Harvard University
Harvard University Biostatistics Working Paper Series

Year 2014 *Paper* 185

Constrained Bayesian Estimation of Inverse Probability Weights for Nonmonotone Missing Data

BaoLuo Sun[*] Eric J. Tchetgen Tchetgen[†]

[*] Harvard School of Public Health, bluosun@gmail.com

[†] Harvard School of Public Health, etchetgen@gmail.com

Constrained Bayesian Estimation of Inverse Probability Weights for Nonmonotone Missing Data

BaoLuo Sun
Department of Biostatistics, Harvard School of Public Health
and
Eric J. Tchetgen Tchetgen
Departments of Biostatistics & Epidemiology, Harvard School of Public Health

November 18, 2014

Abstract

In the analysis of non-monotone missing at random (MAR) data using inverse probability weighting, a straightforward approach towards modeling the missingness mechanism based on simple polytomous logistic regression often imposes more restrictive conditions than what MAR entails. We propose a class of models for the non-monotone missingness mechanism that spans and accommodates the entire MAR model, and the estimation procedure can be easily implemented within this class using existing software. Where unconstrained maximum likelihood (ML) does not converge, we propose a Bayesian estimation of the missing data process which is guaranteed to yield inferences within the model. We illustrate the new methodology in an application evaluating the association between maternal HIV infection and adverse birth outcomes in Botswana.

Keywords: Missing at random; Inverse probability weighting; Propensity score

1 INTRODUCTION

Missing data is a major complication which occurs frequently in empirical research. Nonresponse in sample surveys, dropout or non-compliance in clinical trials and data excision by error or to protect confidentiality are but a few examples of ways in which full data is unavailable and our ability to make accurate inferences may be compromised. Missingness could also be introduced into a study by design, e.g. multi-stage sampling plans in order to reduce the cost associated with measurements for all subjects. In many practical situations, the missing data pattern is non-monotone, that is, there is no nested pattern of missingness such that observing variable X_k implies that variable X_j is also observed, for any $j < k$. Non-monotone missing data patterns may occur, for instance, when individuals who dropped out of a longitudinal study re-enter at later time points. The missing data process is said to be missing-completely-at-random (MCAR) if it is independent of both observed and unobserved variables in the full data, and missing-at-random (MAR) if, conditional on the observed variables, the process is independent of the unobserved ones (Little and Rubin 2002). A missing data process which is neither MCAR nor MAR is said to be missing-not-at-random (MNAR).

While complete-case analysis is the easiest to implement and often employed in practice, the method is generally known to produce biased estimates when the missingness mechanism is not MCAR (Little and Rubin 2002). Efficiency is also lost by discarding samples with incomplete data. Other commonly used procedures include last-observation-carried forward analysis in longitudinal studies and simple imputation techniques, but they typically produce valid inferences only under restrictive and often unrealistic conditions (Molengerghs et al. 2004; Siddiqui and Ali 1998; Little and Rubin 2002). In addition, it may be difficult to account for the variance induced by simply filling in missing values and analyzing the resulting data using procedures originally meant for fully observed data, without explicitly defining a data generating mechanism for the filled in values.

The development of principled methods to appropriately account for missing data has been an area of active and on-going research. The assumptions of MAR or MCAR, together with separability of parameters governing the missingness mechanism and complete data model, provide sufficient conditions for valid inferences based on the observed data likelihood (Little and Rubin 2002). Similar analysis could also be performed within the Bayesian framework by introducing priors for model parameters. While likelihood-based methods generally require specification of the full-data likelihood, the estimators are efficient under the assumed parametric restrictions. Estimation is easier in monotone missing data patterns where the likelihood can be factored naturally into a series of conditional densities. For non-monotone missing data patterns, except under certain special conditions (Rubin 1974), estimation usually requires some iterative procedures such as the expectation-maximization (EM) algorithm (Dempster et al. 1977).

Multiple Imputation (MI) has become an influential technique to account for missing data since its introduction in the survey analysis setting (Rubin 1977), and is widely utilized through its incorporation into mainstream statistical software (Horton and Lipsitz 2001). MI requires an imputation model, typically a regression model, that relates the distribution of the missing data to the observed data. MI is particularly well suited for monotone missing data patterns where missing values can be imputed sequentially under a series of conditional imputation models. For arbitrary missing data patterns, it is often necessary to specify a multivariate distribution of all variables for joint imputation in order to ensure model congeniality, such as the multivariate normal imputation distribution (Schafer 1997). An alternative strategy involving multiple imputation using chained equations (MICE) (van Buuren and Oudshoorn 2000) has become popular due to its simplicity and applicability under a wide variety of settings with different variable types and missing data patterns. Under MICE, MI is essentially accomplished by specifying a series of univariate conditional imputation models. It is possible, however, that the conditional models are incompatible and a joint distribution may not exist (White et al. 2011; van Buuren 2007). At the same time, since in non-monotone missing data the sets of observed and unobserved variables changes among individuals often in an arbitrary manner, it is not clear how conditional densities at the single variable level can be made to depend only on observed variables, which is assumed under MAR. In actual implementations, the issue of MAR is often skirted, although it is a key assumption to ensure validity of multiple imputation (Kenward and Carpenter 2007).

Another popular approach involves weighting the complete cases by the inverse probability of being a complete case, also known as the propensity score (Horvitz and Thompson 1952; Little and Rubin 2002; Robins et al. 1994; van der Laan and Robins 2003; Tsiatis 2006), which creates a pseudo-population of complete cases in which selection bias due to missing data is removed. IPW estimation does not require specification of the full-data likelihood, but the missingness mechanism needs to be modeled. In the case of a monotone missing data pattern, the missingness probability can be naturally factored into a sequence of conditional binomials due to the nested missing data patterns, under the MAR assumption. The missingness mechanism can then be modelled by a corresponding sequence of binary regressions using say logistic or probit regression. This approach does not generally work for non-monotone missing data patterns, and, as we will discuss later in the paper, a straightforward polytomous logistic regression for the collection of possible missing data patterns will impose more restrictive conditions than what MAR strictly entails. The development of coherent models and practical estimation procedures for the missingness probabilities of nonmonotone missing data is challenging, even under the assumption that the data is MAR. To the best of our knowledge, and as discussed in the seminal missing data book of Tsiatis (2006, p. 188), there currently is not available, a general approach to model an arbitrary nonmonotone missing data generating process only imposing MAR. This is an important gap in the missing data literature, which has essentially restricted the use of inverse probability weighted estimation to monotone missing data settings.

As a remedy, Robins and Gill proposed a large class of models for the missing data mechanism, the randomised monotone missingness (RMM) processes, which are guaranteed to be MAR for a non-monotone missing data mechanism without being MCAR (Robins and Gill 1997). This class of models does not span the space of all MAR models and therefore it is possible to test whether the class includes the true missing data mechanism. However, estimation of the missing data mechanism within this class is complex and computationally demanding, even for small to moderate sample size and number of different missing data patterns, and there is currently no available software to implement the approach, which has limited its widespread adoption. In this paper we take a different direction, and we propose a class of models for the non-monotone missing data mechanism that spans the entire MAR model and therefore, with enough data such that non-parametric models can be used reliably, in principle one would not be able to reject MAR based on the oberved data.

In order to estimate the missingness mechanism required for IPW estimation, we discuss two approaches: unconstrained maximum likelihood estimation and constrained Bayesian estimation. The first approach is easily implemented in standard software, say using existing procedures in SAS or R. However, despite this appealing feature, as we illustrate in extensive simulation studies, unconstrained maximum likelihood estimation has a major drawback, in that it is not guaranteed to converge in finite sample, even if all regression models are correctly specified. This problematic feature of the approach is mainly due to the fact that it fails to impose certain natural restrictions of the model. This drawback was previously noted by Robins and colleagues (Robins et al. 1999), who upon pointing out this potential difficulty, abandoned the approach. In this paper, we propose a novel constrained Bayesian strategy (Gelfand et al. 1992) as a viable alternative to unconstrained maximum likelihood estimation, which largely resolves any convergence difficulty and is easily implemented in standard Bayesian software packages. Constrained Bayesian estimation has been used previously to estimate risk ratio and relative excess risk regressions (Chu and Cole 2010, 2011); however, to the best of our knowledge, it has not previously been used in the current context. We present a simulation study to investigate the finite-sample properties of both constrained and unconstrained inferences in the context of IPW of logistic regression with non-monotone missing outcome and covariates, followed by an analysis of adverse birth outcomes on a cohort of women in Botswana to illustrate an application of the methods.

2 NOTATION AND ASSUMPTIONS

Let $\boldsymbol{L} = (L_1, ..., L_K)'$ be a random K-vector representing the complete data. In addition, let $\boldsymbol{R} = m$ be the scalar random variable encoding the different missing data patterns where $1 \leq m \leq 2^K$. For individuals with $\boldsymbol{R} = m$, we observe $\boldsymbol{L}_{(m)}$ where $\boldsymbol{L}_{(m)} \subseteq \boldsymbol{L}$. For example, suppose the complete data for each person consists of two random variables $L = (L_1, L_2)$. Then we may encode the missing data patterns as follows: $\boldsymbol{R} = 1$ if we observe $\boldsymbol{L}_{(1)} = \boldsymbol{L} = (L_1, L_2)$; $\boldsymbol{R} = 2$ if we observe $\boldsymbol{L}_{(2)} = L_1$; $\boldsymbol{R} = 3$ if we observe

$\boldsymbol{L}_{(3)} = L_2$; and $\boldsymbol{R} = 4$ if neither variable is observed. For non-parametric identification of the missingness parameters, we assume that the missing data process is MAR (Robins et al. 1994).

$$\Pr\{\boldsymbol{R} = m|\boldsymbol{L}\} = \pi_m(\boldsymbol{L}_{(m)}) \quad \text{for} \quad m = 1, ..., M \tag{1}$$

If we were to relax the above assumption by considering MNAR processes, we would lose non-parametric identification, and we would require an additional assumption to successfully identify parameters indexing the missing data mechanism (Robins et al. 1999; Robins and Rotnitzky 1997). In addition, we assume that the probability of being a complete case is bounded away from zero, a necessary assumption for identification of a full data functional (Robins et al. 1994).

$$\pi_1(\boldsymbol{L}) > \sigma > 0 \quad \text{with probability 1} \tag{2}$$

for a non-zero positive constant $\sigma > 0$. A key implication of the MAR assumption is that the missing data process is itself nonparametrically identified under the assumption, without imposing any restriction on the full data distribution of $\boldsymbol{L}$. This also implies that provided on using separate parameters to index the missing data mechanism and the full data distribution, efficient estimation of the parameters of the missing data process can be obtained by maximizing its partial likelihood, ignoring the part of the likelihood corresponding to the full data likelihood.

3 IPW INFERENCE

Suppose we observe n i.i.d. vector $\boldsymbol{L}_i$, and we wish to make inferences about the parameter β_0 which is the unique solution of the full data population estimating equation

$$\mathrm{E}\{M(\boldsymbol{L}; \beta_0)\} = 0 \tag{3}$$

where expectation is taken over the distribution of the complete data $\boldsymbol{L}$. Note that we do not require a model for the distribution of the full data $\boldsymbol{L}$; in fact, estimation is possible under certain weak regularity conditions (van der Vaart 1998) as long as full data unbiased estimating functions exist. Since the empirical version of the estimating function will involve non-monotone missing variables, we only use complete cases weighted by the inverse probability of being a complete case observation to remove any possible selection bias associated with the missing data mechanism. We have

$$\mathrm{E}\left\{\frac{\mathbf{1}(\boldsymbol{R}=1)}{\pi_1(\boldsymbol{L})}M(\boldsymbol{L};\beta_0)\right\}=0 \tag{4}$$

where $\mathbf{1}(\boldsymbol{R}=1)$ is the indicator of a complete case, and $\pi_1(\boldsymbol{L})$ is the corresponding probability of observing complete data. The above unbiasedness of the estimating function (4) holds by straightforward iterated expectations.

The above framework encompasses a great variety of settings under which investigators may wish to account for non-monotone missing data. This includes IPW of the full data score equation, where the score function is such an unbiased estimating function, given a model $f(\boldsymbol{L}|\beta)$ for the law of the full data, in which case (4) reduces to

$$\mathrm{E}\left\{\frac{\mathbf{1}(\boldsymbol{R}=1)}{\pi_1(\boldsymbol{L})}\left.\frac{\partial\log f(\beta|\boldsymbol{L})}{\partial\beta}\right|_{\beta_0}\right\}=0 \tag{5}$$

Note that equation (5) does not necessarily correspond to the observed data score equation, and will therefore generally not achieve the efficiency bound for the model. Estimation can also be extended to classes of semi-parametric models which specify only certain marginal relationships in $\boldsymbol{L}$ and in which scientific interest focuses on some low dimentional functional $\beta=\beta(F_{\boldsymbol{L}})$ of the distribution $F_{\boldsymbol{L}}$ of the full data $\boldsymbol{L}$. For instance, in many health related applications it is common to specify a model $g(\boldsymbol{X},\beta)$ for the conditional mean of the outcome response Y given a set of covariates $\boldsymbol{X}=(X_0,X_1,...,X_P)'$ with $X_0\equiv 1$ for the intercept. Here $\boldsymbol{L}=(Y,\boldsymbol{X})$ and the data could be missing in any of the outcome variables or covariates. Then the parameters of interest can be identified by the population IPW estimating equation

$$\mathrm{E}\left\{\frac{\mathbf{1}(\boldsymbol{R}=1)}{\pi_1(\boldsymbol{L})}[Y-g(\boldsymbol{X},\beta_0)]\boldsymbol{X}\right\}=0 \tag{6}$$

Likewise, if we were to model the conditional median of Y given $\boldsymbol{X}$ with the model $m(\boldsymbol{X},\beta)$, then one could instead identify the true parameter of interest by the IPW population estimating equation

$$\mathrm{E}\left\{\frac{\mathbf{1}(R_1=1)}{\pi_1(\boldsymbol{L})}[\mathbf{1}\{Y<m(\boldsymbol{X},\beta_0)\}-0.5]\boldsymbol{X}\right\}=0 \tag{7}$$

Regression parameters in semi-parametric models for right censored failure time data can likewise be identified by similar IPW population estimating equations, e.g. Cox proportional hazards regression and Aalen's additive hazards regression. Analogous estimating equations are also available for longitudinal and clustered data. Feasible finite sample estimating equations are obtained by replacing population expectations with their empirical counterparts, and $\pi_1(\boldsymbol{L})$ with a consistent estimator. In the next section, we consider several strategies for estimating $\pi_1(\boldsymbol{L})$; we discuss both, certain strategies that although straightforward to implement, are not guaranteed to converge, as well as more principled strategies that do not suffer such limitation.

Briefly, we note that in some instances, IPW estimating functions may be augmented for greater efficiency and double robustness (Robins et al. 1994; van der Laan and Robins 2003; Tsiatis 2006). Unfortunately, AIPW estimating functions for non-monotone missing data are not available in closed form and require solving a complicated integral equation numerically (Robins et al. 1994), which is considerably more computationally intensive than the simple IPW estimators we plan to focus on (Tsiatis 2006). However, we note that estimation of the missing data process is equally relevant for AIPW, therefore the methods described below are very relevant to the construction of doubly robust semi-parametric locally efficient estimators.

3.1 ESTIMATION

3.1.1 The failure of polytomous regression

Unless we have missingness by design, the missingness probability $\pi_1(\boldsymbol{L}, \gamma)$ is estimated with a parametric specification of the missing data mechanism (1). In the case of non-monotone missing data, a straightforward approach to model $\pi_m(\boldsymbol{L}_{(m)}; \gamma)$ using simple polytomous logistic regression will have the unintended consequence of imposing more restrictive conditions than what MAR strictly entails (Robins and Gill 1997). For example, suppose we have 2 variables $\boldsymbol{L} = (L_1, L_2)$ with 4 possible missingness patterns encoded as in the previous section. Polytomous logistic regression produces, for $m = 2, 3, 4$

$$\Pr\{\boldsymbol{R} = m | L_1, L_2\} = \frac{\exp(\gamma_{0m} + \gamma_{1m} L_1 + \gamma_{2m} L_2)}{1 + \sum_{k=2}^{4} \exp(\gamma_{0k} + \gamma_{1k} L_1 + \gamma_{2k} L_2)} \tag{8}$$

By the MAR assumption, since for $\boldsymbol{R} = 4$ neither variable is observed, the probability $\Pr\{\boldsymbol{R} = 4|L_1, L_2\}$ does not depend on either (L_1, L_2) so that $\gamma_{1j} = \gamma_{2j} = 0$ for $j = 2, 3, 4$. Therefore model (6) is also MCAR. In general, it is shown in appendix A.1 using a similar argument that the missing data pattern probabilities modeled using polytomous logistic regression can only depend on the intersection of the sets of observed variables $\boldsymbol{L}_{(m)}$, $m = 2, 3, ..., K$, which imposes more restrictions than MAR only. In the example above, the intersection of variables in the 4 missingness patterns is the null set, hence the missing data process does not depend on any variable and is MCAR.

In the context of binary longitudinal data analysis with non-monotone missing response and missing covariates (Chen and Zhou 2011), the conditional probability for the missing data pattern k at time point j of the i^{th} individual, $\lambda_{ijk} = \Pr(\boldsymbol{R}_{ij} = k|\bar{\boldsymbol{R}}_{ij}, \boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, is modeled using the generalized logistic link, with λ_{ij1} as the reference level

$$\log\left(\frac{\lambda_{ijk}}{\lambda_{ij1}}\right) = \boldsymbol{\mu}_{ijk}^T \alpha_k, \quad k = 2, 3, 4 \tag{9}$$

where $\boldsymbol{\mu}_{ijk}^T$ may be a subset of $(\bar{\boldsymbol{R}}_{ij}, \bar{\boldsymbol{Y}}_{ij}^o, \bar{\boldsymbol{X}}_{ij}^o, \boldsymbol{Z}_i)$, that is, the assumption that the conditional probability for missing data process at time j depend only on previously observed missing data indicators, outcomes and covariates up until time $j - 1$, with $\boldsymbol{Z}$ being covariates that are always observed. This convenient approach has been previously used to model longitudinal MAR processes (Robins et al. 1995; Chen et al. 2010). However, it is natural to assume that the missing data process at time j should also depend on observed outcome and covariate values up to and including time j, and in this sense, the model will be necessarily restricted under MAR for the same reason given earlier, in that for $\boldsymbol{R}_{ij} = 4$, neither $\boldsymbol{X}_{ij}$ nor $\boldsymbol{Y}_{ij}$ is observed and so the other missing data patterns also cannot depend on these variables.

3.1.2 Proposed missing data model

Our approach involves modelling the missingness probability for each missing data pattern separately as a series of logistic regressions

$$\Pr\{\boldsymbol{R} = m|\boldsymbol{L}; \gamma_m\} = \pi_m(\boldsymbol{L}_{(m)}; \gamma_m) = \text{expit}(\gamma_m{}'\boldsymbol{L}_{(m)}) \quad m = 2, ..., M \tag{10}$$

and the probability of observing complete data is

$$\Pr\{\boldsymbol{R} = 1|\boldsymbol{L}; \gamma\} = \pi_1(\boldsymbol{L}; \gamma) = 1 - \sum_{m=2}^{M} \text{expit}({\gamma_m}'\boldsymbol{L}_{(m)}) \tag{11}$$

The above complete case probability depends on the union set of observed variables

$$\bigcup_{m=2}^{M} \boldsymbol{L}_{(m)} \tag{12}$$

Consider the estimator defined as the value which maximizes the unconstrained log-likelihood function corresponding to equations (10) and (11).

$$\sum_{i=1}^{N} \left[\left\{ \sum_{m=2}^{M} \mathbf{1}(\boldsymbol{R}_i = m) \log \pi_m(\boldsymbol{L}_{i(m)}; \gamma_m) \right\} + \mathbf{1}(\boldsymbol{R}_i = 1) \log \left\{ 1 - \sum_{k=2}^{M} \pi_k(\boldsymbol{L}_{i(k)}; \gamma_k) \right\} \right] \tag{13}$$

with the score function

$$\boldsymbol{S}_{\gamma_m} = \sum_{i=1}^{N} \left\{ \mathbf{1}(\boldsymbol{R}_i = m) - \frac{\mathbf{1}(\boldsymbol{R}_i = 1)\pi_m(\boldsymbol{L}_{i(m)})}{\pi_1(\boldsymbol{L}_{i(1)}; \gamma)} \right\} (1 - \pi_m(\boldsymbol{L}_{i(m)})) \left(1, \boldsymbol{L}_{(m)}\right)^T \tag{14}$$

for the parameters γ_m in missing data pattern m, where $\boldsymbol{S}_{\gamma_m}$ and $\left(1, \boldsymbol{L}_{(m)}\right)^T$ have the same dimensions. Parametric specifications (10) and (11) do not define a proper probability mass function unless the following constraints also hold

$$\sum_{k=2}^{M} \pi_k(\boldsymbol{L}_{i(k)}; \gamma_k) < 1 \text{ for } i = 1, 2, ..., N \tag{15}$$

Thus, it may be in practice that the unconstrained maximum likelihood estimator that maximizes (13) will be undefined, if there is at least one complete case for which the empirical version of restriction (15) is not satisfied in the process of finding the maximum likelihood estimate. In actual implementation of the procedure, this translates into failure to converge during maximization of the observed log-likelihood (13). For this reason, we will refer to the equation (13) as the unconstrained log-likelihood function, as it does not naturally impose restrictions (15).

Note that even if the missingness mechanism were known, not all constraints listed in (15) can be observed, since only those for complete cases can be observed to satisfy the restriction. In fact, only the restrictions for complete cases are strictly needed to be enforced in order to ensure that the maximum likelihood estimate can be computed in practice. Thus, one could in principle attempt to maximize the observed data log-likelihood (13) together with the observable constraints

$$\mathbf{1}(\boldsymbol{R}_i = 1) \sum_{k=2}^{M} \pi_k(\boldsymbol{L}_{i(k)}; \gamma_k) < 1 \text{ for } i = 1, 2, ..., N \tag{16}$$

which is potentially computationally prohibitive, since there are as many constraints as complete case observations. Instead, we propose an alternative, more attractive solution upon noting that the above constraints can be reformulated using a Bayesian constrained estimation approach where samples are drawn from the unconstrained posterior conditional distribution for $\boldsymbol{\gamma}$ and only those draws that fall into the constrained parameter space (16) are retained (Gelfand et al. 1992).

To implement the approach, we specify a diffuse prior distribution $\pi(\boldsymbol{\gamma})$ for $\boldsymbol{\gamma} = (\gamma_2, ..., \gamma_M)$ under model (10) and incorporate constraint (16) for each individual with complete data. It is natural to think of the constraint as built into the log-likelihood function (13). The posterior distribution of $\boldsymbol{\gamma}$ under the contrained Bayesian model is proportional to

$$f(\boldsymbol{\gamma}|data) \propto f(data|\boldsymbol{\gamma}) \times \pi(\boldsymbol{\gamma}) = \prod_{i=1}^{N} \left\{ \prod_{m=2}^{M} \{\pi_m(\boldsymbol{L}_{i(m)}; \gamma_m)\}^{\mathbf{1}(\boldsymbol{R}_i = m)} \times \right.$$
$$\left. \left\{ \{1 - \sum_{k=2}^{M} \pi_k(\boldsymbol{L}_{i(k)}; \gamma_k)\} \times \mathbf{1}\{\sum_{k=2}^{M} \pi_k(\boldsymbol{L}_{i(k)}; \gamma_k) < 1\} \right\}^{\mathbf{1}(\boldsymbol{R}_i = 1)} \right\} \times \pi(\boldsymbol{\gamma}) \tag{17}$$

Let $\hat{\pi}_1(\boldsymbol{L}) = 1 - \sum_{m=2}^{M} \text{expit}(\hat{\gamma}'_m \boldsymbol{L}_{(m)})$ where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_2, ...\hat{\gamma}_M)$ is the posterior mode (or mean) from the Bayesian constrained approach. Then an estimate for the parameter of interest β_0 is given by the solution $\hat{\beta}$ to the inverse probability weighted estimating equation

$$\sum_{i=1}^{N} \left\{ \frac{\mathbf{1}(\boldsymbol{R}_i = 1)}{\pi_1(\boldsymbol{L}_i; \hat{\gamma})} M(\boldsymbol{L}_i; \hat{\beta}) \right\} = 0 \tag{18}$$

Subject to standard regularity conditions and assuming that the missing data model given in (10) and (11) is correctly specified, we show in appendix section A.2 that $\hat{\beta}$ is consistent and asymptotically normal

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N\left(0, \text{E}\{\nabla_\beta U(\beta_0, \gamma_0)\}^{-1} \text{Var}\left[U(\beta_0, \gamma_0) - W(\beta_0, \gamma_0)\right] \text{E}\{\nabla_\beta U(\beta_0, \gamma_0)\}^{-1^T}\right) \tag{19}$$

where $U(\beta,\gamma) = \{\mathbf{1}(R_1 = 1)/\pi_1(\boldsymbol{L};\gamma)\}M(\boldsymbol{L};\beta)$, S_{γ_0} is the score function for the missing data mechanism and

$$W(\beta_0,\gamma_0) = \mathrm{E}[U(\beta_0,\gamma_0)S_{\gamma_0}^T]\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S\gamma_0$$

The asymptotic variance in (19) can be consistently estimated by replacing the terms under expectation with empirical averages evaluated at $(\hat{\beta},\hat{\gamma})$.

$$\hat{\mathrm{E}}\{\nabla_\beta U(\hat{\beta},\hat{\gamma})\}^{-1}\hat{\mathrm{Var}}\left[U(\hat{\beta},\hat{\gamma}) - \hat{W}(\hat{\beta},\hat{\gamma})\right]\hat{\mathrm{E}}\{\nabla_\beta U(\hat{\beta},\hat{\gamma})\}^{-1^T} \tag{20}$$

Although the posterior mode (or mean) is asymptotically efficient by the Bernstein-von Mises Theorem, in finite sample the posterior estimate may not necessarily correspond to the solution of the score function (14). For inference under the Bayesian constrained approach, we therefore apply a finite-sample correction to the variance estimate

$$\hat{\mathrm{E}}\{\nabla_\beta U(\hat{\beta},\hat{\gamma})\}^{-1}\hat{\mathrm{Var}}\left[U(\hat{\beta},\hat{\gamma}) - \hat{W}(\hat{\beta},\hat{\gamma}) + \hat{\mathrm{E}}\{W(\hat{\beta},\hat{\gamma})\}\right]\hat{\mathrm{E}}\{\nabla_\beta U(\hat{\beta},\hat{\gamma})\}^{-1^T} \tag{21}$$

so that the term in $\hat{\mathrm{Var}}[\cdot]$ has mean zero empirically. The correction term $\hat{\mathrm{E}}\{W(\hat{\beta},\hat{\gamma})\}$ is expected to vanish as sample size increases. A conservative, albeit more easily implementable, estimate of the asymptotic variance in (19) is obtained by the standard sandwich variance formula (Robins et al. 1994)

$$\hat{\mathrm{E}}\{\nabla_\beta U(\hat{\beta},\hat{\gamma})\}^{-1}\hat{\mathrm{Var}}\left[U(\hat{\beta},\hat{\gamma})\right]\hat{\mathrm{E}}\{\nabla_\beta U(\hat{\beta},\hat{\gamma})\}^{-1^T} \tag{22}$$

4 SIMULATION

In this section we present a simulation study to investigate the finite-sample properties of the proposed estimator. Full data consists of independent and identically distributed $\boldsymbol{L} = (Y, A, \boldsymbol{C})$ with exposure A, binary outcome Y and confounders $\boldsymbol{C} = (C_1, C_2)$. Although there are 16 possible missing data patterns, in actual applications with missing data we rarely observe all missing data patterns, and here we consider the case where only 4 of such patterns are observed: $\boldsymbol{R} = 1$ if we observe $\boldsymbol{L}_{(1)} = \boldsymbol{L}$; $\boldsymbol{R} = 2$ if we observe $\boldsymbol{L}_{(2)} = (Y, A, C_1)$; $\boldsymbol{R} = 3$ if we observe $\boldsymbol{L}_{(3)} = (Y, C_1)$; and $\boldsymbol{R} = 4$ if we observe $\boldsymbol{L}_{(4)} = (A, C_2)$.

The vector (X_1, X_2, X_3) is generated from a multivariate standard normal distribution with correlation coefficent $\rho = 0.3$ between X_1 & X_2 and $\rho = -0.2$ between X_1 & X_3. Then we take $A = \Phi(X_1), C_1 = \Phi(X_2)$ and $C_2 = \Phi(X_3)$ where $\Phi(\cdot)$ is the CDF of the standard normal distribution. Finally, the outcome variable Y is generated as

$$\operatorname{logit} \Pr(Y = 1|A, C_1, C_2) = \alpha_0 + \alpha_1 A + \alpha_2 C_1 + \alpha_3 C_2$$

Where $\boldsymbol{\alpha} = (-0.55, 0.4, -0.35, 0.2)$. Following the MAR assumption, the probabilities of the 4 missing data patterns are generated as

$$\begin{aligned}
\Pr\{\boldsymbol{R} = 2|\boldsymbol{L}\} &= \operatorname{expit}\{\gamma_{2,0} + \gamma_{2,1}Y + \gamma_{2,2}A + \gamma_{2,3}C_1\} \\
\Pr\{\boldsymbol{R} = 3|\boldsymbol{L}\} &= \operatorname{expit}\{\gamma_{3,0} + \gamma_{3,1}Y + \gamma_{3,2}C_1\} \\
\Pr\{\boldsymbol{R} = 4|\boldsymbol{L}\} &= \operatorname{expit}\{\gamma_{4,0} + \gamma_{4,1}A + \gamma_{4,2}C_2\} \\
\Pr\{\boldsymbol{R} = 1|\boldsymbol{L}\} &= 1 - \sum_{m=2}^{4} \Pr\{\boldsymbol{R} = m|\boldsymbol{L}\}
\end{aligned}$$

The observed missing data mechanism is generated from a multinomial distribution based on the above probabilities, and only the corresponding observed data for the sampled pattern contributes to estimation. We performed 1000 replicates each with sample size $n = 500, 1000$ or 2000. The true parameters for the missing data mechanism are

$$\begin{aligned}
\boldsymbol{\gamma_a} &= (-0.68, -0.50, -0.10, 0.10, -0.80, -0.40, -0.30, -1.00, -0.20, 0.30) \\
\boldsymbol{\gamma_b} &= (-0.70, -1.50, -0.60, -0.30, -0.90, -0.60, -0.50, -0.80, -0.70, -0.40) \\
\boldsymbol{\gamma_c} &= (-1.00, -1.50, -1.00, -0.80, -1.00, -1.50, -1.50, -1.00, -1.50, -1.20)
\end{aligned}$$

so that each simulation replicate has approximately 15-25% of complete cases generated under $\boldsymbol{\gamma_a}$, 35-45% of complete cases generated under $\boldsymbol{\gamma_b}$ and 65-75% of complete cases generated under $\boldsymbol{\gamma_c}$.

The parameters $\gamma_{i,j}$ in the missingness mechanism are estimated using both the constrained Bayesian model as well as unconstrained maximum likelihood. The main results for the estimation of parameters $\gamma_{i,j}$ based on simulated data under $\boldsymbol{\gamma_b}$ are summarized in Table 1, while the results under $\boldsymbol{\gamma_a}$ and $\boldsymbol{\gamma_c}$ respectively are included in the supplementary material. For posterior computation, we specify the diffuse priors $\gamma_{i,j} \sim \text{normal}(0, 10^3)$. Adaptive Gibbs sampling (Gilks et al. 1995) was implemented through BRugs, the R interface to the OpenBUGS MCMC software (Lunn et al. 2009). We assessed convergence by visually inspecting the trace plots as well as through the Gelman-Rubin convergence

Table 1: Estimation for the parameters in the simulation missing data model for the scenario with 35-45% complete-cases under $\boldsymbol{\gamma_b}$. Bias and SE refer to relative bias and Monte Carlo standard error respectively. Posterior mean is used as estimate in constrained Bayesian estimation. Results for unconstrained maximum likelihood estimation are restricted to those runs that did converge.

	$n = 500$				$n = 1000$				$n = 2000$			
	Cons.	Bayes	Uncons.	MLE	Cons.	Bayes	Uncons.	MLE	Cons.	Bayes	Uncons.	MLE
	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$\gamma_{2,0}$	0.04	0.28	0.07	0.26	0.04	0.20	0.03	0.19	0.01	0.14	0.00	0.13
$\gamma_{2,1}$	0.05	0.34	0.02	0.32	0.02	0.23	0.01	0.23	0.01	0.16	0.01	0.16
$\gamma_{2,2}$	-0.06	0.43	-0.08	0.41	-0.03	0.29	-0.00	0.29	-0.04	0.20	-0.01	0.20
$\gamma_{2,3}$	0.11	0.41	-0.05	0.40	-0.01	0.29	-0.08	0.28	0.06	0.21	0.02	0.20
$\gamma_{3,0}$	0.04	0.23	0.01	0.22	0.02	0.16	0.01	0.16	0.01	0.11	0.00	0.11
$\gamma_{3,1}$	-0.01	0.24	-0.00	0.24	0.03	0.17	0.02	0.17	0.00	0.11	0.00	0.11
$\gamma_{3,2}$	-0.00	0.37	-0.00	0.37	-0.02	0.27	-0.02	0.27	-0.02	0.19	-0.00	0.19
$\gamma_{4,0}$	0.02	0.29	0.07	0.27	0.00	0.21	0.04	0.19	-0.01	0.14	0.02	0.13
$\gamma_{4,1}$	-0.00	0.38	-0.05	0.37	0.00	0.27	-0.03	0.26	0.02	0.19	-0.00	0.19
$\gamma_{4,2}$	0.01	0.35	-0.15	0.32	-0.01	0.24	-0.08	0.23	0.01	0.17	-0.03	0.17

statistic (Gelman and Rubin 1992), and included an adaptive phase of 10000 iterations out of a total of 20000 iterations. The effect of exposure α_1 in the outcome regression is estimated using weighted logistic regression. We performed unweighted complete-case regression to evaluate the magnitude of selection bias, and also carried out maximum likelihood estimation on the original full data without missing variables as a gauge for the efficiency bound (Table 2). The results for the estimation of parameters $(\alpha_0, \alpha_2, \alpha_3)$ are included in the supplementary material.

The bias of the IPW estimators using constrained Bayesian estimation or unconstrained maximum likelihood procedure generally decreases with increasing sample size or proportion of complete-cases in the data, and the bias becomes negligible at moderate complete case sample sizes and higher. Our estimator of the asymptotic variance given by equation (19) performs quite well and is close to the Monte Carlo standard errors. The standard deviation estimated using the standard sandwich estimator (22) is typically larger than the asymptotic variance estimate, which follows from the theoretical result that the sandwich estimator produces a conservative estimate of the true asymptotic variance. The coverage using the consistent estimator of the asymptotic variance is close to the nominal 95% using Wald confidence intervals.

Table 2: Estimation for effect of exposure $\alpha_1 = 0.4$ in the logistic regression model from 1000 simulation replicates. Posterior mean is used as estimate in constrained Bayesian estimation. Results for unconstrained maximum likelihood estimation are restricted to those runs that did converge. Sandwich variance estimators are given in (22). The asymptotic variance estimators are given in (20) and (21). For complete-case analysis and full data maximum likelihood, the asymptotic variance estimates are obtained via Fisher scoring. Coverages are based on nominal 95% Wald confidence intervals.

IPW Bayesian Constrained Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sand. SE	% Cover	Asymp. SE	% Cover
15-25	500	100	-0.36	0.83	0.92	96.8	0.82	95.0
	1000	100	-0.25	0.58	0.63	96.2	0.56	94.8
	2000	100	-0.12	0.39	0.44	97.3	0.38	95.3
35-45	500	100	-0.10	0.50	0.55	97.2	0.50	94.7
	1000	100	-0.08	0.36	0.39	96.6	0.35	94.8
	2000	100	-0.02	0.25	0.37	96.2	0.25	94.9
65-75	500	100	-0.02	0.41	0.41	95.6	0.40	94.8
	1000	100	0.00	0.28	0.29	95.1	0.28	94.3
	2000	100	0.00	0.20	0.20	95.2	0.20	94.5

IPW Unconstrained ML Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sand. SE	% Cover	Asymp. SE	% Cover
15-25	500	56.9	-0.08	0.80	0.91	96.8	0.79	95.4
	1000	72.2	-0.04	0.57	0.63	97.0	0.55	95.0
	2000	83.3	-0.03	0.39	0.44	97.3	0.38	94.9
35-45	500	81.1	-0.03	0.49	0.55	97.5	0.50	95.5
	1000	91.7	-0.02	0.35	0.39	97.0	0.35	94.7
	2000	97.2	-0.01	0.25	0.35	96.3	0.24	94.8
65-75	500	96.3	0.00	0.41	0.41	95.4	0.40	94.7
	1000	99.7	0.01	0.28	0.29	94.9	0.28	94.3
	2000	100	0.00	0.20	0.20	94.8	0.20	94.5

Complete-Case Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sand. SE	% Cover	Asymp. SE	% Cover
15-25	500	100	-1.08	0.91			0.89	92.5
	1000	100	-1.10	0.62			0.61	88.7
	2000	100	-1.04	0.43			0.42	83.2
35-45	500	100	-1.19	0.54			0.54	84.7
	1000	100	-1.24	0.39			0.38	72.7
	2000	100	-1.20	0.27			0.26	55.9
65-75	500	100	-0.55	0.41			0.40	90.8
	1000	100	-0.52	0.29			0.28	88.7
	2000	100	-0.54	0.20			0.20	82.0

Full Data Maximum Likelihood

% CC	n	%Converge	Rel. Bias	MC SE	Sand. SE	% Cover	Asymp. SE	% Cover
15-25	500	100	-0.01	0.35			0.34	94.8
	1000	100	0.02	0.25			0.24	94.8
	2000	100	-0.02	0.16			0.17	95.2
35-45	500	100	0.03	0.35			0.34	95.0
	1000	100	-0.01	0.24			0.24	95.3
	2000	100	0.02	0.17			0.17	94.9
65-75	500	100	0.01	0.35			0.34	94.6
	1000	100	0.02	0.24			0.24	94.7
	2000	100	0.00	0.17			0.17	94.5

The proportion of simulated samples where the unconstrained maximum likelihood estimate converged increased both with total sample size, and with the proportion of complete cases. We note that the bias of the estimator using constrained Bayesian estimation is larger than that using unconstrained maximum likelihood estimation when the latter converges, particularly when sample size is small and the proportion of complete cases is low. Similar finite sample bias has been reported in previous implementations of the constrained Bayesian estimation in a log-linear model of risk (Chu and Cole 2010). However, even so, as noted above the coverage of 95% confidence intervals does not appear to be affected and the bias appears to vanish as sample size increases. The bias could be potentially due to the small number of available complete cases for which constraints (16) are imposed. The constrained Bayesian estimation can be adapted to impose the constraints over a range of possible covariate combinations for all individuals, if we were to assume bounds on the domain of the full data. Nevertheless, the constrained Bayesian procedure is guaranteed to produce an estimate for $\pi_1(\boldsymbol{L})$ within the parameter space of the model which may be used for IPW. Restricting our analysis to only those simulated samples in which the unconstrained maximum likelihood procedure converged, then constrained Bayesian estimation presents similar bias compared to the unconstrained maximum likelihood method (results not shown). The simulation results indicate that both IPW estimators have smaller bias and greater efficiency than the complete-case only analysis.

5 APPLICATION

The empirical application concerns a study of highly active antiretroviral therapy (HAART) and adverse birth outcomes among HIV-infected women in Botswana. A detailed description of the study cohort has been presented elsewhere (Chen et al. 2012). The entire study cohort consists of 33148 obstetrical records abstracted from 6 sites in Botswana for 24 months. Our current analysis focuses on the subset of women who were known to be HIV positive ($n = 9711$).

Two adverse birth outcomes of interest are the binary variables stillbirth and small for gestational age. Stillbirth was defined as fetal death with an Apgar score of 0 and small for gestational age as below the 10th percentile of birthweight by gestational age.

Table 3: Tabulation of non-monotone missing data patterns as a percentage of total data ($n = 9711$). Missing variables are indicated by 0. Complete-cases are given in the first pattern.

Analysis with outcome stillbirth					
pattern	Stillbirth	Hypertension	Low CD4$^+$	Continued HAART	% of data
1	1	1	1	1	45.7
2	1	0	1	1	0.9
3	1	1	0	1	47.8
4	1	0	0	1	5.6
Analysis with outcome small for gestational age					
pattern	Small gest. age	Hypertension	Low CD4$^+$	Continued HAART	% of data
1	1	1	1	1	45.0
2	0	1	1	1	0.7
3	1	0	1	1	0.7
4	0	0	1	1	0.2
5	1	1	0	1	46.6
6	0	1	0	1	1.2
7	1	0	0	1	3.9
8	0	0	0	1	1.7
Analysis with outcome birthweight					
pattern	Birthweight	Hypertension	Low CD4$^+$	Continued HAART	% of data
1	1	1	1	1	45.4
2	0	1	1	1	0.2
3	1	0	1	1	0.9
4	1	1	0	1	47.6
5	0	1	0	1	0.2
6	1	0	0	1	5.6
7	0	0	0	1	0.1

The data contains a number of predictors of interest with unobserved values (Table 3): maternal hypertension in pregnancy (6.5% missing), whether CD4$^+$ cell count is less than 200 μL (53.4% missing) and whether a woman continued HAART in pregnancy or not. Our goal is to correlate these factors with the risks of the two adverse birth outcomes. The outcome variable small for gestational age itself has 3.8% of values missing, while stillbirth data is available for all. We applied the proposed IPW estimator in logistic regression as well as performed complete case analysis and multivariate imputation by chained equations (MICE) (van Buuren and Groothuis-Oudshoorn 2011) with $M = 50$ imputed samples for comparison. In addition, we investigated the effects of the predictors on the continuous variable birthweight (0.5% missing) through IPW linear regression (Table 4).

Unconstrained maximum likelihood estimation of the missing data model converged in

Table 4: Analysis for outcomes stillbirth, small for gestational age and birthweight. The values presented for the first 2 outcomes are estimated odds ratios from logistic regression, while the values for birthweight are estimated coefficients (in grams) from linear regression. Wald 95% confidence intervals are constructed with standard errors for both IPW estimators using (20) and (21), and using Fisher scoring for complete-case analysis. The standard error for MICE is estimated by Rubin's formula with $M = 50$ imputed samples.

		Stillbirth		
Predictor	Complete-case	IPW Max. Likelihood	IPW Cons. Bayesian	MICE
l Hypertension	3.56 (2.59, 4.88)	4.04 (2.87, 5.71)	4.03 (2.86, 5.70)	3.34 (2.72, 4.10)
Low CD4$^+$	1.71 (1.16, 2.54)	1.63 (1.09, 2.45)	1.64 (1.09, 2.45)	1.59 (1.05, 2.41)
Cont. HAART	1.86 (1.25, 2.79)	2.01 (1.38, 2.94)	2.00 (1.37, 2.93)	1.55 (1.26, 1.91)
		Small for Gestational Age		
Predictor	Complete-case	IPW Max. Likelihood	IPW Cons. Bayesian	MICE
Hypertension	1.58 (1.30, 1.92)	1.62 (1.31, 1.99)	1.62 (1.30, 1.99)	1.68 (1.47, 1.92)
Low CD4$^+$	1.70 (1.37, 2.11)	1.61 (1.28, 2.03)	1.62 (1.29, 2.03)	1.49 (1.05, 2.10)
Cont. HAART	1.90 (1.52, 2.37)	1.97 (1.58, 2.46)	1.97 (1.58, 2.46)	1.85 (1.64, 2.09)
		Birthweight		
Predictor	Complete-case	IPW Max. Likelihood	IPW Cons. Bayesian	MICE
Hypertension	-162 (-217, -106)	-249 (-278, -220)	-248 (-277, -220)	-233 (-269, -197)
Low CD4$^+$	-118 (-174, -63)	-108 (-162, -56)	-109 (-162, -56)	-115 (-174, -56)
Cont. HAART	-226 (-288, -165)	-270 (-289, -251)	-269 (-288, -250)	-213 (-244, -181)

all 3 analyses. Given the reasonably large sample size ($n = 9711$), the results from unconstrained maximum likelihood estimation are similar to those obtained by constrained Bayesian estimation, consistent with findings from both the simulation study and asymptotics theory. For the first outcome risk of stillbirth, we note that IPW produces an estimated odds ratio for maternal hypertension which is greater by 13.5% compared with complete-case results. However, the 95% confidence intervals for IPW, complete-case and MICE logistic regression estimates overlapped. For the second outcome small for gestational age, results were very similar across different methods, suggesting little selection bias might be present, that can be accounted for by IPW or MICE. The efficiency of IPW estimators is similar to that of complete-case estimators, while the efficiency of MICE estimators is more variable depending on the proportion of missing data in each predictor.

The estimates for the linear regression of birthweight are more affected by IPW and MICE adjustments. The IPW estimates of the average decrease in birthweight increases by 53.7% with maternal hypertension and 19.5% with continued HAART treatment, compared with complete-case analysis. Differences between MICE and IPW estimates may reflect differences of modeling assumptions between the methods because one is making assumptions on the full data univariate conditional laws in the former (which as previously noted may be vulnerable to bias due to model incompatibility) and the missing data model in the latter, although validity for either method requires MAR.

6 DISCUSSION

We have proposed a simple yet general class of missing data models for non-monotone MAR mechanisms which makes no assumption about the full data distribution. Our models are explicit in their dependence on only the observed variables, and the proposed inverse probability weighted estimators can be easily implemented using existing software. An important contribution of the paper is a proposed strategy to estimate the missing data mechanism under MAR while circumventing potential convergence difficulties encountered with unconstrained maximum likelihood estimation of the missing data process.

Assuming no model misspecification, the proposed IPW estimator corrects the bias of complete-case analysis and may be used whenever one has available a full data estimating equation. While constrainted Bayesian estimation is guaranteed to produce valid probability weights for subsequent estimation of a full data regression or other functionals of interest, we found that there was non-negligible finite sample bias in small samples. However, this bias appears to vanish at moderate to large sample sizes. The bias may be due to the fact that constraints (16) are imposed on complete-cases only, and thus the constraints may not be satisfied for incomplete cases. The constrained Bayesian approach could be adapted to impose the constraints over a finite range of possible values for the full data $\boldsymbol{L}$, if bounds for the sample space were known.

Lastly, Robins and Gill have argued that the class of RMM models represents the most general plausible physical mechanism for generating non-monotone missing data (Robins and Gill 1997). Therefore, they have effectively argued that any model within our class that is not RMM may be difficult to motivate scientifically. However we emphasize that the perspective we have presented is completely agnostic as to whether a particular submodel of MAR may be more scientifically meaningful than another; in fact, RMM, like any other submodel of MAR, can be accommodated by the proposed approach, but would require placing additional appropriate constraints while sampling from the posterior, to ensure that one remains within the submodel. This will necessarily result in a more complicated fitting procedure, with little apparent benefit for bias reduction or efficiency gain. This is because, as is well known in the missing data literature, it is generally advisable for efficiency considerations in IPW estimation under MAR, that one estimates the probability of a complete-case using as richly parameterized a regression as empirically feasible (Robins et al. 1994). This implies that even if RMM is correctly specified, one would generally benefit from including correlates of the full data estimating equation into a model for the missing data mechanism, even if such variables do not necessarily correlate with the missing data process. We believe such efficiency considerations trump any concern for scientific interpretation of the model for the missing data process, particularly since after all, the missing data process is technically a nuisance parameter not of primary scientific interest.

APPENDIX: PROOFS

A.1 Restrictions imposed by polytomous logistic regression model

Suppose there are M missingness patterns, each with observed variables $\boldsymbol{L}_{(m)}$, $m = 1, ..., M$. Choosing pattern j as the baseline category, we model the other missingness pattern probabilities as

$$\Pr\{\boldsymbol{R} = m|\boldsymbol{L}\} = \frac{\exp(\boldsymbol{\gamma_m}'\boldsymbol{L}_{(m)})}{1 + \sum_{k\in\{1,...,M\}\setminus\{j\}} \exp(\boldsymbol{\gamma_k}'\boldsymbol{L}_{(k)})} \quad \text{for} \quad m \in \{1, ..., M\} \setminus \{j\}$$

Let $\boldsymbol{L}_I = \bigcap_{m\in\{1,...,M\}\setminus\{j\}} \boldsymbol{L}_{(m)}$. Then by the MAR assumption, each of the above probabilities $\Pr\{\boldsymbol{R} = m|\boldsymbol{L}\}$ depends on $\boldsymbol{L}_{(m)}$ respectively. But they can only depend on $\boldsymbol{L}_I$. If not, then the probability for one of the missing data patterns h will depend on variables $\boldsymbol{L}_{(h)} \setminus \boldsymbol{L}_I$ that another pattern does not have. This is not possible due to the linked nature of the terms in the denominator of the probability expression.

A.2 Asymptotic results for IPW estimator

The consistency of $\hat{\beta}$ can be established under general conditions for 2-step estimators (Newey and McFadden 1993) to show uniform convergence of estimating equation (18) in β, where we make use of the fact that $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}$. Typically one would need to impose moment assumptions on $\pi_1(\boldsymbol{L};\gamma)$ and $M(\boldsymbol{L};\beta)$ (Wooldridge 2007).

To investigate the asymptotic distribution of $\hat{\beta}$, under suitable regularity conditions expand (18) around the true values β_0 and subsequently γ_0,

$$\sqrt{n}(\hat{\beta} - \beta_0) = -\left[\frac{1}{n}\sum_{i=1}^{n} \nabla_\beta U_i(\beta^*, \hat{\gamma})\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(\beta_0, \hat{\gamma})$$

$$= -\left[\frac{1}{n}\sum_{i=1}^{n} \nabla_\beta U_i(\beta^*, \hat{\gamma})\right]^{-1} \times \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(\beta_0, \gamma_0) + \left(\frac{1}{n}\sum_{i=1}^{n} \nabla_\gamma U_i(\beta_0, \gamma^*)\right)\sqrt{n}(\hat{\gamma} - \gamma_0)\right]$$

where β^* and γ^* are the mean values and $U(\beta, \gamma) = \{\mathbf{1}(R_1 = 1)/\pi_1(\boldsymbol{L};\gamma)\}M(\boldsymbol{L};\beta)$. When $\hat{\gamma}$ is the maximum likelihood estimator or a Bayes point estimator satisfying conditions in the Bernstein-von Mises Theorem, it is an asymptotically linear estimator with the influence function

$$\sqrt{n}(\hat{\gamma} - \gamma_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathrm{E}\left[S_{\gamma_0} S_{\gamma_0}^T\right]^{-1} S_i\gamma_0 + o_p(1) \tag{23}$$

where S_γ is the score function with respect to the missing data model parameters γ. Substituting the influence function representation into previous expansion gives

$$\begin{aligned}&\sqrt{n}(\hat{\beta}-\beta_0)\\ &= -\mathrm{E}\{\nabla_\beta U(\beta_0,\gamma_0)\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{U_i(\beta_0,\gamma_0)+\mathrm{E}\{\nabla_\gamma U(\beta_0,\gamma_0)\}\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S_i\gamma_0\right\}+o_p(1)\end{aligned} \tag{24}$$

In addition, from the assumption that the parameters governing full data and the missing data process are separable, under standard regularity conditions we have for observed data $\boldsymbol{O}$

$$\begin{aligned}\mathrm{E}[U(\beta,\gamma)] &= \int U(\beta,\gamma)f(\boldsymbol{O};\beta,\gamma)\,d\boldsymbol{O}=0\\ \frac{\partial}{\partial\gamma}\mathrm{E}[U(\beta,\gamma)] &= \int\frac{\partial}{\partial\gamma}U(\beta,\gamma)f(\boldsymbol{O};\beta,\gamma)\,d\boldsymbol{O}+\int U(\beta,\gamma)\frac{\partial}{\partial\gamma}f(\boldsymbol{O};\beta,\gamma)\,d\boldsymbol{O}=0\\ \implies \mathrm{E}\{\nabla_\gamma U(\gamma,\beta)\} &= -\int U(\beta,\gamma)\frac{\frac{\partial}{\partial\gamma}f(\boldsymbol{O};\beta,\gamma)}{f(\boldsymbol{O};\beta,\gamma)}f(\boldsymbol{O};\beta,\gamma)\,d\boldsymbol{O}=-\mathrm{E}[U(\beta,\gamma)S_\gamma]\end{aligned}$$

Substituting the above equality to (24)

$$\begin{aligned}&\sqrt{n}(\hat{\beta}-\beta_0)=\\ &-\mathrm{E}\{\nabla_\beta U(\beta_0,\gamma_0)\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{U_i(\beta_0,\gamma_0)-\mathrm{E}[U(\beta_0,\gamma_0)S_{\gamma_0}^T]\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S_i\gamma_0\right\}+o_p(1)\end{aligned}$$

An application of Slutsky's theorem shows that

$$\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{d}N\left(0,\mathrm{E}\{\nabla_\beta U(\beta_0,\gamma_0)\}^{-1}\mathrm{Var}\left[U(\beta_0,\gamma_0)-W(\beta_0,\gamma_0)\right]\mathrm{E}\{\nabla_\beta U(\beta_0,\gamma_0)\}^{-1^T}\right) \tag{25}$$

where

$$W(\beta_0,\gamma_0)=\mathrm{E}[U(\beta_0,\gamma_0)S_{\gamma_0}^T]\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S\gamma_0$$

The sandwich estimator is consistent for $\mathrm{E}\{\nabla_\beta U(\beta_0,\gamma_0)\}^{-1}\mathrm{E}\left[U(\beta_0,\gamma_0)^{\bigotimes 2}\right]\mathrm{E}\{\nabla_\beta U(\beta_0,\gamma_0)\}^{-1^T}$. In the Hilbert space of mean-zero random functions, $\mathrm{E}\left[U(\beta_0,\gamma_0)S_{\gamma_0}^T\right]\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S\gamma_0$ is the projection of $U(\beta_0,\gamma_0)$ onto the linear subspace spanned by elements of S_{γ_0}. Therefore by Pythagorean Theorem

$$\mathrm{E}\left[U(\beta_0,\gamma_0)^{\bigotimes 2}\right]-\mathrm{E}\left[\left\{U(\beta_0,\gamma_0)-\mathrm{E}\left[U(\beta_0,\gamma_0)S_{\gamma_0}^T\right]\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S\gamma_0\right\}^{\bigotimes 2}\right]$$

is positive semi-definite and the sandwich estimator provides conservative estimate for the true asymptotic variance.

References

Chen, B., Yi, G. Y. and Cook, R. J. (2010), 'Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random', *Journal of the American Statistical Association* **105(489)**, 336–353.

Chen, B. and Zhou, X.-H. (2011), 'Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates', *Biometrics* **67(3)**, 830–842.

Chen, J. Y., Ribaudo, H. J., Souda, S., Parekh, N., Ogwu, A., Lockman, S., Powis, K., Dryden-Peterson, S., Creek, T., Jimbo, W., Madidimalo, T., Makhema, J., Essex, M. and Shapiro, R. L. (2012), 'Highly active antiretroviral therapy and adverse birth outcomes among hiv-infected women in botswana', *The Journal of Infectious Diseases* **206(11)**, 1695–1705.

Chu, H. and Cole, S. R. (2010), 'Estimation of risk ratios in cohort studies with common outcomes: A bayesian approach', *Epidemiology* **21(6)**, 855–862.

Chu, H. and Cole, S. R. (2011), 'Estimating the relative excess risk due to interaction: A bayesian approach', *Epidemiology* **22(2)**, 242–248.

Dempster, A., Laird, N. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society. Series B (Methodological)* **39(1)**, 1–38.

Gelfand, A. E., Smith, A. F. M. and Lee, T.-M. (1992), 'Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling', *Journal of the American Statistical Association* **87(418)**, 523–532.

Gelman, A. and Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7(4)**, 457–511.

Gilks, W., Best, N. and Tan, K. (1995), 'Adaptive rejection metropolis sampling within gibbs sampling', *Applied Statistics* **44(4)**, 455–472.

Horton, N. J. and Lipsitz, S. R. (2001), 'Multiple imputation in practice: Comparison of software packages for regression models with missing variables', *The American Statistician* **55(3)**, 244–254.

Horvitz, D. and Thompson, D. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47(260)**, 663–685.

Kenward, M. and Carpenter, J. (2007), 'Multiple imputation: Current perspectives', *Statistical Methods in Medical Research* **16**, 199–218.

Little, R. J. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Wiley.

Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009), 'The bugs project: Evolution, critique and future directions', *Statistics in Medicine* **28(25)**, 3049–3067.

Molengerghs, G., Thijs, H., Jansen, I. and Beunckens, C. (2004), 'Analyzing incomplete longitudinal clinical trial data', *Biostatistics* **5(3)**, 445–464.

Newey, W. and McFadden, D. (1993), Large sample estimation and hypothesis testing, *in* D. McFadden and R. Engler, eds, 'Handbook of Econometrics', Vol. 4, North-Holland.

Robins, J. M. and Gill, R. D. (1997), 'Non-response models for the analysis of non-monotone ignorable missing data', *Statistics in Medicine* **16**, 39–56.

Robins, J. M. and Rotnitzky, A. (1997), 'Analysis of semiparametric regression models with non-ignorable non-response', *Statistics in Medicine* **16**, 81–102.

Robins, J. M., Rotnitzky, A. and Scharfstein, D. O. (1999), Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models, *in* M. E. Halloran and D. Berry, eds, 'Statistical Models in Epidemiology: The Environment and Clinical Trials', Springer-Verlag.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American Statistical Association* **89(427)**, 846–866.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995), 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association* **90**, 106–121.

Rubin, D. B. (1974), 'Characterizing the estimation of parameters in incomplete-data problems', *Journal of the American Statistical Association* **69(346)**, 467–474.

Rubin, D. B. (1977), 'Formalizing subjective notions about the effect of nonrespondents in sample surveys', *Journal of the American Statistical Association* **72**, 538–543.

Schafer, J. (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall.

Siddiqui, O. and Ali, M. W. (1998), 'A comparison of the random-effects pattern mixture model with last-observation-carried-forward (locf) analysis in longitudinal clinical trials with dropouts', *Journal of Biopharmaceutical Statistics* **8(4)**, 545–563.

Tsiatis, A. (2006), *Semiparametric Theory and Missing Data*, Springer.

van Buuren, S. (2007), 'Multiple imputation of discrete and continuous data by fully conditional specification', *Statistical Methods in Medical Research* **16**, 219–242.

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

van Buuren, S. and Oudshoorn, C. (2000), 'Multivariate imputation by chained equations: Mice v1.0 users manual', *Leiden: TNO Prevention and Health* .

van der Laan, M. J. and Robins, J. M. (2003), *Unified Methods for Censored Longitudinal Data and Causality*, Springer.

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press.

White, I. R., Royston, P. and Wood, A. M. (2011), 'Multiple imputation using chained equations: Issues and guidance for practice', *Statistics in Medicine* **30(4)**, 377–399.

Wooldridge, J. (2007), 'Inverse probability weighted m-estimation for general missing data problems', *Journal of Econometrics* **141**, 1281–1301.

SUPPLEMENTARY MATERIALS

Estimation of parameters $\gamma_{i,j}$ generated under $\boldsymbol{\gamma_a}$ and $\boldsymbol{\gamma_c}$

Table 5: Estimation for the parameters in the simulation missing data model for the scenario with 15-25% complete-cases under $\boldsymbol{\gamma_a}$. Bias and SE refer to relative bias and Monte Carlo standard error respectively. Posterior mean is used as estimate in constrained Bayesian estimation. Results for unconstrained maximum likelihood estimation are restricted to those runs that did converge.

	$n = 500$				$n = 1000$				$n = 2000$			
	Cons.	Bayes	Uncons.	MLE	Cons.	Bayes	Uncons.	MLE	Cons.	Bayes	Uncons.	MLE
	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$\gamma_{2,0}$	0.03	0.24	0.06	0.23	0.01	0.16	0.03	0.16	0.01	0.11	0.01	0.11
$\gamma_{2,1}$	0.02	0.20	-0.00	0.20	0.01	0.13	-0.00	0.13	-0.00	0.10	-0.01	0.09
$\gamma_{2,2}$	-0.30	0.33	-0.31	0.31	-0.20	0.22	-0.22	0.22	-0.13	0.16	-0.06	0.16
$\gamma_{2,3}$	-0.24	0.33	0.37	0.31	-0.19	0.22	0.07	0.22	-0.06	0.15	0.08	0.15
$\gamma_{3,0}$	0.04	0.21	0.04	0.20	0.02	0.14	0.02	0.14	0.01	0.10	0.01	0.10
$\gamma_{3,1}$	0.02	0.21	-0.01	0.22	-0.01	0.15	0.00	0.14	-0.00	0.11	0.01	0.10
$\gamma_{3,2}$	-0.04	0.35	-0.16	0.33	-0.04	0.23	-0.10	0.23	-0.02	0.16	-0.03	0.16
$\gamma_{4,0}$	-0.02	0.24	-0.03	0.22	0.00	0.17	0.00	0.17	-0.01	0.12	-0.00	0.11
$\gamma_{4,1}$	0.23	0.33	0.05	0.32	0.08	0.23	-0.04	0.23	0.04	0.16	-0.02	0.16
$\gamma_{4,2}$	-0.05	0.25	-0.17	0.20	-0.02	0.17	-0.06	0.16	-0.04	0.12	-0.05	0.11

Table 6: Estimation for the parameters in the simulation missing data model for the scenario with 65-75% complete-cases under $\boldsymbol{\gamma_c}$. Bias and SE refer to relative bias and Monte Carlo standard error respectively. Posterior mean is used as estimate in constrained Bayesian estimation. Results for unconstrained maximum likelihood estimation are restricted to those runs that did converge.

	$n = 500$				$n = 1000$				$n = 2000$			
	Cons.	Bayes	Uncons.	MLE	Cons.	Bayes	Uncons.	MLE	Cons.	Bayes	Uncons.	MLE
	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$\gamma_{2,0}$	0.04	0.36	0.04	0.35	0.02	0.24	0.01	0.25	0.00	0.18	-0.00	0.18
$\gamma_{2,1}$	0.08	0.49	0.04	0.46	0.04	0.32	0.02	0.33	0.02	0.21	0.01	0.21
$\gamma_{2,2}$	0.03	0.61	0.01	0.60	-0.02	0.41	-0.02	0.41	0.01	0.30	0.02	0.30
$\gamma_{2,3}$	0.03	0.59	-0.01	0.58	0.04	0.40	0.02	0.40	0.03	0.29	0.01	0.29
$\gamma_{3,0}$	0.04	0.29	0.01	0.28	0.02	0.20	0.00	0.21	0.02	0.14	0.00	0.14
$\gamma_{3,1}$	0.06	0.45	0.04	0.42	0.03	0.28	0.01	0.30	0.01	0.20	0.01	0.20
$\gamma_{3,2}$	0.02	0.54	0.01	0.54	-0.00	0.38	-0.00	0.40	-0.01	0.28	0.00	0.28
$\gamma_{4,0}$	0.01	0.40	0.01	0.38	0.02	0.28	0.01	0.28	-0.01	0.20	-0.00	0.20
$\gamma_{4,1}$	0.03	0.60	0.02	0.58	0.00	0.41	-0.01	0.42	0.01	0.28	0.01	0.28
$\gamma_{4,2}$	0.04	0.53	0.01	0.51	0.00	0.38	-0.01	0.38	0.01	0.27	0.00	0.27

Estimation of parameters $(\alpha_0, \alpha_2, \alpha_3)$

Table 7: Estimation for intercept $\alpha_0 = -0.55$ in the logistic regression model from 1000 simulation replicates. Posterior mean is used as estimate in constrained Bayesian estimation. Results for unconstrained maximum likelihood estimation are restricted to those runs that did converge. Sandwich variance estimators are given in (22). The asymptotic variance estimators are given in (20) and (21). For complete-case analysis and full data maximum likelihood, the asymptotic variance estimates are obtained via Fisher scoring. Coverages are based on nominal 95% Wald confidence intervals.

IPW Bayesian Constrained Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	100	-0.25	0.68	0.78	96.6	0.66	94.1
	1000	100	-0.15	0.46	0.53	98.3	0.45	95.6
	2000	100	-0.09	0.32	0.37	97.5	0.32	94.9
35-45	500	100	-0.16	0.44	0.50	97.4	0.44	94.2
	1000	100	-0.08	0.29	0.35	98.0	0.30	95.3
	2000	100	-0.02	0.22	0.25	96.4	0.21	94.0
65-75	500	100	-0.05	0.32	0.36	97.6	0.33	95.7
	1000	100	-0.02	0.24	0.26	96.7	0.24	94.9
	2000	100	-0.01	0.17	0.18	96.9	0.17	95.4

IPW Unconstrained ML Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	56.9	-0.05	0.66	0.76	97.8	0.64	94.6
	1000	72.2	-0.03	0.45	0.53	98.7	0.45	95.7
	2000	83.3	-0.01	0.32	0.37	97.4	0.32	94.5
35-45	500	81.1	-0.03	0.43	0.50	97.7	0.43	95.3
	1000	91.7	-0.02	0.29	0.35	98.0	0.30	95.2
	2000	97.2	0.01	0.22	0.25	96.8	0.21	94.2
65-75	500	96.3	-0.01	0.32	0.37	97.8	0.33	94.7
	1000	99.7	-0.01	0.24	0.26	96.7	0.24	94.9
	2000	100	0.00	0.17	0.18	96.9	0.17	95.4

Complete-Case Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	100	-2.24	0.80			0.79	63.4
	1000	100	-2.17	0.53			0.52	35.7
	2000	100	-2.16	0.36			0.36	8.5
35-45	500	100	-2.10	0.48			0.48	33.0
	1000	100	-2.08	0.33			0.34	6.6
	2000	100	-2.06	0.24			0.24	0.1
65-75	500	100	-1.19	0.33			0.36	55.9
	1000	100	-1.18	0.25			0.25	27.7
	2000	100	-1.17	0.18			0.18	4.2

Full Data Maximum Likelihood

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	100	-0.01	0.29			0.29	95.6
	1000	100	0.02	0.21			0.20	94.8
	2000	100	-0.01	0.14			0.14	95.3
35-45	500	100	0.01	0.29			0.29	95.4
	1000	100	0.00	0.20			0.20	94.9
	2000	100	0.01	0.14			0.14	95.2
65-75	500	100	-0.01	0.28			0.29	95.5
	1000	100	0.00	0.20			0.20	94.2
	2000	100	0.00	0.14			0.14	95.0

Table 8: Estimation for effect of first confounder $\alpha_2 = -0.35$ in the logistic regression model from 1000 simulation replicates. Posterior mean is used as estimate in constrained Bayesian estimation. Results for unconstrained maximum likelihood estimation are restricted to those runs that did converge. Sandwich variance estimators are given in (22). The asymptotic variance estimators are given in (20) and (21). For complete-case analysis and full data maximum likelihood, the asymptotic variance estimates are obtained via Fisher scoring. Coverages are based on nominal 95% Wald confidence intervals.

IPW Bayesian Constrained Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	100	0.35	0.83	0.90	96.4	0.80	94.5
	1000	100	0.10	0.55	0.62	97.8	0.54	95.0
	2000	100	0.04	0.38	0.43	97.8	0.37	95.4
35-45	500	100	0.12	0.50	0.54	97.2	0.51	94.6
	1000	100	0.04	0.34	0.37	97.7	0.34	94.3
	2000	100	0.01	0.24	0.26	97.2	0.24	94.8
65-75	500	100	0.02	0.38	0.40	96.3	0.38	94.8
	1000	100	0.03	0.26	0.28	96.4	0.26	95.1
	2000	100	-0.01	0.19	0.20	95.4	0.19	94.3

IPW Unconstrained ML Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	56.9	-0.04	0.79	0.89	97.6	0.76	94.6
	1000	72.2	-0.02	0.54	0.60	97.6	0.54	94.7
	2000	83.3	-0.03	0.38	0.43	97.8	0.38	94.8
35-45	500	81.1	0.03	0.49	0.53	97.5	0.48	95.2
	1000	91.7	0.01	0.33	0.37	97.7	0.34	95.5
	2000	97.2	0.00	0.23	0.26	97.3	0.24	94.9
65-75	500	96.3	-0.02	0.38	0.40	96.2	0.37	94.7
	1000	99.7	0.02	0.26	0.28	96.6	0.26	94.5
	2000	100	-0.01	0.19	0.20	95.5	0.19	94.3

Complete-Case Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	100	0.91	0.90			0.87	94.0
	1000	100	0.74	0.59			0.60	93.7
	2000	100	0.73	0.43			0.42	89.9
35-45	500	100	1.08	0.55			0.52	89.4
	1000	100	0.98	0.37			0.37	84.5
	2000	100	0.99	0.26			0.26	73.1
65-75	500	100	1.08	0.41			0.40	83.3
	1000	100	1.10	0.28			0.28	72.3
	2000	100	1.06	0.20			0.20	52.1

Full Data Maximum Likelihood

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	100	0.02	0.35			0.34	94.2
	1000	100	-0.03	0.23			0.24	95.4
	2000	100	-0.03	0.17			0.17	94.6
35-45	500	100	0.01	0.35			0.34	94.6
	1000	100	-0.01	0.24			0.24	94.8
	2000	100	-0.01	0.16			0.16	94.7
65-75	500	100	0.01	0.33			0.33	94.5
	1000	100	0.02	0.23			0.23	94.9
	2000	100	-0.01	0.17			0.16	94.2

Table 9: Estimation for effect of second confounder $\alpha_3 = 0.2$ in the logistic regression model from 1000 simulation replicates. Posterior mean is used as estimate in constrained Bayesian estimation. Results for unconstrained maximum likelihood estimation are restricted to those runs that did converge. Sandwich variance estimators are given in (22). The asymptotic variance estimators are given in (20) and (21). For complete-case analysis and full data maximum likelihood, the asymptotic variance estimates are obtained via Fisher scoring. Coverages are based on nominal 95% Wald confidence intervals.

IPW Bayesian Constrained Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	100	0.98	0.88	0.89	94.5	0.82	92.2
	1000	100	0.37	0.55	0.60	97.3	0.55	95.4
	2000	100	0.09	0.39	0.42	96.4	0.38	94.8
35-45	500	100	-0.14	0.53	0.52	95.0	0.51	94.3
	1000	100	-0.10	0.36	0.37	96.0	0.36	94.8
	2000	100	-0.03	0.26	0.26	94.9	0.25	94.5
65-75	500	100	-0.05	0.37	0.39	96.6	0.39	95.8
	1000	100	-0.02	0.28	0.27	94.4	0.27	94.3
	2000	100	0.00	0.19	0.19	94.8	0.19	94.6

IPW Unconstrained ML Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	56.9	-0.12	0.83	0.87	95.8	0.78	93.4
	1000	72.2	-0.11	0.54	0.60	97.6	0.54	95.9
	2000	83.3	-0.10	0.38	0.42	96.1	0.38	95.1
35-45	500	81.1	-0.08	0.53	0.52	95.1	0.50	94.5
	1000	91.7	-0.07	0.36	0.37	95.7	0.36	94.6
	2000	97.2	-0.04	0.26	0.26	94.7	0.25	94.3
65-75	500	96.3	-0.04	0.37	0.39	96.7	0.38	94.7
	1000	99.7	-0.01	0.28	0.28	94.4	0.27	94.2
	2000	100	0.00	0.19	0.19	94.7	0.19	94.5

Complete-Case Estimation

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	100	2.14	0.89			0.86	92.1
	1000	100	1.90	0.58			0.58	89.9
	2000	100	1.79	0.41			0.41	86.2
35-45	500	100	-0.57	0.53			0.51	94.6
	1000	100	-0.60	0.36			0.36	94.1
	2000	100	-0.56	0.26			0.25	91.5
65-75	500	100	-0.14	0.37			0.39	96.5
	1000	100	-0.11	0.27			0.27	94.5
	2000	100	-0.10	0.19			0.19	94.2

Full Data Maximum Likelihood

% CC	n	%Converge	Rel. Bias	MC SE	Sandwich SE	% Coverage	Asymp. SE	% Coverage
15-25	500	100	0.01	0.34			0.33	94.4
	1000	100	0.01	0.24			0.23	94.6
	2000	100	-0.03	0.16			0.16	94.8
35-45	500	100	0.02	0.33			0.33	94.9
	1000	100	0.00	0.23			0.23	94.7
	2000	100	-0.02	0.17			0.16	94.6
65-75	500	100	-0.03	0.33			0.33	95.3
	1000	100	-0.01	0.23			0.23	94.1
	2000	100	0.00	0.16			0.16	94.6