\mathbf{W}_i, \mathbf{1})} - 1 \right) \hat{E}\left[Y_i \mid \mathbf{R}_i = \mathbf{1}, \mathbf{W}_i\right] \right\}.$$

It is worth noting that $\hat{\beta}_{\varphi_{eff}}$ is doubly robust (DR) in the sense that it is consistent for $\boldsymbol{\beta}^*$ if either the working model for the missing data process $\pi(\mathbf{W}_i, \mathbf{1})$ or the working model for the outcome regression function $E[Y \mid \mathbf{R} = 1, \mathbf{W}]$ is correctly specified, but not necessarily both.[38] This nice property offers analysts two chances of making correct inference. Furthermore, the specified working models are practically certain to be incorrect especially in the presence of high-dimensional covariates. But as long as at least one model is nearly correct, the bias of $\hat{\beta}_{\varphi_{eff}}$ will be small by theory and simulation results.[38] The variance estimates of $\hat{\beta}_{\varphi_{eff}}$ can be obtained using either the asymptotic theory and delta methods or bootstrap re-sampling approaches.

4.2. MNAR

The MAR assumption cannot be empirically tested using observed data except under limited scenarios.[39] Subject matter expertise is usually required to judge its plausibility. When MAR does not appear to be reasonable, then additional assumptions on the missing data process need to be imposed to make the parameters of interest identifiable. Since these additional assumptions are not verifiable under a nonparametric full data model for ($\mathbf{W}$, $\mathbf{V}$), a sensitivity analysis is recommended. There are different ways of conducting a sensitivity analysis for MNAR (i.e., nonignorable) data. We focus on the selection bias function approach for IPW estimators.[27,30] This approach decomposes the nonignorable missing data process in a natural and straightforward manner, and thus makes it relatively easy to impose sensitivity assumptions using background information and substance knowledge.

Under MNAR, $\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{0})$ depends on both $\mathbf{W}_i$ and $\mathbf{V}_i$. The selection bias function approach uses a user-specified function to quantify the residual association between the missingness probability and the possibly unobserved components of $\mathbf{V}$ conditioning on observed data. Specifically, we assume that

$$\frac{\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{0})}{\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1})}=\frac{P(\mathbf{R}_i=\mathbf{0}|\mathbf{W}_i,\mathbf{V}_i)}{P(\mathbf{R}_i=1|\mathbf{W}_i,\mathbf{V}_i)}=\exp\{h(\mathbf{W}_i)+q(\mathbf{W}_i,\mathbf{V}_i)\} \quad (4)$$

where $h(\mathbf{W}_i)$ is an unrestricted function of $\mathbf{W}_i$ and $q(\mathbf{W}_i, \mathbf{V}_i)$ is the selection bias function. In other words, the “odds” of having missing data depends on the possibly unobserved components $\mathbf{V}_i$ through the selection bias function $q(\mathbf{W}_i, \mathbf{V}_i)$. Note that $q(\mathbf{W}_i, \mathbf{V}_i)$ needs to be specified by investigators, e.g., $q(\mathbf{W}_i, \mathbf{V}_i; \mathbf{c}) = \mathbf{c}^T\mathbf{V}_i$ where $\mathbf{c}$ is a given constant vector. When the model for the full data is nonparametric, the functional form chosen for $q(\mathbf{W}_i, \mathbf{V}_i)$ and the value of the parameter $\mathbf{c}$ are not empirically testable. In this paper, we do not dwell on the choice of the selection bias function $q(\mathbf{W}_i, \mathbf{V}_i)$ as it depends heavily on the study setting and existing substance knowledge about the missing mechanism. [27,30]

Assuming eq. (4) holds and $q(\mathbf{W}_i, \mathbf{V}_i)$ has been specified, we still need to estimate $h(\mathbf{W}_i)$ to obtain an estimated missing probability $\hat{\pi}_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1})$. To do so, we usually impose a parametric working model $h(\mathbf{W}_i; \boldsymbol{\alpha})$ indexed by a unknown parameter $\boldsymbol{\alpha}$, e.g., $h(\mathbf{W}_i; \boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\mathbf{W}_i$. If $\mathbf{W}$ is categorical and the sample size is large, then we can use a saturated model to avoid model misspecification. The parameter estimate $\hat{\boldsymbol{\alpha}}$ is obtained by solving the unbiased estimating equation

$$\sum_{i=1}^{n}\mathbf{A}_i(\psi)=\sum_{i=1}^{n}\left(\frac{I(\mathbf{R}_i=\mathbf{1})}{\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1};\boldsymbol{\alpha},q)}-1\right)\psi(\mathbf{W}_i)=0,$$

where $\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1}; \boldsymbol{\alpha}, q) = [1 + \exp\{h(\mathbf{W}_i; \boldsymbol{\alpha}) + q(\mathbf{W}_i, \mathbf{V}_i)\}]^{-1}$ and $\boldsymbol{\psi}$ is a vector of selected functions of $\mathbf{W}_i$ (e.g., $\boldsymbol{\psi}(\mathbf{W}_i) = \mathbf{W}_i$). Note that the dimension of $\boldsymbol{\psi}$ needs to be equal to the dimension of $\boldsymbol{\alpha}$. Under regularity conditions, the corresponding $\hat{\boldsymbol{\alpha}}$ is consistent for the true value $\boldsymbol{\alpha}^*$ as long as the parametric working model is correct and eq. (4) holds. However, the variance of $\hat{\boldsymbol{\alpha}}$ depends on $\boldsymbol{\psi}$.

As with MAR settings, the IPW estimator $\hat{\boldsymbol{\beta}}_0$ can be obtained as the solution to eq. (3) using the estimated missing probability $\hat{\pi}_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1}) = \pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{1}; \hat{\boldsymbol{\alpha}}, q)$. See references[27,28,40] for details on doubly-robust estimators and other, more efficient augmented estimators.

5. Monotone missing pattern

We now introduce the weighting approach for monotone missing patterns. Without loss of generality, we assume $R_{s,i} \geq R_{t,i}$ for any $1 \leq s < t \leq q$. Equivalently, for each subject i, if the sth element $V_{s,i}$ is missing, then all subsequent elements $\{V_{t,i}: t > s\}$ are missing.

We first focus on example 2 and then present general results. Specially, we consider the setting in which $\mathbf{W}_i$ contains the treatment indicator and a vector of baseline covariates that are recorded for each subject (e.g., age, sex, comorbidity conditions); while $\mathbf{V}_i = (V_{1,i}, Y_i)^T$ denotes the BP measured at baseline and at 12 months. We make the data "monotone" by ignoring $Y = V_2$ on subjects missing V_1 ($R_{2,i} = 0$ if $R_{1,i} = 0$). We will estimate the coefficients $\boldsymbol{\beta}$ in the outcome regression model

$$E\left[Y|\mathbf{W},V_1;\boldsymbol{\beta}\right]=b\left(\mathbf{W},V_1;\boldsymbol{\beta}\right)=\left(\mathbf{W}^T,V_1\right)\boldsymbol{\beta}\ \text{with}\ \mathbf{M}\left(\mathbf{W},\mathbf{V};\beta\right)=\left[Y-b\left(\mathbf{W},V_1;\boldsymbol{\beta}\right)\right]\begin{pmatrix}\mathbf{W}\\ V_i\end{pmatrix}.$$

Monotone missing data can be analyzed by applying the weighting approach for a uniform missing pattern in a nested fashion; that is, a monotone missing pattern can be decomposed into multiple uniform missing data models. For example, in example 2, since we have two missing components, we derive our estimators in two steps. In the first step, we derive estimators under an artificial missing data model in which the full data is $\mathbf{L}_i=\left(\mathbf{W}_i^T,\mathbf{V}_i^T\right)^T$ but the observed data is $\mathbf{O}_i^{\diamond}=\left(\mathbf{W}_i^T,R_{1,i},R_{1,i}V_{1,i},R_{1,i}Y_i\right)^T$. That is, both $V_{1,i}$ and Y_i are observed whenever the missing indicator $R_{1,i}$ is 1. In the second step, we consider a second artificial missing data model with $\mathbf{O}_i^{\diamond}$ now the full data and $\mathbf{O}_i = \mathbf{W}_i, R_{1,i}, R_{2,i}, R_{1,i}V_{1,i}, R_{2,i}Y_i$ the observed data. Our final estimator will only depend on the actual data $\{\mathbf{O}_i, i =1,\ldots,n\}$.

Specifically, let $E_{1,i} = e_{1,i}(\mathbf{W}_i, V_{1,i}, Y) \equiv P(R_{1,i} = 1 \mid \mathbf{W}_i, V_{1,i}, Y_i)$ and $E_{2,i} = e_{2,i}(\mathbf{W}_i, V_{1,i}, Y_i) \equiv P(R_{2,i} = 1 \mid R_{1,i} = 1, \mathbf{W}_i, V_{1,i}, Y_i)$. Then, under monotone missingness, $\pi_i\,(\mathbf{W}_i,V_{1,i},Y,\mathbf{1})= P(\mathbf{R}_i =\mathbf{1} \mid \mathbf{W}_i,V_{1,i},Y_i)= E_{1,i}\,E_{2,i}$, $\pi_i\,(\mathbf{W}_i,V_{1,i},Y,(1,0)^T)= P(\mathbf{R}_i = (1,0)^T \mid \mathbf{W}_i,V_{1,i},Y_i)= E_{1,i}(1-E_{2,i})$ and $\pi_i\,(\mathbf{W}_i,V_{1,i},Y,\mathbf{0})= P(\mathbf{R}_i = \mathbf{0} \mid \mathbf{W}_i,V_{1,i},Y_i)=1-E_{1,i}$. As above, suppose $e_{1,i}$ and $e_{2,i}$ are known functions. Later we relax this assumption.

The first step of our estimation procedure is to apply the IPW approach to the first artificial missing data model. In Section 4, we obtain a first-stage class of estimators $\{\tilde{\boldsymbol{\beta}}_{\boldsymbol{\varphi}_1}: \boldsymbol{\varphi}_1\}$ by solving the estimating equation $\sum_{i=1}^{n}\tilde{\mathbf{D}}_i\left(\boldsymbol{\beta},\boldsymbol{\varphi}_1\right)=0$ where

$$\tilde{\mathbf{D}}_i\left(\boldsymbol{\beta},\boldsymbol{\varphi}_1\right)=\frac{R_{1,i}}{E_{1,i}}\mathbf{M}\left(\mathbf{W}_i,V_{1,i},Y_i;\boldsymbol{\beta}\right)-\left(\frac{R_{1,i}}{E_{1,i}}-1\right)\boldsymbol{\varphi}_1\left(\mathbf{W}_i\right).$$

Here $\boldsymbol{\varphi}_1$ is a vector of selected functions of the observed components $\mathbf{W}_i$. However, the first term in $\tilde{\mathbf{D}}_i$ $(\boldsymbol{\beta},\boldsymbol{\varphi}_1)$depends on the outcome Y_i which might still be missing in the actual data even if $R_{1,i}$ =1. To obtain unbiased estimating equations that depend only on the observed data $\mathbf{O}_i$, in the second stage of our estimation procedure, we apply the IPW approach to the second artificial missingness model, where $\mathbf{O}_i^\diamond$ is now the full data and $\mathbf{O}_i$ is the observed data. Note that in this artificial missingness model, the missing indicator does not equal $R_{2,i}$. Rather, the missing indicator equals one when the “full” data and the observed data are the same. Since $\mathbf{O}_i=\mathbf{O}_i^\diamond$ if $R_{1,i}$ = 0 or $R_{1,i}$ = $R_{2,i}$ =1, we define a new missing indicator

$$\tilde{R}_i=\left(1-R_{1,i}\right)+R_{2,i}$$

with $\tilde{E}_i\equiv P\left(\tilde{R}_i=1|\mathbf{O}_i^\diamond\right)=\left(1-R_{1,i}\right)+R_{1,i}E_{2,i}$. Thus, our second-stage IPW estimators $\{\hat{\boldsymbol{\beta}}_{(\boldsymbol{\varphi}_1\boldsymbol{\varphi}_2)}$: $\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2)\}$ are solutions to the estimating equation $\sum_{i=1}^{n}\mathbf{D}_i\left(\boldsymbol{\beta},\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2\right)=0$ where

$$\begin{gathered}\mathbf{D}_i\left(\boldsymbol{\beta},\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2\right)=\frac{\tilde{R}_i}{\tilde{E}_i}\tilde{\mathbf{D}}_i\left(\boldsymbol{\beta},\boldsymbol{\varphi}_1\right)-\left(\frac{\tilde{R}_i}{\tilde{E}_i}-1\right)\boldsymbol{\varphi}_2\left(\mathbf{W}_i,R_{1,i},R_{1,i}V_{1,i}\right)\\ =\frac{\tilde{R}_i}{\tilde{E}_i}\frac{R_{1,i}}{E_{1,i}}\mathbf{M}\left(\mathbf{W}_i,V_{1,i},Y_i;\boldsymbol{\beta}\right)-\frac{\tilde{R}_i}{\tilde{E}_i}\left(\frac{R_{1,i}}{E_{1,i}}-1\right)\boldsymbol{\varphi}_1\left(\mathbf{W}_i\right)-\left(\frac{\tilde{R}_i}{\tilde{E}_i}-1\right)\boldsymbol{\varphi}_2\left(\mathbf{W}_i,R_{1,i},R_{1,i}V_{1,i}\right).\end{gathered}$$

By definition, $\frac{\tilde{R}_i}{\tilde{E}_i}\frac{R_{1,i}}{E_{1,i}}=\frac{R_{1,i}R_{2,i}}{E_{1,i}E_{2,i}}$ and $\frac{\tilde{R}_i}{\tilde{E}_i}=\left(1-R_{1,i}\right)+R_{1,i}\frac{R_{2,i}}{E_{2,i}}$. Thus, $(\tilde{R}_i\ \tilde{E}_i^{-1}-1)\boldsymbol{\varphi}_2(\mathbf{W}_i$, $R_{1,i}$, $R_{1,i}V_{1,i})$= $R_{1,i}$ $(R_{2,i}\ E_{2,i}{}^{-1}-1)$ $\boldsymbol{\varphi}_2$ $(\mathbf{W}_i$, $R_{1,i}$ $=1,V_{1,i})$. For simplicity, we denote $\boldsymbol{\varphi}_2$ $(\mathbf{W}_i$, $R_{1,i}$ $=1,V_{1,i})$ as $\boldsymbol{\varphi}_2$ $(\mathbf{W}_i,V_{1,i})$. After some algebra, one has

$$\begin{aligned}&\mathbf{D}_i\left(\boldsymbol{\beta},\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2\right)\\ &\quad=\frac{R_{1,i}R_{2,i}}{E_{1,i},E_{2,i}}\mathbf{M}\left(\mathbf{W}_i,V_{1,i},Y_i;\boldsymbol{\beta}\right)\\ &\quad-\frac{R_{1,i}}{E_{1,i}}\left(\frac{R_{2,i}}{E_{2,i}}-1\right)\left[\boldsymbol{\varphi}_2\left(\mathbf{W}_i,V_{1,i}\right)\pi_{1,i}+\left(1-\pi_{1,i}\right)\boldsymbol{\varphi}_1\left(\mathbf{W}_i\right)\right]\\ &\quad+\left(\frac{R_{1,i}}{E_{1,i}}-1\right)\boldsymbol{\varphi}_1\left(\mathbf{W}_i\right)\end{aligned}\qquad(5)$$

Under regularity conditions, it can be proved that $\hat{\boldsymbol{\beta}}_{(\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2)}$ is a CAN estimator of $\boldsymbol{\beta}^*$.[21] Let $\boldsymbol{\varphi}_1^\diamond\left(\mathbf{W}_i\right)\equiv\boldsymbol{\varphi}_1\left(\mathbf{W}_i\right)$ and $\boldsymbol{\varphi}_2^\diamond\left(\mathbf{W}_i,V_{1,i}\right)\equiv\boldsymbol{\varphi}_2\left(\mathbf{W}_i,V_{1,i}\right)\pi_{1,i}+\left(1-\pi_{1,i}\right)\boldsymbol{\varphi}_1\left(\mathbf{W}_i\right)$. We can rewrite $\mathbf{D}_i$ $(\boldsymbol{\beta},\boldsymbol{\varphi}_1$, $\boldsymbol{\varphi}_2)$ as

$$\frac{R_{1,i}R_{2,i}}{E_{1,i}E_{2,i}}\mathbf{M}\left(\mathbf{W}_i,V_{1,i},Y_i;\boldsymbol{\beta}\right)-\frac{R_{1,i}}{E_{1,i}}\left(\frac{R_{2,i}}{E_{2,i}}-1\right)\boldsymbol{\varphi}_2^{\diamond}\left(\mathbf{W}_i,V_{1,i}\right)+\left(\frac{R_{1,i}}{E_{1,i}}-1\right)\boldsymbol{\varphi}_1^{\diamond}\left(\mathbf{W}_i\right).$$

To maximize efficiency under MAR, we select $\boldsymbol{\varphi}_2^{\diamond}\left(\mathbf{W}_i,V_{1,i}\right)$ to be $E[\mathbf{M}(\mathbf{W}_i,V_{1,i},Y_i;\boldsymbol{\beta})|\ \mathbf{R}_i=\mathbf{1},\mathbf{W}_i,V_{1,i}]$ and $\boldsymbol{\varphi}_1^{\diamond}\left(\mathbf{W}_i\right)$ to be $E\left[\boldsymbol{\varphi}_2^{*}\left(\mathbf{W}_i,V_{1,i}\right)|R_{1,i}=1,\mathbf{W}_i\right]$. See Robins, Rotnitzky, and others for further discussions of efficiency.[5,6,20–22,24,25,40]

Next we consider how to estimate $E_{1,i}$ and $E_{2,i}$ under MAR and MNAR mechanisms respectively.

5.1. MAR

If MAR holds, then for $\mathbf{r} = (r_1, r_2)^T \in \{\mathbf{1},\mathbf{0},(1,0)^T\}$,

$$\pi_1\left(\mathbf{W}_i,V_{1,i},Y,\mathbf{r}\right)=P\left(\mathbf{R}_i=\mathbf{r}|\mathbf{W}_i,V_{1,i},Y_i\right)=P\left(\mathbf{R}_i=\mathbf{r}|\mathbf{W}_i,r_1V_{1,i},r_2Y_i\right).$$

Thus $E_{1,i} =1-P(\mathbf{R}_i = \mathbf{0} \mid \mathbf{W}_i)= P(R_{1,i} =1|\ \mathbf{W}_i)$ is a function of $\mathbf{W}_i$ only, whereas

$$\begin{aligned}E_{2,i}&=P\left(R_{2,i}=1|R_{1,i}=1,\mathbf{W}_i,V_{1,i},Y\right)\\&=1-P\left(R_{2,i}=0|R_{1,i}=1,\mathbf{W}_i,V_{1,i},Y\right)\\&=1-P\left(\mathbf{R}_i=(1,0)^T|\mathbf{W}_i,V_{1,i}\right)E_{1,i}^{-1}\end{aligned}$$

depends on $(\mathbf{W}_i,V_{1,i})$. That is, $E_{2,i} = P(R_{2,i} =1|\ R_{1,i} =1,\ \mathbf{W}_i,V_{1,i})$. Therefore, $E_{1,i}$ can be estimated using the observed data $\{(R_{1,i},\ \mathbf{W}_i):\ i =1,\ldots,n\}$ by regressing $R_{1,i}$ on $\mathbf{W}_i$ using either a parametric working model or data-adaptive nonparametric techniques. Similarly, $E_{2,i}$ can be estimated using the observed data $\{(R_{2,i},\ \mathbf{W}_i,V_{1,i}):\ i \in \{1,\ldots,n\}$ and $R_{1,i} =1\}$ by regressing $R_{2,i}$ on $(\mathbf{W}_i,V_{1,i})$ among those with $R_{1,i} =1$.

5.2. MNAR

When the missing data process depends on possibly unobserved data and the full data model is nonparametric, we must impose additional assumptions to make the parameters of interest identifiable. We extend the sensitivity analysis approach for the uniform missing pattern and assume that

$$\begin{aligned}\frac{1-E_{1,i}}{E_{1,i}}&=\frac{P(R_{1,i}=0|\mathbf{W}_i,V_{1,i},Y_i)}{P(R_{1,i}=1|\mathbf{W}_i,V_{1,i},Y_i)}=\exp\left(h_1\left(\mathbf{W}_i\right)+q_1\left(\mathbf{W}_i,V_{1,i},Y_i\right)\right)\\\frac{1-E_{2,i}}{E_{2,i}}&=\frac{P(R_{2,i}=0|R_{1,i}=1,\mathbf{W}_i,V_{1,i},Y_i)}{P(R_{2,i}=1|R_{1,i}=1,\mathbf{W}_i,V_{1,i},Y_i)}=\exp\left(h_2\left(\mathbf{W}_i,V_{1,i}\right)+q_2\left(\mathbf{W}_i,V_{1,i},Y_i\right)\right).\end{aligned}$$

Here $q_1\ (\mathbf{W}_i,V_{1,i},Y_i)$ and $q_2\ (\mathbf{W}_i,V_{1,i},Y_i)$ are investigator-specified selection bias functions. To estimate $h_1\ (\mathbf{W}_i)$ and $h_2\ (\mathbf{W}_i,V_{1,i})$, we impose parametric working models $h_1\ (\mathbf{W}_i;\boldsymbol{\alpha})$ and

h_2 ($\mathbf{W}_{\boldsymbol{i}}$, $V_{1,i}$;$\boldsymbol{\alpha}$), and obtain the estimated parameter $\hat{\boldsymbol{\alpha}}$ by solving the unbiased estimating equation $\sum_{i=1}^{n}\mathbf{A}_i(\boldsymbol{\psi})=0$ where

$$\mathbf{A}_i(\boldsymbol{\psi})=\sum_{\mathbf{r}\in\left\{\mathbf{0},(1,0)^T\right\}}\left\{I(\mathbf{R}_i=\mathbf{r})-\frac{I(\mathbf{R}_i=\mathbf{1})}{\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1};\widehat{\boldsymbol{\alpha}},q_1,q_2)}\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{r};\widehat{\boldsymbol{\alpha}},q_1,q_2)\right\}\boldsymbol{\psi}_r\left(\mathbf{W}_i,\mathbf{r}(V_{1,i},Y_i)^T\right)$$

Here

$$\begin{aligned}\widehat{E}_{1,i}=E_{1,i}(\widehat{\boldsymbol{\alpha}},q_1)=[1+\exp\{h_1(\mathbf{W}_i;\widehat{\boldsymbol{\alpha}})+q_1(\mathbf{W}_i,\mathbf{V}_i)\}]^{-1}\\ \widehat{E}_{2,i}=E_{2,i}(\widehat{\boldsymbol{\alpha}},q_2)=[1+\exp\{h_2(\mathbf{W}_i,V_{1,i};\widehat{\boldsymbol{\alpha}})+q_2(\mathbf{W}_i,\mathbf{V}_i)\}]^{-1},\end{aligned}$$

$\boldsymbol{\pi}_i$($\mathbf{W}_i$, $\mathbf{V}_i$,$\mathbf{1}$;$\hat{\boldsymbol{\alpha}}$, q_1, q_2)= $\hat{E}_{1,i}$ $\hat{E}_{2,i}$, $\boldsymbol{\pi}_i$($\mathbf{W}_i$, $\mathbf{V}_i$,$(1,0)^T$;$\hat{\boldsymbol{\alpha}}$, q_1, q_2)= $\hat{E}_{1,i}$(1−$\hat{E}_{2,i}$), and $\boldsymbol{\pi}_i$ ($\mathbf{W}_i$, $\mathbf{V}_i$, $\mathbf{0}$;$\hat{\boldsymbol{\alpha}}$, q_1, q_2)=1−$\hat{E}_{1,i}$. Moreover, $\boldsymbol{\psi}_{\mathbf{r}}$ ($\mathbf{W}_i$,$\mathbf{r}$($V_{1,i}$,Y_i)^T) is a vector of functions of the variables that are observed when $\mathbf{R}_i$ = $\mathbf{r}$.

5.3. General monotone results

The results we introduced above for example 2 can be extended to multiple-occasion monotone missing data models. In such models, $\mathbf{V}_i$ consists $q \geq 2$ elements and $\mathbf{R}_i$ indicates the corresponding vector of missing indicators. If the s th component ($1 \leq s \leq q$) $V_{s,i}$ is missing ($R_{s,i} = 0$), all subsequent components of $\mathbf{V}_i$ are missing ($R_{t,i} = 0$ for any $s < t \leq q$). Let $\mathbf{r}_s\equiv\left(\mathbf{1}_{q-s}^T,\mathbf{0}_s^T\right)^T$ indicate a q-dimensional vector with the first $q - s$ elements being 1 and the remaining s elements being 0 (i.e., the first $q - s$ elements of $\mathbf{V}_i$ are observed while the remaining s elements are missing). The class of IPW estimators $\hat{\boldsymbol{\beta}}$ is constructed based on the estimating equations $\sum_{i=1}^{n}\mathbf{D}_i\left(\boldsymbol{\beta},\boldsymbol{\varphi}_1,\ldots,\boldsymbol{\varphi}_q\right)$ where

$$\begin{aligned}&\mathbf{D}_i\left(\boldsymbol{\beta},\boldsymbol{\varphi}_1,\ldots,\boldsymbol{\varphi}_q\right)\\ &\quad=\frac{I(\mathbf{R}_i=\mathbf{1})}{\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1})}\mathbf{M}(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta})\\ &\quad+\sum_{s=1}^{q}\left\{I(\mathbf{R}_i=\mathbf{r}_s)-\frac{I(\mathbf{R}_i=\mathbf{1})}{\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1})}\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{r}_s)\right\}\boldsymbol{\varphi}_s\left(\mathbf{W}_i,\mathbf{V}_{(\mathbf{r}_s),i}\right),\end{aligned}$$

where $\boldsymbol{\varphi}_s$ ($\mathbf{W}_i$,$\mathbf{V}_{(r_s)i}$) is a vector of selected functions of the variables $\mathbf{W}_i$ and ($V_{1,i}$, …,$V_{q-s,i}$)^T, which are observed when $\mathbf{R}_i$ = $\mathbf{r}_s$. For any $1 \leq s \leq q$, let $E_{s,i} \equiv P(R_{s,i}=1|R_{s-1,i}=1, \mathbf{W}_i, \mathbf{V}_i)$ denote subject i's conditional probability of observing the sth element $V_{s,i}$ given the full data ($\mathbf{W}_i$,$\mathbf{V}_i$) and the event that all previous elements ($V_{1,i}$,…,$V_{s-1,i}$) are observed. Due to monotone missingness, $\pi_i(\mathbf{W}_i,\mathbf{V}_i,\mathbf{r}_s)=\prod_{t=1}^{q-s}E_{t,i}(1-E_{q-s+1,i})$.

Under MAR, $E_{s,i}$ depends on $(\mathbf{W}_i,V_{1,i},\ldots,V_{s-1,i})$ only, i.e., $P(R_{s,i}=1|R_{s-1,i}=1, \mathbf{W}_i,\mathbf{V}_i)=P(R_{s,i}=1|R_{s-1,i}=1, \mathbf{W}_i,V_{1,i},\ldots,V_{s-1,i})$. Then $E_{s,i}$ can be estimated from the observed data $\{R_{s,i}, \mathbf{W}_i,V_{1,i},\ldots,V_{s-1,i}: i=1,\ldots,n$ and $R_{s-1,i}=1\}$ by regressing $R_{s,i}$ on $(\mathbf{W}_i,V_{1,i},\ldots,V_{s-1,i})$ among those with $R_{s-1,i}=1$.

The estimation of the missing data process under MNAR is much more complicated. As before, selection bias functions need to be specified for the “odds” of having missing data. Specifically, for any $1 \le s \le q$,

$$\frac{1-E_{s,i}}{E_{s,i}}=\frac{P(R_{s,i}=0|R_{s-1,i}=1,\mathbf{W}_i,\mathbf{V}_i)}{P(R_{s,i}=1|R_{s-1,i}=1,\mathbf{W}_i,\mathbf{V}_i)} =\exp\left(h_s\left(\mathbf{W}_i, V_{1,i},\ldots,V_{s-1,i},\boldsymbol{\alpha}\right)+q_s\left(\mathbf{W}_i,\mathbf{V}_i\right)\right)$$

Then, $\pi_i\left(\mathbf{W}_i,\mathbf{V}_i,\mathbf{r}_s;\boldsymbol{\alpha}\right)=\prod_{t=1}^{q-s}E_{t,i}\times\left(1-E_{q-s+1,i}\right)$ and $\pi_i\left(\mathbf{W}_i,\mathbf{V}_i,\mathbf{r}_s;\boldsymbol{\alpha}\right)=\prod_{t=1}^{q-s}E_{t,i}$. The estimated $\hat{\boldsymbol{\alpha}}$ solves the estimating equation $\sum_{i=1}^{n}\mathbf{A}_i\left(\boldsymbol{\psi},\alpha\right)=0$ where

$$A_i\left(\boldsymbol{\psi},\boldsymbol{\alpha}\right)=\sum_{s=1}^{q}\left\{I\left(\mathbf{R}_i=\mathbf{r}_s\right)-\frac{I\left(\mathbf{R}_i=\mathbf{1}\right)}{\pi_i\left(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1};\boldsymbol{\alpha}\right)}\pi_i\left(\mathbf{W}_i,\mathbf{V}_i,\mathbf{r}_s;\boldsymbol{\alpha}\right)\right\}\boldsymbol{\psi}_s\left(\mathbf{W}_i,V_{1,i},\ldots,V_{q-s,i}\right),$$

and $\boldsymbol{\psi}_s(\mathbf{W}_i,V_{1,i},\ldots,V_{q-s,i})$ is a vector of functions of $(\mathbf{W}_i,V_{1,i},\ldots,V_{q-s,i})$.

6. Non-monotone missing pattern

In non-monotone missing data models, the q-dimensional vector of missing indicators $\mathbf{R}_i$ can take 2^q possible values as each element can be either 0 or 1. For example, when $q = 2$, $\mathbf{R}_i = \mathbf{r} \in \{(0,0)^T,(0,1)^T,(1,0)^T,(1,1)^T\}$. In such models, the estimation of the missing data process is substantially more challenging.

The estimation of the parameter of interest $\boldsymbol{\beta}$ when the missing probabilities $\{\boldsymbol{\pi}_i(\mathbf{W}_i, \mathbf{V}_i,\mathbf{r}): \mathbf{r}\}$ are known is similar to the estimation in monotone missing data models. Specifically, the IPW estimator $\hat{\boldsymbol{\beta}}$ is obtained by solving the estimating equation $\sum_{i=1}^{n}\mathbf{D}_i\left(\boldsymbol{\beta},\{\boldsymbol{\varphi}_\mathbf{r}\}\right)$ where

$$\mathbf{D}_i\left(\boldsymbol{\beta},\{\boldsymbol{\varphi}_\mathbf{r}\}\right)=\frac{I\left(\mathbf{R}_i=\mathbf{1}\right)}{\pi_i\left(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1}\right)}\mathbf{M}\left(\mathbf{W}_i,\mathbf{V}_i;\boldsymbol{\beta}\right)+\sum_{\mathbf{r}\neq\mathbf{1}}\left\{I\left(\mathbf{R}_i=\mathbf{r}\right)-\frac{I\left(\mathbf{R}_i=\mathbf{1}\right)}{\pi_i\left(\mathbf{W}_i,\mathbf{V}_i,\mathbf{1}\right)}\pi_i\left(\mathbf{W}_i,\mathbf{V}_i,\mathbf{r}\right)\right\}\boldsymbol{\varphi}_\mathbf{r}\left(\mathbf{W}_i,\mathbf{V}_{(\mathbf{r}),i}\right),$$

and $\boldsymbol{\varphi}_\mathbf{r}(\mathbf{W}_i, \mathbf{V}_{(\mathbf{r}),i})$ is a selected $m \times 1$ vector of functions of the observed components $(\mathbf{W}_i, \mathbf{V}_{(\mathbf{r}),i})$ when $\mathbf{R}_i = \mathbf{r}$. Unlike in Section 5.3, $\mathbf{r}$ is no longer restricted to $\left\{\left(\mathbf{1}_{q-s}^T,\mathbf{0}_s^T\right)^T:1\le s\le q\right\}$.

In most applications, π_i ($\mathbf{W}_i$, $\mathbf{V}_i$, $\mathbf{r}$) is unknown and must be estimated from the observed data. Robins and colleagues proposed the randomized monotone missingness (RMM) processes[41] to analyze non-monotone ignorable missing data, and the selection bias permutation missingness (PM) models[42,43] to analyze non-monotone nonignorable missing data. These approaches are sometimes plausible. However, they are quite complex and computationally intensive. There currently exists no user-friendly software program to facilitate their implementation. These limitations likely contribute to lack of wide adoption. Through introducing the heuristic ideas behind these approaches, we hope to encourage researchers to develop user-friendly software tools for these methods.

We use two motivating examples for MAR and MNAR mechanisms respectively; PM models are best explained in the context of a longitudinal study. In contrast, RMM models do not apply to longitudinal data. Both examples share common notation. The full data is denoted by $\mathbf{L}_i = \{(\mathbf{W}_i^T, V_{1,i}, V_{2,i}, V_{3,i})^T, i = 1, \ldots, n\}$, and the observed data is denoted by $\{\mathbf{O}_i = (\mathbf{W}_i^T, R_{1,i}, R_{2,i}, R_{3,i}, R_{1,i}V_{1,i}, R_{2,i}V_{2,i}, R_{3,i}V_{3,i})^T, i = 1, \ldots, n\}$ where $\mathbf{R}_i = (R_{1,i}, R_{2,i}, R_{3,i})^T$ is the vector of missing indicators. The parameter of interest $\boldsymbol{\beta}^*$ is the unique solution to $E[\mathbf{M}(\mathbf{W}, \mathbf{V}; \boldsymbol{\beta}^*)] = 0$

6.1. MAR

We consider example 3. Under MAR, for any $\mathbf{r} = (r_1, r_2, r_3)^T$,

$$\pi_i\left(\mathbf{W}_i, \mathbf{V}_i, \mathbf{r}\right) = P\left(\mathbf{R}_i = \mathbf{r} | \mathbf{W}_i, \mathbf{V}_i\right) = P\left(\mathbf{R}_i = \mathbf{r} | \mathbf{W}_i, r_1 V_{1,i}, r_2 V_{2,i}, r_3 V_{3,i}\right).$$

If ($\mathbf{W}_i$, $\mathbf{V}_i$) is discrete with few levels, the estimated missing probabilities $\hat{\pi}_i$ ($\mathbf{W}_i$, $\mathbf{V}_i$, $\mathbf{r}$) can be obtained as the empirical proportions within each covariate level. In practice, we need to impose parametric working models for π_i ($\mathbf{W}_i$, $\mathbf{V}_i$, $\mathbf{r}$) to reduce dimension and borrow information across different covariate levels. To simultaneously satisfy the restrictions imposed by MAR, the inequalities $0 \le \pi_i(\mathbf{r}) \le 1$, and the equality $\sum_{\mathbf{r}} \pi_i\left(\mathbf{r}\right) = 1$, it will be difficult, if not impossible, to directly model $\{\pi_i(\mathbf{W}_i, \mathbf{V}_i, \mathbf{r}): \mathbf{r}\}$.

Robins & Gill [41] proposed an algorithm to estimate π_i ($\mathbf{W}_i$, $\mathbf{V}_i$, $\mathbf{r}$) under a sub-model of MAR models, which they referred to as a RMM model. This model is assumed to be generated as follows. For each subject i, $\mathbf{W}_i$ is observed. Then one of the three elements of $\mathbf{V}_i$, $V_{s,i}$, $1 \le s \le 3$ is observed with probability $p_s = p_s(\mathbf{W}_i)$, or one quits without observing any element of $\mathbf{V}_i$ with probability $q = 1 - \sum_{s=1}^{3} p_s$. If, for example, $V_{1,i}$ is observed, then in a second step, we observe $V_{2,i}$ with a conditional probability $p_{12}(V_{1,i})$, or observe $V_{3,i}$ with a conditional probability $p_{13}(V_{1,i})$, or quit with probability $1 - p_{12}(V_{1,i}) - p_{13}(V_{1,i})$. Note that the conditional probabilities $p_{12}(V_{1,i})$ and $p_{13}(V_{1,i})$ depend both on $\mathbf{W}_i$ and the value of $V_{1,i}$ observed at the first step. For simplicity, we suppress the dependence on $\mathbf{W}_i$ when no ambiguity arises. Suppose $V_{2,i}$ is observed at the second step, then in the third step, we observe the third component $V_{3,i}$ with a conditional probability $p_{123}(V_{1,i}, V_{2,i})$ or quit with probability 1 − p_{123} ($V_{1,i}$, $V_{2,i}$). The following figure is similar to Figure 1 in Robins & Gill[41] to help understanding.

An RMM process satisfies MAR. For example, the overall probability of observing ($V_{1,i}$, $V_{2,i}$), $\boldsymbol{\pi}_i$ ($\mathbf{r} = (1,1,0)^T$), equals p_1 p_{12} ($V_{1,i}$)(1 − p_{123} ($V_{1,i}$, $V_{2,i}$)) + p_2 p_{21} ($V_{2,i}$)(1 − p_{213} ($V_{2,i}$, $V_{1,i}$)), since we either observe $V_{1,i}$ at the first step and then $V_{2,i}$ at the second step and then quit without observing $V_{3,i}$, or observe $V_{2,i}$ at the first step and then $V_{1,i}$ at the second step and then quit without observing $V_{3,i}$. This overall probability depends on ($\mathbf{W}_i$, $V_{1,i}$, $V_{2,i}$) which are observed when $\mathbf{R}_i = (1,1,0)^T$. It can be shown that the probabilities sum to 1.

Gill & Robins[44] showed that there do exist ignorable (i.e., MAR) missing data processes that are not RMM. However, such processes are often unrealistic “due to the subtle and precise manner in which the data must be ‘hidden’ to insure that the process is MAR”.

The estimation of $\boldsymbol{\pi}_i$ ($\mathbf{W}_i$, $\mathbf{V}_i$, $\mathbf{r}$) is non-trivial for RMM processes. To reduce the dimension, the authors considered Markov RMM processes in which the conditional probabilities do not depend on the order in which the variables were observed. For example, p_{123} ($V_{1,i}$, $V_{2,i}$) = p_{213} ($V_{1,i}$, $V_{2,i}$) and will be denoted as $p_3^{12}\left(V_{1,i},V_{2,i}\right)$. Parametric working models are imposed for these conditional probabilities. For example, for any $k \in \{1,2,3\}$, we model the first-step probabilities with a multinomial logistic regression model

$$p_k=\rho_k/\left(1+\sum_{k=1}^{3}\rho_k\right)\ \text{ where }\ \rho_k=\rho_k\left(\mathbf{W}_i\right)=\exp\left[\gamma_{0,k}+\boldsymbol{\gamma}_{1,k}^T\mathbf{W}_i\right].$$

The second step probabilities are modeled by

$$\begin{aligned}p_{kl}\left(V_{k,i}\right)=\rho_{kl}\left(V_{k,i}\right)/\left(1+\sum_{l\neq k}\rho_{kl}\left(V_{k,i}\right)\right)\ \text{ for }\ l\neq k,\ \text{where}\\ \rho_{kl}\left(V_{k,i}\right)=\exp\left[\gamma_{0,kl}+\boldsymbol{\gamma}_{1,kl}^T\mathbf{W}_i+\gamma_{2,kl}V_{k,i}\right].\end{aligned}$$

Finally, the third step probabilities are modeled by,

$$\log it\left[p_k^{\{1,2,3\}\backslash k}\left(\mathbf{V}_{(-k),i}\right)\right]=\zeta_{0,k}+\boldsymbol{\zeta}_{1,k}^T\mathbf{W}_i+\boldsymbol{\zeta}_{2,k}^T\mathbf{V}_{(-k),i},k\in\{1,2,3\},$$

where $\mathbf{V}_{(-k),i}$ indicates the two elements other than $V_{k,i}$ (e.g., $\mathbf{V}_{(-1),i} = (V_{2,i}, V_{3,i})^T$). When appropriate, we can further decrease the dimension of the parameter space by assuming, for example, ($\gamma_{0,k}$, $\boldsymbol{\gamma}_{1,k}^T$) does not depend on k.

The maximum likelihood estimates (MLEs) of the unknown parameters cannot be directly obtained as the order in which variables were observed is missing. For example, there are two paths in the figure above by which $V_{1,i}$ and $V_{2,i}$ could be observed: $V_{1,i}$ − $V_{2,i}$ − *quit*, or

$V_{2,i} - V_{1,i} - quit$. The authors suggest treating the path information as missing and to obtain the MLE with the Expectation-Maximization (EM) algorithm. See [41] for details.

6.2. MNAR

For non-monotone nonignorable missing data processes, Robins et al.[43] propose selection bias PM models. Consider our motivating example 4, a longitudinal study with three BP measurements. In longitudinal studies, the PM order is the reverse of the temporal order. Under a PM model, we assume that the conditional probability of observing $V_{s,i}$ at the *s*th visit depends (i) on the observed components from previous visits (i.e., $\mathbf{L}_{s,i} \equiv (\mathbf{W}_i, R_{1,i}, \ldots R_{s-1,i}, R_{1,i}V_{1,i}, \ldots, R_{s-1,i}V_{s-1,i})$) but not on the unobserved components of $(V_{1,i}, \ldots, V_{s-1,i})$; (ii) on the value of $V_{s,i}$ through a specified selection bias function; and (iii) on both observed and unobserved components in future visits ($(V_{s+1,i}, \ldots, V_{q,i})$). In our motivating example 4, we consider a simplified PM model in which the conditional probability of observing $V_{s,i}$ does not depend on any future data. Thus,

$$\begin{gathered}\pi_i\left(\mathbf{W}_i, \mathbf{V}_i, \mathbf{r}\right)=P\left(\mathbf{R}_i=\mathbf{r} | \mathbf{W}_i, \mathbf{V}_i\right) \text{ is } \prod_{s=1}^{3} E_{s,i}\left(r_s\right), \text{ where} \\ E_{s,i}\left(r_s\right) \equiv P\left(R_{s,i}=r_s | R_{1,i}=r_1, \ldots, R_{s-1,i}=r_{s-1}, \mathbf{W}_i, \mathbf{V}_i\right) \text{ satisfies} \\ E_{s,i}\left(1\right)=P\left(R_{s,i}=1 | R_{1,i}, \ldots, R_{s-1,i}, R_{1,i}V_{1,i}, \ldots, R_{s-1,i}V_{s-1,i}, V_{s,i}\right) \\ =\text{exp}it\left\{h_s\left(\mathbf{L}_{s,i}\right)+q_s\left(V_{s,i}, \mathbf{L}_{s,i}\right)\right\}\end{gathered} \tag{6}$$

Here q_s $(V_{s,i}, \mathbf{L}_{s,i})$ is an investigator specified selection bias function and $h_s(\mathbf{L}_{s,i})$ is an unrestricted function to be estimated. By eq. (6), the conditional probability $E_{s,i}$ (r_s) depends on the possibly unobserved value of $V_{s,i}$ through q_s $(V_{s,i}, \mathbf{L}_{s,i})$.

In most applications, we impose parametric working models h_s $(\mathbf{L}_{s,i}; \boldsymbol{\delta}_s)$ for h_s $(\mathbf{L}_{s,i})$ to overcome the curse of dimensionality. The parameter $\boldsymbol{\delta}_s$ can be estimated by solving

$$\sum_{i=1}^{n}\left\{\frac{R_{s,i}}{\text{exp}it\left[h_s\left(\mathbf{L}_{s,i};\boldsymbol{\delta}_s\right)+q_s\left(\mathbf{L}_{s,i}, V_{s,i}\right)\right]}-1\right\}\boldsymbol{\varphi}_s\left(\mathbf{W}_i\right)=0 \tag{7}$$

where $\boldsymbol{\varphi}_s$ $(\mathbf{W}_i)$ is a vector of selected known functions of $\mathbf{W}_i$ and has the same dimension as $\boldsymbol{\delta}_s$. See Vansteelandt et al.[30] for an extension of this approach to estimate the mean vector of repeated outcomes in a nonignorable, non-monotone missing data model.

Although a subject's decision to miss the *s*th visit cannot directly depend on future data. But R_s, the indicator variable indicating whether V_s was observed, might be statistically associated with future data, when some factors that affect the decision are not recorded in $(\mathbf{L}_s, V_s)$ but are associated with $(V_{s+1}, \ldots, V_q)$. See Robins, Rotnitzky, and Scharfstein[43] for further discussions.

7. Discussion

We have introduced the IPW approaches in a wide range of settings with different missing data patterns and mechanisms. These weighting approaches share the same basic idea. However, different strategies are needed to estimate the missing probabilities depending on the missing data pattern and mechanism. Our goal in this review paper was to provide a conceptual overview of existing weighting approaches.

Our review began with a simple uniform missing data model; for each subject *i*, either the entire vector $\mathbf{V}_i$ is observed or it is completely missing. We then discussed monotone missing data patterns. We show these models can be decomposed into multiple “artificial” uniform missing data models and estimators are obtained by applying weighting approaches for uniform missing data models in a nested fashion. In Section 6, we discussed non-monotone missing patterns and notice the estimation of the missingness probabilities is substantially more challenging and complex. We then introduced the RMM processes for non-monotone MAR data and the selection bias PM approach for non-monotone MNAR data. User-friendly software programs need to be developed to make these methods useful for practice.

We considered both MAR and MNAR mechanisms. IPW estimators for MNAR are natural extensions of IPW estimators for MAR in which selection bias functions quantify the residual association of the missing probabilities and unobserved data conditional on observed data. The MAR assumption cannot be empirically tested when the model of the full data is nonparametric. Subject matter expertise and prior information are typically required to judge its plausibility. In uniform and monotone missing patterns, MAR sometimes is reasonable if data on a large set of variables are collected. The MAR assumption is less likely to hold with non-monotone missingness.[30] Unless strong prior information is available, we recommend analysts consider the possibility that the missingness mechanism is nonignorable and conduct a sensitivity analysis.

References

1. Little, R.; Rubin, D. Statistical Analysis with Missing Data. New York: John Wiley & Sons; 1987.
2. Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. Annual Review of Public Health. 2004; 25:99–117.
3. Rubin D. Inference and missing data (with discussion). Biometrika. 1976; 63:581–592.
4. Horvitz DG, Thompson DJ. A Generalization of Sampling Without Replacement from A Finite Universe. Journal of the American Statistical Association. 1952; 47:663–685.
5. Robins J, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association. 1994; 89:846–866.
6. Robins J, Rotnitzky A, Zhao L. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association. 1995; 90:106–121.
7. Robins J, Hernan M, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000; 11:550–560. [PubMed: 10955408]
8. Hernan M, Brumback B, Robins J. Marginal structural models to estimate the causal effect of Zidovudine on the survival of HIV-positive men. Epidemiology. 2000; 11:561–570. [PubMed: 10955409]
9. Horton NJ, Laird NM. Maximum likelihood analysis of generalized linear models with missing covariates. Statistical Methods in Medical Research. 1999; 8:37–50. [PubMed: 10347859]
10. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: A comparative review. Journal of the American Statistical Association. 2005; 100:332–346.

11. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test. 2009; 18:1–43. [PubMed: 21218187]
12. Ibrahim JG, Chen MH. Power prior distributions for regression models. Statistical Science. 2000; 15:46–60.
13. Chen MH, Ibrahim JG, Lipsitz SR. Bayesian methods for missing covariates in cure rate models. Lifetime Data Analysis. 2002; 8:117–146. [PubMed: 12048863]
14. Ibrahim JG, Chen MH, Lipsitz SR. Bayesian methods for generalized linear models with covariates missing at random. Canadian Journal of Statistics-Revue Canadienne de Statistique. 2002; 30:55–78.
15. Harel O, Zhou XH. Multiple imputation: Review of theory, implementation and software. Statistics in Medicine. 2007; 26:3057–3077. [PubMed: 17256804]
16. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.
17. Schafer JL. Multiple imputation: a primer. Statistical Methods in Medical Research. 1999; 8:3–15. [PubMed: 10347857]
18. Tsiatis, A. Semiparametric theory and missing data. New York: Springer; 2006.
19. van der Laan, M.; Robins, J. Unified methods for censored longitudinal data and causality. New York: Springer; 2003.
20. Robins JM, Rotnitzky A. Semiparametric Efficiency in Multivariate Regression-Models with Missing Data. Journal of the American Statistical Association. 1995; 90:122–129.
21. Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. Journal of the American Statistical Association. 1998; 93:1321–1339.
22. Robins JM, Rotnitzky A, Zhao LP. Analysis of Semiparametric Regression-Models for Repeated Outcomes in the Presence of Missing Data. Journal of the American Statistical Association. 1995; 90:106–121.
23. Rotnitzky A, Robins JM. Semiparametric regression estimation in the presence of dependent censoring. Biometrika. 1995; 82:805–820.
24. Rotnitzky A, Robins JM. Semiparametric Estimation of Models for Means and Covariances in the Presence of Missing Data. Scandinavian Journal of Statistics. 1995; 22:323–333.
25. Rotnitzky A, Holcroft CA, Robins JM. Efficiency comparisons in multivariate multiple regression with missing outcomes. Journal of Multivariate Analysis. 1997; 61:102–128.
26. Bickel, PJ.; Klaassen, CA.; Ritov, Y.; Wellner, JA. Efficient and adaptive estimation for semiparametric models. New York: Springer Verlag; 1998.
27. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association. 1999; 94:1096–1120.
28. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models - Rejoinder. Journal of the American Statistical Association. 1999; 94:1135–1146.
29. Robins JM, Rotnitzky A. Inference for semiparametric models: Some questions and an answer - Comments. Statistica Sinica. 2001; 11:920–936.
30. Vansteelandt S, Rotnitzky A, Robins J. Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. Biometrika. 2007; 94:841–860.
31. Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and regression trees. Belmont, CA: Wadsworth International Group; 1984.
32. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting. Annals of Statistics. 2000; 28:337–374.
33. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting - Rejoinder. Annals of Statistics. 2000; 28:400–407.
34. Breiman L. Random forests. Machine Learning. 2001; 45:5–32.
35. Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning: data mining, inference, and prediction. 2. New York: Springer; 2009.

36. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data. Epidemiology. 2009; 20:512–522. [PubMed: 19487948]

37. Therneau, TM.; Atkinsoon, EJ. An introduction to recursive partitioning using the RPART routines. 1997.

38. Bang H, Robins J. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005; 61:962–972. [PubMed: 16401269]

39. Potthoff RF, Tudor GE, Pieper KS, Hasselblad V. Can one assess whether missing data are missing at random in medical studies? Statistical Methods in Medical Research. 2006; 15:213–234. [PubMed: 16768297]

40. Rotnitzky A, Robins J. Analysis of semi-parametric regression models with non-ignorable non-response. Statistics in Medicine. 1997; 16:81–102. [PubMed: 9004385]

41. Robins JM, Gill RD. Non-response models for the analysis of non-monotone ignorable missing data. Statistics in Medicine. 1997; 16:39–56. [PubMed: 9004382]

42. Robins JM. Non-response models for the analysis of non-monotone non-ignorable missing data. Statistics in Medicine. 1997; 16:21–37. [PubMed: 9004381]

43. Robins, JM.; Rotnitzky, A.; Scharfstein, D. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran, M.; Berry, D., editors. Statistical Models in Epidemiology: The Environment and Clinical Trials. New York: Springer-Verlag; 1999. p. 1-92.

44. Gill, RD.; van der Laan, M.; Robins, JM. Coarsening at random: characterizations, conjectures and counterexamples. In: Lin, DY., editor. Proceedings of the First Seattle Symposium on Biostatistics: Survival Analysis. New York: Springer Verlag; 1997. p. 255-94.

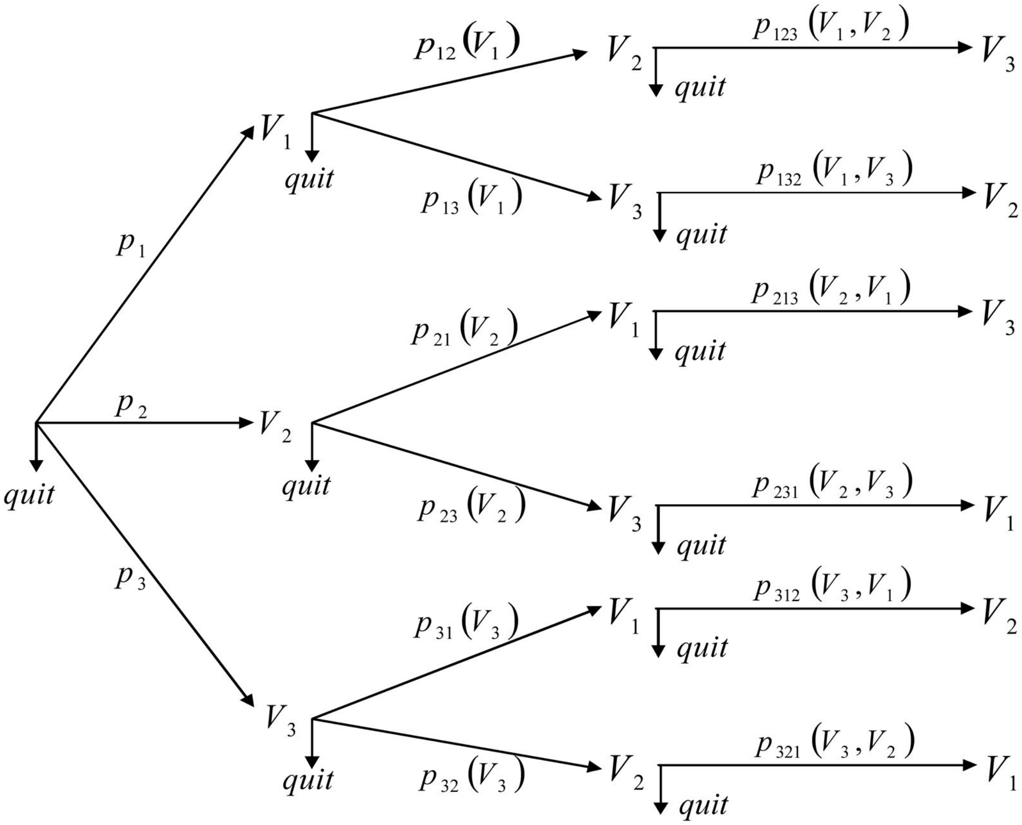

Figure 1.

Missing data process in a RMM process