Identification and Estimation for Nonignorable Missing Data: A Data Fusion Approach

Zixiao Wang [1] **AmirEmad Ghassami** [2] **Ilya Shpitser** [3]

Abstract

We consider the task of identifying and estimating a parameter of interest in settings where data is missing not at random (MNAR). In general, such parameters are not identified without strong assumptions on the missing data model. In this paper, we take an alternative approach and introduce a method inspired by data fusion, where information in the MNAR dataset is augmented by information in an auxiliary dataset subject to missingness at random (MAR). We show that even if the parameter of interest cannot be identified given either dataset alone, it can be identified given pooled data, under two complementary sets of assumptions. We derive inverse probability weighted (IPW) estimators for identified parameters under both sets of assumptions, and evaluate the performance of our estimation strategies via simulation studies, and a data application.

1. Introduction

Missing data is a pervasive and challenging issue in various applications of statistical inference, such as healthcare, economics, and the social sciences. Data are said to be missing at random (MAR) when the mechanism of missingness depends only on the observed data. Strategies to deal with MAR have been extensively investigated in the literature (Dempster et al., 1977; Robins et al., 1994; Tsiatis, 2006; Little & Rubin, 2019). In many practical settings, MAR is not a realistic assumption. Instead, missingness often depends on variables that are themselves missing or unobserved. Such settings are said to exhibit nonignorable missingness, with the resulting data being missing not at random (MNAR) (Fielding et al., 2008; Schafer & Graham, 2002). A classic example of a scenario with MNAR data occurs in longitudinal studies, where, due to the treatment's toxicity, some patients may become too ill to visit the clinic, leading to the situation where the outcome of certain patients with circumstances associated with those outcomes are more likely to be lost to follow-up (Ibrahim et al., 2012).

Existing MNAR models typically impose constraints on target distribution and its missingness mechanism, ensuring the parameter of interest can be identified. This approach goes back to the work of (Heckman, 1979), who proposed an outcome-selection model based on parametric modeling of outcome variable and missing pattern. (Little, 1993) introduced the pattern-mixture model where one needs to specify the distribution for each missing data pattern independently. Other related work involves permutation model (Robins, 1997), the discrete choice model (Tchetgen Tchetgen et al., 2018), the block-sequential MAR model (Zhou et al., 2010), the no self-censoring (NSC) model (Shpitser, 2016; Sadinle & Reiter, 2017; Malinsky et al., 2021), the instrumental variable approaches (Miao et al., 2015; Tchetgen Tchetgen & Wirth, 2017), and approaches based on graphical models (Mohan et al., 2013; Bhattacharya et al., 2019; Nabi et al., 2020) just to name a few.

In some applications where MNAR data is present, researchers may have access to additional auxiliary data on informative variables that are themselves not missing, or missing given a simpler missingness process. Hence, it is of interest to investigate combining the information in the available datasets. Such a *data fusion* strategy is natural in many applications: for example, surveys of HIV patients containing sensitive questions (such as those on sexual history) with high degree of missingness may be augmented with other sources of information with simpler missingness mechanisms, such as electronic health records. Similarly, our methods are inspired by settings where large, poorly structured datasets are enhanced by smaller, well-curated datasets, such as combining electronic health record data on patients, which often exhibits complex non-ignorable missingness patterns, with observational study data on similar patients, where missingness is like at random due to standard study attrition.

[1]Department of Biostatistics Johns Hopkins University, Baltimore, MD [2]Department of Mathematics and Statistics, Boston University, Boston, MA [3]Department of Computer Science, Johns Hopkins University, Baltimore, MD. Correspondence to: Zixiao Wang <zwang383@jh.edu>, AmirEmad Ghassami <ghassami@bu.edu>, Ilya Shpitser <ilyas@cs.jhu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

In this paper, we demonstrate that a parameter of interest may be identified given MNAR data in a primary domain, and data from an auxiliary domain, even though the same parameter is not identified from data from either domain alone. In the primary dataset, information on our target variable of interest as well as other variables in the model is only partially observed with an MNAR mechanism, which is more commonly encountered in real-world data. In the auxiliary dataset, the target variable of interest is fully unobserved, and the rest of the variables exhibit missingness with a MAR mechanism. Using the language of graphical models, we propose two complementary models of the missing mechanisms in the MNAR domain that allow different causal relations among the variables. Specifically, in the proposed Model 1, we consider the case where the missingness of the outcome is directly associated with another variable that may itself be potentially missing. In the proposed Model 2, we consider the case where the missingness is directly associated with a potentially missing outcome itself. We introduce a novel data fusion technique that combines information from the primary and the auxiliary datasets to achieve identifiaction of the parameter of interest. We illustrate our method by estimating the hospitalization rate in New York during the initial shortage stage of the COVID-19 pandemic in March 2020 where hospitalization status was often unrecorded with a complex censoring mechanism, with the auxiliary data from March 2023, a period characterized by improved conditions and thus reduced level of missingness with a simpler censoring mechanism.

Similar data fusion approaches exist in the literature of causal inference considering settings where the presence of unobserved confounders in the system may render the causal effect of a treatment on an outcome variable unidentified. For instance, works such as (Athey et al., 2016; 2020; Ghassami et al., 2022; Imbens et al., 2022) demonstrated that while information is often missing for long-term effects in randomized trials (Bouguen et al., 2019), auxiliary observational studies may contain information on long-term effects of treatments that can render the causal effect identified. Note that this is fundamentally different from the case that the parameter is identified in at least one of the domains and the purpose of combining datasets is improvement in estimation efficiency, as opposed to identification (see, for example, (Kallus & Mao, 2020)). To the best of our knowledge, our work is the first to pursue data fusion for the purpose of identification in an MNAR setting.

The rest of the paper is organized as follows. In Section 2, we introduce the concepts of missing data and directed acyclic graphs (DAGs), providing an overview of the problem setup and the parameters of interest. In Section 3, we describe the assumption of our data fusion setting and present an identification approach from the pooled data under two complementary sets of assumptions. Graphical models are used to illustrate the idea of data fusion and missing mechanisms. For estimation, we propose an inverse probability weighted (IPW) estimator for the target parameter in both models, discussed in Section 4. The study of the performance of the proposed estimators and a comparison with a MAR estimator and a multiple imputation by chained equations (MICE) approach (van Buuren & Groothuis-Oudshoorn, 2011) via a series of simulations is presented in Section 5. We describe the COVID-19 application analysis in Section 6. We conclude and discuss future work in Section 7. All the proofs are provided in the Appendix.

2. Preliminaries

Missing Data. We consider a missing data model which is a collection of distributions that are defined over a set of random variables $\{X, R, Y^{(1)}, Y, M^{(1)}, M\}$. In this context, X represents a set of covariates that are always observed, $Y^{(1)}$ and $M^{(1)}$ represent the underlying outcome variable of interest and another covariate (or set of covariates), respectively, that could be potentially missing, R is a binary indicator variable of missingness for the variables $Y^{(1)}$ and $M^{(1)}$, and Y and M represent the observed versions of $Y^{(1)}$ and $M^{(1)}$: when $R = 1$, the corresponding observed variables are $Y \equiv Y^{(1)}$ and $M \equiv M^{(1)}$, when $R = 0$, $Y =$ "?" and $M =$ "?". The full distribution (law) of a missing data model is $p(X, Y^{(1)}, M^{(1)}, R)$, and it is generally partitioned into two pieces: the target distribution $p(X, Y^{(1)}, M^{(1)})$ and the missingness mechanism $p(R|X, Y^{(1)}, M^{(1)})$. While the target distribution consists of potentially missing random variables, the missingness mechanism denotes the patterns exhibited by missingness indicators given the observed and missing variables. The observed distribution is $p(X, R, Y, M)$. When dealing with missing data problems, the objective is to obtain estimations or inferences about functions of variables in $\{X, Y^{(1)}, M^{(1)}\}$ based on the observed data.

Directed Acyclic Graphs (DAGs). Several widely used missing data models (Robins, 1997; Shpitser, 2016; Miao et al., 2015; Zhou et al., 2010) can be thought of as different factorizations of the complete data distribution and represented by DAGs (Mohan et al., 2013). A DAG, denoted by $\mathcal{G}(V)$, is a graph with a vertex set V connected by direct edges, ensuring that no cycles exist within the structure. Statistical models associated with a DAG $\mathcal{G}$ entail probability distributions that factorize as $p(V) = \prod_{V_i \in V} p\left(V_i \mid \text{pa}_{\mathcal{G}}(V_i)\right)$, where $\text{pa}_{\mathcal{G}}(V_i)$ are the parents of node V_i within the DAG $\mathcal{G}$. Conditional independences in any distribution obeying the above factorization may be read off via the d-separation criterion (Pearl, 1988). Graphical models that allow for context-specific dependence structures have been considered in (Nyman et al., 2014) and will be

employed in our work.

Problem Setup. We consider a setting with two available datasets, a primary dataset and an auxiliary dataset. The primary dataset is drawn from a primary domain referred to as Domain 1, where the outcome whose mean we would like to estimate is observed but potentially missing not at random. The auxiliary dataset is drawn from an auxiliary domain referred to as Domain 2, where the outcome is not recorded, but an auxiliary variable set is recorded, but potentially missing at random.

Let G be the binary indicator of the domain: $G = 1$ indicates data in Domain 1 which is MNAR, and $G = 2$ indicates data in the auxiliary MAR domain. Let R_1, R_2 denote indicators for missingness in Domain 1 and Domain 2, respectively. Therefore, observed variables in the Domain 2 are $\{G = 2, X, M, R_2\}$. In the Domain 1, we can observe $\{G = 1, X, M, R_1, Y\}$, where the missing indicator R_1 is subject to potentially missing variables $\{M^{(1)}, Y^{(1)}\}$, indicating that M, Y are missing not at random. R_1 is the missing indicator for both $M^{(1)}$ and $Y^{(1)}$, i.e., if $R_1 = 0$, $\{M^{(1)}, Y^{(1)}\}$ will be missing at the same time, and therefore $M =$ "?", $Y =$ "?". The cause of missingness in Domain 1 can be either $M^{(1)}$, or the outcome itself $Y^{(1)}$; we discuss the identification and estimation of these two cases in what we will call Model 1 and Model 2, respectively. In the pooled dataset containing the collection of random variables $\{G, X, M, R, Y\}$, where $R = R_1$ if $G = 1$ and $R = R_2$ if $G = 2$.

Estimands. Our target of inference is the mean of a (potentially missing) outcome $Y^{(1)}$ in the primary domain (denoted by $G = 1$). That is, the parameter of interest in this work is

$$\beta = \mathbb{E}[Y^{(1)} \mid G = 1].$$

As mentioned earlier, the outcome variable is exclusively present in the primary domain (Domain 1) where we have identification challenges due to MNAR condition. In order to overcome this obstacle, we utilize information from the auxiliary domain (Domain 2) to construct a framework for identification and estimation.

3. Identification

In this section, we examine two complementary sets of assumptions pertaining to the identification of Model 1 and Model 2. Subsequently, we establish the corresponding identification theorems. Our discussion commences with assumptions shared by both models in Section 3.1, followed by a separate exploration of additional assumptions and identification for each model. Before proceeding further, we note that our assumptions lead to graphical models of missing data spanning two domains, with the conditional independences defining the model illustrated (via the d-separation criterion) in the graphs shown in Fig. 1. Note that any graphical model implying the same conditional independence restrictions as these figures is also consistent with our model.

3.1. Data Fusion Assumptions

The full data distribution in the MNAR dataset (Domain 1), denoted as $p(X, M^{(1)}, Y^{(1)}, R_1, G = 1)$, cannot be deduced solely from the observed data distribution $p(X, M, Y, R_1, G = 1)$, unless certain constraints are placed on the mechanism responsible for data missingness. In our analysis, we avoid positing strong assumptions for identification and instead, leverage the information in Domain 2. In this domain, data is missing at random which is formalized in the following.

Assumption 1 (Auxiliary domain MAR). In auxiliary domain ($G = 2$), $M^{(1)}$ is missing at random, i.e., the missing indicator variable is independent of the potentially missing variable $M^{(1)}$ conditional on observed covariates X. That is,

$$M^{(1)} \perp\!\!\!\perp R \mid X, G = 2. \tag{1}$$

In order for us to be able to leverage the auxiliary domain, the information encoded in this domain must be relevant to the primary domain. Such a relevance requirement is usually stated by a *external validity*-type assumption in the literature (Hotz et al., 2005). We require the relevance of the information in the two domains in the form of the following *selection* assumption.

Assumption 2 (Selection at random). Given covariates X, the domain indicator G is conditionally independent of $M^{(1)}$. That is,

$$M^{(1)} \perp\!\!\!\perp G \mid X. \tag{2}$$

Assumption 2 limits the differences of $M^{(1)}$ between the two populations by requiring that conditioned on the rest of the covariates, it is as if units are randomly selected to belong to the domains. In practice, this assumption should be discussed and justified by domain experts when combining the datasets.

3.2. Model 1

In Model 1, we consider the scenario where the covariates X and the potentially missing variable $M^{(1)}$ contribute to the missingness in the primary domain, yet the outcome variable is not directly involved in the missingness mechanism, i.e., it is independent of the missing indicator conditioned on $M^{(1)}$ and X. This restriction is formulatted in the following Assumption 3.

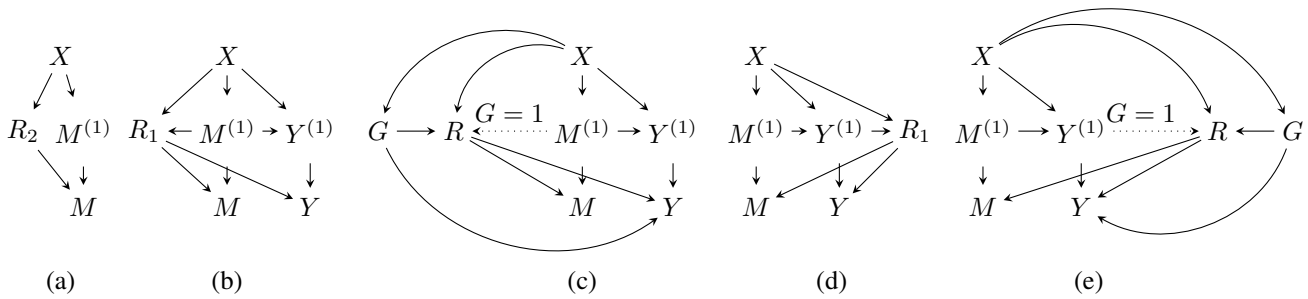


Figure 1. Graphical models: (a) The auxiliary MAR data domain in both Model 1 and Model 2. (b) Primary MNAR domain in Model 1. (c) The pooled data for Model 1, including the selection at random mechanism. (d) Primary MNAR domain in Model 2. (e) The pooled data for Model 2, including the selection at random mechanism. Notice the text $G = 1$ on the dotted arrow denotes a context-dependent relationship. Every vertex in the graph is assumed to have full support.

Assumption 3 (Primary domain MNAR). In primary domain ($G = 1$), the missing mechanism is independent of the outcome $Y^{(1)}$ given both $M^{(1)}$ and covariates X. That is,

$$Y^{(1)} \perp\!\!\!\perp R \mid X, M^{(1)}, G = 1. \tag{3}$$

The data fusion approach represents the two domains and the selection among them by a single graphical model is shown in Fig. 1: Fig. 1 (a), (b) and (c) show graphical representations for Domain 1, Domain 2 and the pooled dataset, respectively, which satisfy Assumptions 1 and 3. Fig. 1 (a) represents the graphical model in the Domain 2 (which is the pooled data set conditioned on G = 2). Notice that variable Y is absent in the second Domain. Fig. 1 (b) represents the graphical model in Domain 1 (which is the pooled data set conditioned on G = 1). X is the same list of covariates as that in Domain 1, and R_1 is the missing indicator for both $M^{(1)}, Y^{(1)}$ in Domain 1. We represent the graphical model of both domains pooled together in Fig. 1 (c), where the text $G = 1$ on the dotted arrow from $M^{(1)}$ to R denotes a conditional relationship that $M^{(1)}$ is parent of R if and only if $G = 1$. It is important to note that the presented graphical models in Fig. 1 are only one example of models that satisfy our assumptions. For instance, the edges between $Y^{(1)}$ and $M^{(1)}$ and X and G can be reversed without changing the model, and thus the identifying assumptions.

Theorem 1 (Identification in Model 1). Under Assumptions 1, 2, and 3, parameter $\beta = \mathbb{E}[Y^{(1)}|G = 1]$ is identified using the following functional

$$\mathbb{E}[\mathbb{E}[g_1(X, M) \mid X, G = 2, R = 1] \mid G = 1], \tag{4}$$

where $g_1(X, M) \equiv \mathbb{E}[Y|X, M, G = 1, R = 1]$.

The identifying functional involves three distributions: $p(Y \mid X, M, G = 1, R = 1)$, $p(M \mid X, G = 2, R = 1)$, and $p(X \mid G = 1)$. Notice that $p(X \mid G = 1)$ involves no potentially missing variable and is thus identified. Distributions $p(Y \mid X, M, G = 1, R = 1)$ and $p(M \mid X, G = 2, R = 1)$ are also identified as they are conditioned on $R = 1$. As a result, the mean of the potentially missing outcome variable in the first domain can be identified from the observational data.

3.3. Model 2

In Model 2, we consider a scenario where the potentially missing outcome $Y^{(1)}$ is directly associated with the missingness indicator in Domain 1. This necessitates assumptions complementary to those in Model 1 for the identification and estimation of the outcome mean. Inspired by the shadow variable approach in the study of MNAR data (Miao et al., 2015), we consider the following assumption.

Assumption 4. The potentially missing variable $M^{(1)}$ satisfies the following conditional independence requirements

$$M^{(1)} \perp\!\!\!\perp R \mid X, Y^{(1)}, G = 1, \tag{5}$$

$$M^{(1)} \not\perp\!\!\!\perp Y^{(1)} \mid X, R = 1, G = 1. \tag{6}$$

Assumption 4 formalizes the idea that the missingness process in Domain 1 may depend on $(X, Y^{(1)})$, but not on the potentially missing variable $M^{(1)}$ after conditioning on $(X, Y^{(1)})$. One notable distinction from the assumption in (Miao et al., 2015) is that the counterpart of variable $M^{(1)}$ in their setting, which is called the shadow variable, is fully observed, whereas our framework encompasses scenarios where this variable may potentially be missing. Furthermore, the outcome variable serves as both the cause of its own missingness and the missingness of the variable $M^{(1)}$. This, in turn, creates a situation characterized by not at random missingness.

The graphical model for Model 2 is shown in Fig. 1 (a), (d) and (e). The structure of Domain 2 of Model 2 (Fig. 1 (a)) is the same as that in Model 1. However, for Domain 1, a key difference is that $Y^{(1)}$ contributes to the missingness instead of $M^{(1)}$, as is shown in Figure 1 (d), while Figure 1 (e) shows both domains pooled together.

As in the shadow variable approach of (Miao et al., 2015), we leverage the odds ratio function to encode the deviation between the observed and missing data distributions in Domain 1, which is defined as

$$\begin{aligned} & OR\left(Y^{(1)}, X, M^{(1)}\right) \\ &= \frac{p\left(Y^{(1)} \mid R=0, X, M^{(1)}, G=1\right)}{p\left(Y^{(1)} \mid R=1, X, M^{(1)}, G=1\right)} \\ &\times \frac{p\left(Y^{(1)}=0 \mid R=1, X, M^{(1)}, G=1\right)}{p\left(Y^{(1)}=0 \mid R=0, X, M^{(1)}, G=1\right)}. \end{aligned} \tag{7}$$

The following proposition shows that certain parts of the full data distribution are identified under the assumptions in Model 2. These results will allow us to obtain identification of the target parameter.

Proposition 1 (Miao et al. (2015)). Under Assumption 4, for all $(X, Y^{(1)}, M^{(1)})$ in Domain 1, we have

$$\begin{aligned} OR\left(Y^{(1)}, X, M^{(1)}\right) &= OR\left(Y^{(1)}, X\right) \\ &= \frac{p\left(R=0 \mid X, Y^{(1)}, G=1\right)}{p\left(R=1 \mid X, Y^{(1)}, G=1\right)} \\ &\times \frac{p\left(R=1 \mid X, Y^{(1)}=0, G=1\right)}{p\left(R=0 \mid X, Y^{(1)}=0, G=1\right)}, \end{aligned} \tag{8}$$

$$\begin{aligned} & p\left(Y^{(1)} \mid R=0, X, M^{(1)}, G=1\right) = \\ & \frac{p\left(Y^{(1)} \mid R=1, X, M^{(1)}, G=1\right) OR\left(X, Y^{(1)}\right)}{\mathbb{E}\left[OR\left(X, Y^{(1)}\right) \mid R=1, X, M^{(1)}, G=1\right]}, \end{aligned} \tag{9}$$

$$\begin{aligned} & p\left(R=1 \mid X, Y^{(1)}, G=1\right)^{-1} \\ &= 1+\frac{OR\left(X, Y^{(1)}\right) p\left(R=0 \mid X, Y^{(1)}=0, G=1\right)}{p\left(R=1 \mid X, Y^{(1)}=0, G=1\right)}, \end{aligned} \tag{10}$$

$$\begin{aligned} & p\left(R=1 \mid X, Y^{(1)}=0, G=1\right) = \\ & \frac{\mathbb{E}\left[OR\left(X, Y^{(1)}\right) \mid R=1, X, G=1\right]}{\mathbb{E}\left[OR\left(X, Y^{(1)}\right) \mid R=1, X, G=1\right]+\frac{p(R=0 \mid X, G=1)}{p(R=1 \mid X, G=1)}}, \end{aligned} \tag{11}$$

$$\begin{aligned} & \mathbb{E}\{\widetilde{OR}\left(X, Y^{(1)}\right) \mid R=1, X, M^{(1)}, G=1\} = \\ & \frac{p\left(M^{(1)} \mid X, R=0, G=1\right)}{p\left(M^{(1)} \mid X, R=1, G=1\right)}, \end{aligned} \tag{12}$$

where,

$$\widetilde{OR}\left(X, Y^{(1)}\right)=\frac{OR\left(X, Y^{(1)}\right)}{\mathbb{E}\{OR\left(X, Y^{(1)}\right) \mid R=1, X, G=1\}}. \tag{13}$$

We present the proof of these results in the Appendix. Equation (8) shows that the odds ratio function in the MNAR domain is only related to $\{X, Y^{(1)}\}$ under Assumption 4. Equation (9) shows that the missing data distribution of the outcome can be recovered by imposing odds ratio function and the complete case distribution. Equation (10) shows that the propensity of missingness given the outcome, $p\left(R=1 \mid X, Y^{(1)}, G=1\right)$, can be recovered by the odds ratio function and baseline propensity score $p\left(R=1 \mid X, Y^{(1)}=0, G=1\right)$, while the baseline propensity score depend on the odds ratio function with $p(R=1|X, G=1)$ obtained as stated in Equation (11).

In light of Proposition 1, it becomes evident that the crux of the matter lies in the identification of the odds ratio function. Equation (12) serves as a pivotal mathematical expression for $OR(X, Y^{(1)})$. With $p(M^{(1)}|X, R=0, G=1), p(M^{(1)}|X, R=1, G=1)$ and $p(Y^{(1)}|M^{(1)}, X, R=1,)$, Equation (12) is a Fredholm integral equation of the first kind with $\widetilde{OR}(X, Y^{(1)})$ to be solved for. However, the distribution of $p\left(M^{(1)} \mid X, R=0, G=1\right)$ cannot be observed directly from Domain 1. Therefore, to identify the missing distribution $p\left(M^{(1)} \mid X, R=0, G=1\right)$ effectively, we employ the observed distribution in Domain 2 by noticing that under Assumption 1 and 2, we have

$$\begin{aligned} & p(M^{(1)}|X, R=0, G=1) p(R=0|X, G=1) \\ &= p(M|X, R=1, G=2)-p(M, R=1, |X, G=1). \end{aligned} \tag{14}$$

As the distributions on the right-hand side are all observed, we conclude that $\widetilde{OR}(X, Y^{(1)})$ can be solved for, and hence, odds ratio $OR(X, Y^{(1)})$ can be obtained by the formula in Equation (13). Note that Equation (14) serves as a bridge, allowing us to leverage information from Domain 2 to recover the missing distribution in Domain 1. Yet, to guarantee unique identification of $OR(X, Y^{(1)})$, we need to guarantee the uniqueness of the solution for Equation (12). Hence, we assume the condition below stands.

Assumption 5 (Completeness of $p(Y|R=1, X, M, G=1)$). For all square integrable functions $h(X, Y^{(1)})$, we have $\mathbb{E}[h(X, Y^{(1)}) \mid R=1, X, M, G=1]=0$ almost surely if and only if $h(X, Y^{(1)})=0$ almost surely.

Theorem 2. (Identification in Model 2) Under Assumptions 1, 2, 4 and 5, odds ratio $OR(X, Y^{(1)})$ is uniquely identified and parameter $\beta=\mathbb{E}[Y^{(1)}|G=1]$ is identified using the following formula:

$$\begin{aligned} & \mathbb{E}[Y^{(1)}|G=1] \\ &= \sum_{y, m, x} y p(Y=y|X=x, R=1, G=1, M=m) \\ &\times \Bigg(p(M=m, X=x, R=1|G=1) \end{aligned}$$

$$+\frac{OR\left(X=x,Y^{(1)}=y\right)}{\mathbb{E}\left[OR\left(X,Y^{(1)}\right)\mid R=1,X=x,M=m,G=1\right]}$$
$$\times\Big[\frac{p(M=m|X=x,R=1,G=2)}{p(R=0|X=x,G=1)}$$
$$-\frac{p(M=m,R=1|X=x,G=1)}{p(R=0|X=x,G=1)}\Big]$$
$$\times p(X=x,R=0|G=1)\Bigg). \tag{15}$$

4. Estimation

In this section, we focus on the estimation aspect of the problem and propose estimation strategies for the identified functionals for our target parameter under Model 1 and Model 2. We propose inverse probability weighting (IPW) estimators for both models which rely on the propensity scores. In what follows, $\hat{\mathbb{E}}$ denotes the empirical mean operator.

4.1. Model 1

Let $q(X,M^{(1)})=1/p(R=1\mid X,M^{(1)},G=1)$. Our estimation strategy for the parameter in Equation (4) is as follows.

Proposition 2. Under Assumptions 1, 2, and 3, a correctly specified working model with parameters α, $p(R=1\mid X,M^{(1)},G=1;\alpha)$; and the regularity conditions for estimating equations described by (Newey & McFadden, 1994), using a user-specified vector function $h(X,M^{(1)})$, the IPW estimator, denoted by β^{IPW}, obtained by solving the estimating equations below is consistent.

$$\hat{\mathbb{E}}\left[\left(q(X,M^{(1)};\hat{\alpha})R\cdot h(X,M^{(1)})\right)\mid G=1\right], \tag{16}$$
$$-\hat{\mathbb{E}}\left[\hat{\mathbb{E}}\left[h(X,M^{(1)})\mid X,R=1,G=2\right]\mid G=1\right]=0$$
$$\hat{\mathbb{E}}\left[q(X,M^{(1)};\hat{\alpha})R\cdot Y^{(1)}-\hat{\beta}_{IPW}|G=1\right]=0. \tag{17}$$

4.2. Model 2

We next consider the estimation of the parameter of interest in Model 2. In the context of the shadow variable configuration, (Miao et al., 2015) formerly introduced an estimator relying on odds ratio and baseline propensity score in MNAR data. We extend that strategy here by leveraging information from Domain 2 to compensate for the fact that variable $M^{(1)}$ might be missing.

Our estimation strategy for the parameter in Equation (15) is as follows.

Proposition 3. Under Assumptions 1, 2, 4 and 5, a working model for odds ratio function $OR(X,Y^{(1)};\gamma)$ and the baseline propensity score $p(R=1|X,Y^{(1)}=0,G=1;\alpha)$, we can recover $w(X,Y^{(1)},G=1;\alpha,\gamma)=1/p(R=1|X,Y^{(1)},G=1;\alpha,\gamma)$ using Equation (10). With the regularity conditions for estimating equations described by (Newey & McFadden, 1994) and a user-specified vector function $h(X,M^{(1)})$, the IPW estimator, denoted by β^{IPW}, obtained by solving the estimating equations below is consistent.

$$\hat{\mathbb{E}}\left[\left(w(X,Y^{(1)};\hat{\alpha},\hat{\gamma})R\cdot h(X,M^{(1)})\right)\mid G=1\right] \tag{18}$$
$$-\hat{\mathbb{E}}\left[\hat{\mathbb{E}}\left[h(X,M^{(1)})\mid X,R=1,G=2\right]\mid G=1\right]=0,$$
$$\hat{\mathbb{E}}\left[(w(X,Y^{(1)};\hat{\alpha},\hat{\gamma})R\cdot Y^{(1)}-\hat{\beta}_{IPW}|G=1\right]=0. \tag{19}$$

5. Simulation Studies

We investigate the performance of the proposed framework in Section 4 to estimate the outcome means in the primary domain, $\beta=\mathbb{E}[Y^{(1)}\mid G=1]$, through a comprehensive series of simulation experiments. For each model, we generate data consistent with the correctly specified working model in estimation (T), and a misspecified one (F) to check robustness. In both models, to streamline the estimation process, we notice that flipping the edge from X to G does not alter the underlying graph statistical model.

5.1. Simulation of Model 1

We first generate a binary grouping variable G with n observations. Each observation in G is randomly assigned a value of 1 or 2 with equal probability, dividing the dataset into Domain 1 ($G=1$) and Domain 2 ($G=2$).

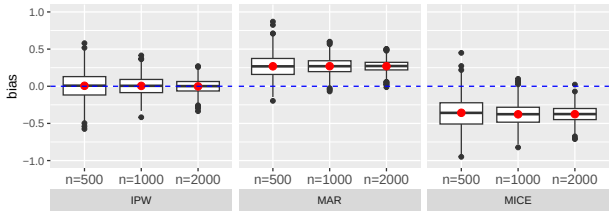
For the auxiliary domain ($G=2$), we assume the following models for covariates X, variable $M^{(1)}$, and the propensity score:

$$X\mid G=2\sim\mathcal{N}(0,1)$$
$$M^{(1)}\mid X,G=2\sim\mathcal{N}\left(0.4X^2,1\right)$$
$$p(R=1\mid X,G=2)=logis(1.4+X)$$

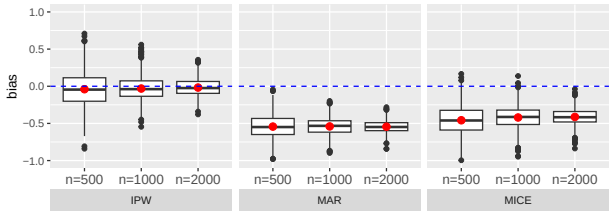
where $logis(x)=(1+\exp(-x))^{-1}$. Under this setting, the missing data proportion of M in Domain 2 is between 50% and 60%.

For primary domain ($G=1$), we posit varying distributions for the covariates X, while maintaining the same distribution for $M^{(1)}$ given X, satisfying Assumption 2. Moreover, we assume the distributions for $Y^{(1)}$ is a function of both X and $M^{(1)}$ by Assumption 3:

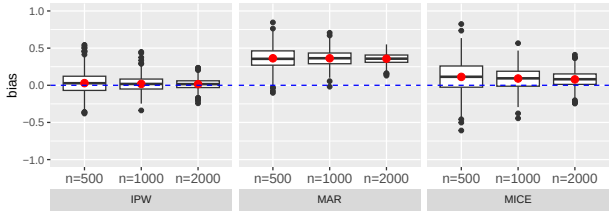
$$X\mid G=1\sim\mathcal{N}(1,1)$$
$$M^{(1)}\mid X,G=1\sim\mathcal{N}\left(0.4X^2,1\right)$$
$$Y^{(1)}\mid X,M^{(1)},G=1\sim\mathcal{N}\left(X+M^{(1)},1\right)$$
$$p(R=1\mid X,M^{(1)},G=1)=logis(0.3+0.1X+M^{(1)})$$



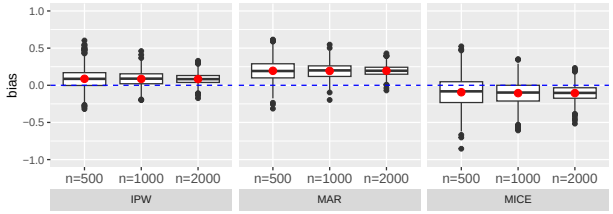
(a) Model1 - T: model is correctly specified.



(b) Model1 - F: model is misspecified.



(c) Model2 - T: model is correctly specified.



(d) Model2 - F: model is misspecified.

Figure 2. Simulation results for Model 1 and Model 2: Bias for estimation of $\mathbb{E}[Y^{(1)} \mid G = 1]$. Boxplots of correct and misspecified settings, calculated from 1000 trials at sample sizes $n \in \{500, 1000, 2000\}$. The red point indicates the mean. Statistics of boxplots are in Table 3 and Table 4. (a) MAR are clearly biased upwards while MICE estimates are biased downwards. (b) IPW estimates, though slightly biased, concentrate around the true value as sample size increases. (c) IPW estimates are less biased than MAR and MICE. (d) IPW estimates are less biased than MAR and MICE estimates.

Thus, the missing data proportion is between 20% and 40%.

For our IPW estimation approach, we specify a working model for $p(R = 1 \mid X, M^{(1)}, G = 1) = logis(\alpha_0 + \alpha_1 X + \alpha_2 M^{(1)})$, and a set of correctly specified estimation equations for $\mathbb{E}[M^{(1)} \mid X, R = 1, G = 2]$. Considering total sample sizes $n \in \{500, 1000, 2000\}$, we summarize the results using boxplot in Fig. 2a. We generate additional misspecified estimation approach by replacing with $p(R = 1 \mid X, M^{(1)}, G = 1) = logis(0.3 + 0.1X - [M^{(1)}]^2)$, and summarized in Fig. 2b.

For each scenario, we compare the result of our IPW estimator with a naive estimator assuming Missing at Random (MAR), which is derived through linear regression for Y given X on the complete cases, and an estimator from the completely imputed dataset using MICE.

For a specific pooled dataset, we used bootstrapping of size $k = 1000$ to report correctly specified IPW estimation result and 95% confidence interval (95% CI), as shown in Table 1. A detailed comparison with MAR and MICE estimators for both correctly (T) and incorrectly specified model is presented in Table 5 in the Appendix.

5.2. Simulation of Model 2

The data generating process for Model 2 is slightly different due to the factorization. We started by generating the binary grouping variable G as that for Model 1. For primary domain ($G = 1$), we generate a covariate $X|G = 1 \sim N(0, 1)$,

Table 1. Bootstrap confidence intervals for Model 1 (T) (True value of $\beta = 1.8$).

n	Est.	95% CI	Width	Bias
500	1.941	[1.594, 2.275]	0.681	0.141
1000	1.744	[1.508, 1.970]	0.462	-0.056
2000	1.767	[1.598, 1.930]	0.332	-0.033

Table 2. Bootstrap confidence intervals for Model 2 (T). (True value of $\beta = -0.659$).

n	Est.	95% CI	Width	Bias
500	-0.557	[-0.815, -0.303]	0.513	0.102
n=1000	-0.760	[-0.949, -0.569]	0.380	-0.101
n=2000	-0.707	[-0.841, -0.561]	0.281	-0.048

and then generate (Y, M, R) as shown below:

$$
\begin{aligned}
&M \mid R = 1, X, G = 1 \sim \mathcal{N}\left(-0.4X^2, 1\right) \\
&Y \mid R = 1, X, M, G = 1 \sim \mathcal{N}\left(X + M, 1\right) \\
&\operatorname{logit} p(R = 1 \mid X, Y^{(1)} = 0, G = 1) = 0.5 + 0.4X \\
&\mathrm{OR}(X, Y^{(1)}; G = 1) = \exp(-0.3Y^{(1)})
\end{aligned}
$$

For the auxiliary domain ($G = 2$), we followed the same generating process for $\{X, M^{(1)}\}$ as above, but a different missing mechanism, $p(R = 1 \mid G = 2, X) = logis(X)$.

In these specified conditions, the missing rate is between 40% and 50% in both domains. We employed correctly spec-

ified estimation equations for $\mathbb{E}[M^{(1)} \mid X, R = 1, G = 2]$, and the above working models in Domain 2. We also generate a misspecified setting (F) by replacing with $logit[p(R = 1|Y^{(1)} = 0, X, G = 1)] = 0.5 + 0.4X + 0.4X^2$ but always impose first-order linear regression. We simulate 1000 replicates, with sample sizes $n \in \{500, 1000, 2000\}$. We present the results using boxplots with a comparison of an MAR estimator, a MICE estimator in Fig. 2c and Fig. 2d, and summarize the statistics of boxplots in in Appendix Table 4. Bootstrapping result is reported in Table 6.

6. Application to COVID-19 case data

In this section, we employ both Model 1 and Model 2 to analyze COVID-19 case data in New York State (NYS), focusing on estimating the hospitalization rate during the initial wave of the pandemic around March 2020.

Due to surge of COVID-19 cases, the situation in NYS hospitals in March 2020 was critical, with reported shortages, and a triage system for patient care, leaving some without adequate care (Schmitt-Grohé et al., 2020; Watkins, 2020). By March 2023, COVID-19 was well on the way to endemicity, and additional resources had been committed to COVID-19 patient care. Both domains exhibit missingness in hospitalization status. However, in our analysis, we assume March 2020 data to exhibit complex MNAR missingness due to systemic early difficulties with COVID-19 respnose, while later data from March 2023 to exhibit a more manageable MAR missingness.

6.1. Data

The data source is the COVID-19 Case Surveillance Public Use Data with Geography maintained by (Centers for Disease Control and Prevention, COVID-19 Response, 2024). New York State was selected as the target population, with data in March 2020 as the primary domain and in March 2023 as the auxiliary domain. We applied filtering procedures as detailed in the Appendix B.4, resulting in a sample size of 78119 in March 2020 and 38237 in March 2023.

In this dataset, let X represent the patient's county, recognizing that spatial information plays a crucial role in accounting for the spread of infectious diseases, Y be a binary indicator of whether the patient reported as 'Hospitalized' and M be the race of the patient. The missing rate is 70% in March 2020 and 46% in March 2023.

6.2. Results

It is challenging to determine whether the cause of missing data is tied to race or hospitalization status in this COVID-19 case report. Therefore, we investigated both the IPW estimator of Model 1 and the IPW estimator of Model 2 within the dataset. Additionally, we conducted a comparative analysis with Missing at Random (MAR) estimator, the Missing Completely at Random (MCAR) estimator, and results from MICE. Results using bootstrap with 1000 samplings are illustrated in Fig. 3. In Model 1, When attributing missing data to race, the estimated hospitalization rate (Est: 0.7396, 95% CI: $[0.7190, 0.7570]$) was found to be lower than the estimate (0.7533) derived from the MCAR analysis. Conversely, in Model 2, assuming that missing mechanisms were linked to hospitalization resulted in an estimated rate (Est: 0.7836, 95% CI: $[0.7558, 0.8071]$) surpassing the MCAR analysis. Same as Model 1, MAR and MICE also showed the MCAR result was overestimated (MAR: Est: 0.7337, 95% CI: $[0.7277, 0.7395]$, MICE: Est: 0.6564, 95% CI: $[0.6041, 0.7085]$). The significant difference in hospitalization rate estimates using the more naive complete case and MAR approaches compared to estimates using model 1 and model 2 underscores the importance of modeling choice in handling missing data in a principled way.

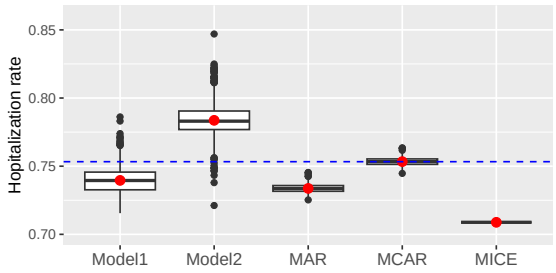


Figure 3. Boxplot of Bootstrap results (size = 1000) using IPW estimator in Model 1 (Est:0.7396, 95%CI: [0.7190,0.7570]), IPW estimator in Model 2 (Est:0.7836, 95%CI: [0.7558,0.8071]), MAR estimator (Est:0.7337, 95%CI: [0.7277,0.7395]), MCAR estimator(Est:0.7533, 95%CI: [0.7477,0.7590]) and MICE (Est:0.6564, 95% CI[0.6041,0.7085]). The statistical summary is calculated in Table 7 in the Appendix.

7. Discussion

In this paper, we introduced a data fusion approach to identification in settings where data is missing not at random (MNAR), but an auxiliary data that is missing at random (MAR) is available. We provided identification results under two complementary models in this setting, as well as a straightforward-to-implement Inverse Probability Weighting (IPW) estimators for the identified parameters in each model. We illustrated the consistency of our estimator via a simulation study. To our knowledge, our work is the first adoption of data fusion ideas for obtaining identification in settings where data is missing not at random (MNAR). We applied both models to the COVID-19 case data in New York state to estimate the hospitalization rate in March 2020. It should be noted that our methodology may extend to a broader class of target parameters, such as the conditional expectation of the outcome given covarites.

A natural extension of our approach is the development of a

semiparametrically efficient estimator under our proposed two models. We leave this extension to future work, as obtaining such an estimator is nontrivial. This is because both of our proposed models impose complex restrictions on the observed data tangent space, in addition to conditional independences implied by the graph.

Acknowledgements

We thank all reviewers who carefully checked the work and gave constructive comments. This work was supported in part by ONR N00014-21-1-2820, NSF CAREER 1942239 and NIH R01 AI127271-01A1. We thank Office of Naval Research (ONR), National Science Foundation (NSF) and National Institutes of Health (NIH).

Impact Statement

This paper aims to advance the area of missing data and data fusion. This work has the potential to improve quality of analyses of retrospective observational data.

References

Athey, S., Chetty, R., Imbens, G., and Kang, H. Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. *arXiv preprint arXiv:1603.09326*, 2016.

Athey, S., Chetty, R., and Imbens, G. Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*, 2020.

Bhattacharya, R., Nabi, R., Shpitser, I., and Robins, J. Identification in missing data models represented by directed acyclic graphs. In *Proceedings of the Thirty Fifth Conference on Uncertainty in Artificial Intelligence (UAI-35th)*. AUAI Press, 2019.

Bouguen, A., Huang, Y., Kremer, M., and Miguel, E. Using randomized controlled trials to estimate long-run impacts in development economics. *Annual Review of Economics*, 11:523–561, 2019.

Centers for Disease Control and Prevention, COVID-19 Response. Covid-19 case surveillance public use data with geography. Version date: January 04, 2024, 2024.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Fielding, S., Fayers, P. M., McDonald, A., McPherson, G., Campbell, M. K., and Group, R. S. Simple imputation methods were inadequate for missing not at random (mnar) quality of life data. *Health and Quality of Life Outcomes*, 6:1–9, 2008.

Ghassami, A., Yang, A., Richardson, D., Shpitser, I., and Tchetgen Tchetgen, E. Combining experimental and observational data for identification and estimation of long-term causal effects. *arXiv preprint arXiv:2201.10743*, 2022.

Heckman, J. J. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pp. 153–161, 1979.

Hotz, V. J., Imbens, G. W., and Mortimer, J. H. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of econometrics*, 125 (1-2):241–270, 2005.

Ibrahim, J. G., Chu, H., and Chen, M.-H. Missing data in clinical studies: issues and methods. *Journal of clinical oncology*, 30(26):3297, 2012.

Imbens, G., Kallus, N., Mao, X., and Wang, Y. Long-term causal inference under persistent confounding via data combination. *arXiv preprint arXiv:2202.07234*, 2022.

Kallus, N. and Mao, X. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*, 2020.

Little, R. J. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Malinsky, D., Shpitser, I., and Tchetgen Tchetgen, E. J. Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association*, pp. 1–9, 2021.

Miao, W., Liu, L., Tchetgen Tchetgen, E., and Geng, Z. Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *arXiv preprint arXiv:1509.02556*, 2015.

Mohan, K., Pearl, J., and Tian, J. Graphical models for inference with missing data. *Advances in neural information processing systems*, 26, 2013.

Nabi, R., Bhattacharya, R., and Shpitser, I. Full law identification in graphical models of missing data: Completeness results. In *International Conference on Machine Learning*, pp. 7153–7163. PMLR, 2020.

Newey, W. K. and McFadden, D. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4: 2111–2245, 1994.

Nyman, H., Pensar, J., Koski, T., and Corander, J. Stratified graphical models-context-specific independence in graphical models. 2014.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufmann, San Mateo, 1988.

Robins, J. M. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in medicine*, 16(1):21–37, 1997.

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

Sadinle, M. and Reiter, J. P. Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104(1):207–220, 2017.

Schafer, J. L. and Graham, J. W. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

Schmitt-Grohé, S., Teoh, K., and Uribe, M. Covid-19: testing inequality in new york city. Technical report, National Bureau of Economic Research, 2020.

Shpitser, I. Consistent estimation of functions of data missing non-monotonically and not at random. *Advances in Neural Information Processing Systems*, 29, 2016.

Tchetgen Tchetgen, E. J. and Wirth, K. E. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics*, 73(4): 1123–1131, 2017.

Tchetgen Tchetgen, E. J., Wang, L., and Sun, B. Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28 (4):2069, 2018.

Tsiatis, A. A. Semiparametric theory and missing data. 2006.

van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

Watkins, A. N.Y.C.'s 911 System Is Overwhelmed. 'I'm Terrified,' a Paramedic Says. *The New York Times*, March 28 2020. Archived from the original on March 31, 2020. Retrieved March 29, 2020.

Zhou, Y., Little, R. J., and Kalbfleisch, J. D. Block-conditional missing at random models for missing data. 2010.

A. PROOFS

Theorem 1 *Under Assumptions 1, 2, and 3, parameter $\beta = \mathbb{E}[Y^{(1)}|G=1]$ is identified using the following functional*

$$
\begin{aligned}
&\mathbb{E}[Y^{(1)}|G=1] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{E}[Y^{(1)}|X, G=1, M^{(1)}]|X, G=1]|G=1] \\
&= \sum_{y,m,x} y \cdot p_Y(Y^{(1)}=y|X=x, G=1, M^{(1)}=m) p_M(M^{(1)}=m|X=x, G=1) p(X=x|G=1) \\
&= \sum_{y,m,x} y \cdot p_Y(Y^{(1)}=y|X=x, G=1, M^{(1)}=m, R=1) p_M(M^{(1)}=m|X=x, G=2) p(X=x|G=1) \\
&= \sum_{y,m,x} y \cdot p_Y(y|X=x, G=1, M=m, R=1) p_M(m|X=x, G=2, R=1) p(X=x|G=1) \\
&= \mathbb{E}[\mathbb{E}[g_1(X,M) \mid X, G=2, R=1] \mid G=1]
\end{aligned}
$$

where $g_1(X,M) \equiv \mathbb{E}[Y|X,M,G=1,R=1]$.

Here, equality 1 holds by the Law of Iterated Expectation, equality 2 is by definition, equality 3 employed both Assumption 2 and Assumption 3, and equality 4 follows Assumption 1.

Theorem 2 *Under Assumption 1, 2,and 4, using equation (9), we have*

$$
\begin{aligned}
&\mathbb{E}[Y^{(1)}|G=1] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{E}[Y^{(1)}|M^{(1)}, X, R, G=1]|X, R, G=1]|G=1] \\
&= \sum_{y,m,x,r} y p(Y^{(1)}=y|X=x, R=r, G=1, M^{(1)}=m) p(M^{(1)}=m|X=x, R=r, G=1) p(X=x, R=r|G=1) \\
&= \sum_{y,m,x} y p(Y^{(1)}=y|M^{(1)}=m, X=x, R=1, G=1) p(M^{(1)}=m|X=x, R=1, G=1) p(X=x, R=1|G=1) \\
&+ \sum_{y,m,x} y p(Y^{(1)}=y|M^{(1)}=m, X=x, R=0, G=1) p(M^{(1)}=m|X=x, R=0, G=1) p(X=x, R=0|G=1) \\
&= \sum_{y,m,x} y p(Y^{(1)}=y|M^{(1)}=m, X=x, R=1, G=1) p(M^{(1)}=m|X=x, R=1, G=1) p(X=x, R=1|G=1) \\
&+ \sum_{y,m,x} y \frac{p\left(Y^{(1)}=y \mid M^{(1)}=m, X=x, R=1, G=1\right) OR\left(X=x, Y^{(1)}=y\right)}{\mathbb{E}\left[OR\left(X, Y^{(1)}\right) \mid R=1, X=x, M^{(1)}=m, G=1\right]} \\
&\times \frac{p(M^{(1)}=m|X=x, R=1, G=2) - p(M^{(1)}=m, R=1|X=x, G=1)}{p(R=0|X=x, G=1)} \\
&\times p(X=x, R=0|G=1) \\
&= \sum_{y,m,x} y p(Y^{(1)}=y|X=x, R=1, G=1, M^{(1)}=m) \\
&\times \Bigg(p(M^{(1)}=m, X=x, R=1|G=1) \\
&+ \frac{OR\left(X=x, Y^{(1)}=y\right)}{\mathbb{E}\left[OR\left(X, Y^{(1)}\right) \mid R=1, X=x, M^{(1)}=m, G=1\right]} \\
&\times \left[\frac{p(M^{(1)}=m|X=x, R=1, G=2)}{p(R=0|X=x, G=1)} - \frac{p(M^{(1)}=m, R=1|X=x, G=1)}{p(R=0|X=x, G=1)}\right] \times p(X=x, R=0|G=1) \Bigg)
\end{aligned}
$$

where the solution of $\widetilde{OR}(X, Y^{(1)})$ in equation (12) is guaranteed by Assumption 5, therefore the solution for $OR(X, Y^{(1)})$ exists by definition.

Here, equality 1 follows by the Law of Iterated Expectation, equality 2 is the definition, equality 3 expands for separating cases $R = 0, 1$, equality 4 uses equation (9) and equation (14) to denote the missing distribution accordingly.

Note that equation (12) can be written as

$$\begin{aligned}
&\mathbb{E}\{\widetilde{OR}\left(X, Y^{(1)}\right) \mid R=1, X, M^{(1)}, G=1\} \\
&= \frac{p\left(M^{(1)} \mid X, R=0, G=1\right)}{p\left(M^{(1)} \mid X, R=1, G=1\right)} \\
&= \frac{p(M|X, R=1, G=2) - p(M|X, R=1, G=1)p(R=1|X, G=1)}{p(R=0|X, G=1)p\left(M^{(1)} \mid X, R=1, G=1\right)}
\end{aligned}$$

by equation (14), which stands since

$$\begin{aligned}
&p(M^{(1)}|X, R=0, G=1)p(R=0|X, G=1) \\
&= p(M^{(1)}|X, G=1) - p(M^{(1)}|X, R=1, G=1)p(R=1|X, G=1) \\
&= p(M^{(1)}|X, G=2) - p(M^{(1)}|X, R=1, G=1)p(R=1|X, G=1) \\
&= p(M^{(1)}|X, R=1, G=2) - p(M^{(1)}, R=1|X, G=1)
\end{aligned} \tag{20}$$

Here, equality 1 is by definition as R is a binary indicator, equality 2 is by Assumption 2, and equality 3 is by Assumption 1.

Proposition 1 *[(Miao et al., 2015)] In the interest of being self-contained, we proved here under Assumption 4, for all $(X, Y^{(1)}, M^{(1)})$ in Domain 1, we have the following properties,*

For equation (8), we note that

$$\begin{aligned}
OR\left(X, Y^{(1)}, M^{(1)}\right) &= \frac{p\left(Y^{(1)} \mid R=0, X, M^{(1)}, G=1\right) p\left(Y^{(1)}=0 \mid R=1, X, M^{(1)}, G=1\right)}{p\left(Y^{(1)} \mid R=1, X, M^{(1)}, G=1\right) p\left(Y^{(1)}=0 \mid R=0, X, M^{(1)}, G=1\right)} \\
&= \frac{p\left(Y^{(1)}, R=0, X, M^{(1)}, G=1\right) p\left(Y^{(1)}=0, R=1, X, M^{(1)}, G=1\right)}{p\left(Y^{(1)}, R=1, X, M^{(1)}, G=1\right) p\left(Y^{(1)}=0, R=0, X, M^{(1)}, G=1\right)} \\
&= \frac{p\left(R=0 \mid Y^{(1)}, X, M^{(1)}, G=1\right) p\left(R=1 \mid Y^{(1)}=0, X, M^{(1)}, G=1\right)}{p\left(R=1 \mid Y^{(1)}, X, M^{(1)}, G=1\right) p\left(R=0 \mid Y^{(1)}=0, X, M^{(1)}, G=1\right)} \\
&= \frac{p\left(R=0 \mid Y^{(1)}, X, G=1\right) p\left(R=1 \mid Y^{(1)}=0, X, G=1\right)}{p\left(R=1 \mid Y^{(1)}, X, G=1\right) p\left(R=0 \mid Y^{(1)}=0, X, G=1\right)} \\
&= OR\left(X, Y^{(1)}\right)
\end{aligned}$$

Equation (9) follows by observing that

$$
\begin{aligned}
&\mathbb{E}\left[OR\left(X, Y^{(1)}\right) \mid R=1, X, M^{(1)}, G=1\right] \\
&= \mathbb{E}\left[\frac{p\left(R=0 \mid Y^{(1)}, X, M^{(1)}, G=1\right) p\left(R=1 \mid Y^{(1)}=0, X, M^{(1)}, G=1\right)}{p\left(R=1 \mid Y^{(1)}, X, M^{(1)}, G=1\right) p\left(R=0 \mid Y^{(1)}=0, X, M^{(1)}, G=1\right)} \mid R=1, X, M^{(1)}, G=1\right] \\
&= \frac{p\left(R=1 \mid Y^{(1)}=0, X, M^{(1)}, G=1\right)}{p\left(R=0 \mid Y^{(1)}=0, X, M^{(1)}, G=1\right)} \mathbb{E}\left[\frac{p\left(R=0 \mid Y^{(1)}, X, M^{(1)}, G=1\right)}{p\left(R=1 \mid Y^{(1)}, X, M^{(1)}, G=1\right)} \mid R=1, X, M^{(1)}, G=1\right] \\
&= \frac{p\left(R=1 \mid Y^{(1)}=0, X, M^{(1)}, G=1\right) p\left(R=0, X, M^{(1)}, G=1\right)}{p\left(R=0 \mid Y^{(1)}=0, X, M^{(1)}, G=1\right) p\left(R=1, X, M^{(1)}, G=1\right)} \\
&\underbrace{\mathbb{E}\left[\frac{p\left(Y^{(1)} \mid R=0, X, M^{(1)}, G=1\right)}{p\left(Y^{(1)} \mid R=1, X, M^{(1)}, G=1\right)} \mid R=1, X, M^{(1)}, G=1\right]}_{=1} \\
&= \frac{p\left(Y^{(1)}=0 \mid R=1, X, M^{(1)}, G=1\right)}{p\left(Y^{(1)}=0 \mid R=0, X, M^{(1)}, G=1\right)}
\end{aligned}
$$

For equation (10),

$$
\begin{aligned}
&p\left(R=1 \mid X, Y^{(1)}, G=1\right)^{-1} \\
&= \frac{1}{p\left(R=1 \mid Y^{(1)}, X, G=1\right)} \\
&= \frac{p\left(R=0 \mid Y^{(1)}, X, G=1\right)+p\left(R=1 \mid Y^{(1)}, X, G=1\right)}{p\left(R=1 \mid Y^{(1)}, X, G=1\right)} \\
&= 1+\frac{p\left(R=0 \mid Y^{(1)}, X, G=1\right)}{p\left(R=1 \mid Y^{(1)}, X, G=1\right)} \\
&= 1+\frac{p\left(R=0 \mid Y^{(1)}, X, G=1\right) p\left(R=1 \mid Y^{(1)}=0, X, G=1\right)}{p\left(R=1 \mid Y^{(1)}, X, G=1\right) p\left(R=0 \mid Y^{(1)}=0, X, G=1\right)} \frac{p\left(R=0 \mid X, Y^{(1)}=0, G=1\right)}{p\left(R=1 \mid X, Y^{(1)}=0, G=1\right)} \\
&= 1+\frac{OR\left(X, Y^{(1)}\right) p\left(R=0 \mid X, Y^{(1)}=0, G=1\right)}{p\left(R=1 \mid X, Y^{(1)}=0, G=1\right)}
\end{aligned}
$$

Equation (11) stands as we first observed

$$
\begin{aligned}
&\mathbb{E}\left[OR\left(X, Y^{(1)}\right) \mid R=1, X, G=1\right] \\
&= \mathbb{E}\left[\frac{p\left(R=0 \mid Y^{(1)}, X, G=1\right) p\left(R=1 \mid Y^{(1)}=0, X, G=1\right)}{p\left(R=1 \mid Y^{(1)}, X, G=1\right) p\left(R=0 \mid Y^{(1)}=0, X, G=1\right)} \mid R=1, X, G=1\right] \\
&= \frac{p\left(R=1 \mid Y^{(1)}=0, X, G=1\right)}{p\left(R=0 \mid Y^{(1)}=0, X, G=1\right)} \mathbb{E}\left[\frac{p\left(R=0 \mid Y^{(1)}, X, G=1\right)}{p\left(R=1 \mid Y^{(1)}, X, G=1\right)} \mid R=1, X, G=1\right] \\
&= \frac{p\left(R=1 \mid Y^{(1)}=0, X, G=1\right) p(R=0, X, G=1)}{p\left(R=0 \mid Y^{(1)}=0, X, G=1\right) p(R=1, X, G=1)} \\
&\underbrace{\mathbb{E}\left[\frac{p\left(Y^{(1)} \mid R=0, X, G=1\right)}{p\left(Y^{(1)} \mid R=1, X, G=1\right)} \mid R=1, X, G=1\right]}_{=1} \\
&= \frac{p\left(Y^{(1)}=0 \mid R=1, X, G=1\right)}{p\left(Y^{(1)}=0 \mid R=0, X, G=1\right)}
\end{aligned}
$$

Then we note that

$$\begin{aligned}
&\frac{\mathbb{E}\left[OR\left(X,Y^{(1)}\right)\mid R=1,X,G=1\right]}{\mathbb{E}\left[OR\left(X,Y^{(1)}\right)\mid R=1,X,G=1\right]+p(R=0\mid X,G=1)/p(R=1\mid X,G=1)}\\
&=\frac{\frac{p\left(Y^{(1)}=0|R=1,X,G=1\right)}{p\left(Y^{(1)}=0|R=0,X,G=1\right)}}{\frac{p\left(Y^{(1)}=0|R=1,X,G=1\right)}{p\left(Y^{(1)}=0|R=0,X,G=1\right)}+\frac{p(R=0,X,G=1)}{p(R=1,X,G=1)}}\\
&=\frac{p\left(Y^{(1)}=0,R=1,X,G=1\right)}{p\left(Y^{(1)}=0,R=1,X,G=1\right)+p\left(Y^{(1)}=0,R=0,X,G=1\right)}\\
&=\frac{p\left(R=1\mid Y^{(1)}=0,X,G=1\right)}{p\left(R=1\mid Y^{(1)}=0,X,G=1\right)+p\left(R=0\mid Y^{(1)}=0,X,G=1\right)}\\
&=p\left(R=1\mid Y^{(1)}=0,X,G=1\right)
\end{aligned}$$

Equation (12) is true because

$$\begin{aligned}
&\widetilde{OR}\left(X,Y^{(1)}\right)\\
&=\frac{OR\left(X,Y^{(1)}\right)}{\mathbb{E}\{OR\left(X,Y^{(1)}\right)\mid R=1,X,G=1\}}\\
&=\frac{\frac{p\left(R=0|Y^{(1)},X,G=1\right)p\left(R=1|Y^{(1)}=0,X,G=1\right)}{p\left(R=1|Y^{(1)},X,G=1\right)p\left(R=0|Y^{(1)}=0,X,G=1\right)}}{\frac{p\left(Y^{(1)}=0|R=1,X,G=1\right)}{p\left(Y^{(1)}=0|R=0,X,G=1\right)}}\\
&=\frac{p\left(Y^{(1)}\mid R=0,X,G=1\right)}{p\left(Y^{(1)}\mid R=1,X,G=1\right)}
\end{aligned}$$

So by leverageing Assumption 4, we have

$$\begin{aligned}
&\mathbb{E}\left[\widetilde{OR}\left(X,Y^{(1)}\right)\mid R=1,X,M^{(1)},G=1\right]=\mathbb{E}\left[\frac{p\left(Y^{(1)}\mid R=0,X,G=1\right)}{p\left(Y^{(1)}\mid R=1,X,G=1\right)}\mid R=1,X,M^{(1)},G=1\right]\\
&=\frac{p\left(M^{(1)}\mid X,R=0,G=1\right)}{p\left(M^{(1)}\mid X,R=1,G=1\right)}\underbrace{\mathbb{E}\left[\frac{p\left(Y^{(1)}\mid M^{(1)},R=0,X,G=1\right)}{p\left(Y^{(1)}\mid M^{(1)},R=1,X,G=1\right)}\mid R=1,X,M^{(1)},G=1\right]}_{=1}
\end{aligned}$$

Proposition 2

we first prove the lemma below:

Lemma 1. Define $q(X,Y^{(1)})=1/p(R=1|X,M^{(1)},G=1)$, for any specific functin $g(X,M^{(1)},Y^{(1)})$, under Assumption 3,

$$\mathbb{E}\left[\left(q(X,M^{(1)})R-1\right)g(X,M^{(1)},Y^{(1)})\mid G=1\right]=0 \tag{21}$$

To prove equation (21), we first notice that

$$\begin{aligned}
&\mathbb{E}\left[\left(q(X,M^{(1)})R-1\right)g(X,M^{(1)},Y^{(1)})\mid X,M^{(1)},G=1\right]\\
&=\underbrace{\mathbb{E}\left[\left(q(X,M^{(1)})R-1\right)\mid X,M^{(1)},G=1\right]}_{=0}\cdot\mathbb{E}\left[g(X,M^{(1)},Y^{(1)}\mid X,M^{(1)},G=1\right]\\
&=0
\end{aligned}$$

Using the Law of Iterated Expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(q(X, M^{(1)})R - 1\right) g(X, M^{(1)}, Y^{(1)}) \mid G = 1\right] \\
&\mathbb{E}\left[\mathbb{E}\left[\left(q(X, M^{(1)})R - 1\right) g(X, M^{(1)}, Y^{(1)}) \mid X, M^{(1)}, G = 1\right] \mid G = 1\right] \\
&= 0
\end{aligned}$$

Noticing $\mathbb{E}[Y^{(1)} - \beta \mid G = 1] = 0$ *by definition, let* $g(X, M^{(1)}, Y^{(1)}) = Y^{(1)} - \beta$ *in equation (21),*

$$\begin{aligned}
&\mathbb{E}\left[(q(X, M^{(1)}; \hat{\alpha})R - 1) \cdot (Y^{(1)} - \beta) | G = 1\right] \\
&= \mathbb{E}\left[(q(X, M^{(1)}; \hat{\alpha})R \cdot Y^{(1)} - \beta | G = 1\right] \\
&= 0
\end{aligned}$$

Additionally, let $g(X, M^{(1)}, Y^{(1)}) = h(X, M^{(1)})$*, where* $h(X, M^{(1)})$ *is any specific function of* $\left(X, M^{(1)}\right)$*, under Assumption 1 and 2, we have*

$$\begin{aligned}
&\mathbb{E}\left[\left(w(X, M^{(1)})R - 1\right) h(X, M^{(1)}) | G = 1\right] \\
&= \mathbb{E}\left[\left(w(X, M^{(1)})R \cdot h(X, M^{(1)})\right) | G = 1\right] - \mathbb{E}\left[h(X, M^{(1)}) | G = 1\right] \\
&= \mathbb{E}\left[\left(w(X, M^{(1)})R \cdot h(X, M^{(1)})\right) | G = 1\right] - \sum_{X, M^{(1)}} h(X, M^{(1)}) p(M^{(1)} \mid X, G = 1) p(X \mid G = 1) \\
&= \mathbb{E}\left[\left(w(X, M^{(1)})R \cdot h(X, M^{(1)})\right) | G = 1\right] - \sum_{X, M^{(1)}} h(X, M^{(1)}) p(M^{(1)} \mid X, G = 2) p(X \mid G = 1) \\
&= \mathbb{E}\left[\left(w(X, M^{(1)})R \cdot h(X, M^{(1)})\right) | G = 1\right] - \sum_{X, M^{(1)}} h(X, M^{(1)}) p(M^{(1)} \mid X, R = 1, G = 2) p(X \mid G = 1) \\
&= \mathbb{E}\left[\left(w(X, M^{(1)})R \cdot h(X, M^{(1)})\right) - \mathbb{E}\left[h(X, M) \mid X, R = 1, G = 2\right] \mid G = 1\right] \\
&= 0
\end{aligned}$$

Proof of IPW estimation approach for Model 2 relies on the following lemma, as stated in (Miao et al., 2015):

Proposition 3 *we first prove the lemma below:*
Lemma 2. Define $w(X, Y^{(1)}) = 1/p(R = 1 | X, Y^{(1)}, G = 1)$, for any specific functin $g(X, M^{(1)}, Y^{(1)})$, under Assumption 4, we have

$$\mathbb{E}\left[\left(w(X, Y^{(1)})R - 1\right) g(X, M^{(1)}, Y^{(1)}) | G = 1\right] = 0 \tag{22}$$

To prove equation (22), we first notice

$$\begin{aligned}
&\mathbb{E}\left[\left(w(X, Y^{(1)})R - 1\right) g(X, M^{(1)}, Y^{(1)}) \mid X, Y^{(1)}, G = 1\right] \\
&= \underbrace{\mathbb{E}\left[\left(w(X, Y^{(1)})R - 1\right) \mid X, Y^{(1)}, G = 1\right]}_{=0} \cdot \mathbb{E}\left[g(X, M^{(1)}, Y^{(1)} \mid X, Y^{(1)}, G = 1\right] \\
&= 0
\end{aligned}$$

Using the Law of Iterated Expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(w(X,Y^{(1)})R-1\right)g(X,M^{(1)},Y^{(1)})\mid G=1\right]\\
&\mathbb{E}\left[\mathbb{E}\left[\left(w(X,Y^{(1)})R-1\right)g(X,M^{(1)},Y^{(1)})\mid X,Y^{(1)},G=1\right]\mid G=1\right]\\
&=0
\end{aligned}$$

Noticing $\mathbb{E}[Y^{(1)}-\beta\mid G=1]=0$*, let* $g(X,M^{(1)},Y^{(1)})=Y^{(1)}-\beta$ *in equation (22),*

$$\begin{aligned}
0&=\mathbb{E}\left[(w(X,Y^{(1)};\hat{\alpha},\hat{\gamma})R-1)\cdot(Y^{(1)}-\beta)|G=1\right]\\
&=\mathbb{E}\left[(w(X,Y^{(1)};\hat{\alpha},\hat{\gamma})R\cdot Y^{(1)}-\beta|G=1\right]
\end{aligned}\tag{23}$$

Also, let $g(X,M^{(1)},Y^{(1)})=h(X,M^{(1)})$*, where* $h(X,M^{(1)})$ *is any specific function of* $(X,M^{(1)})$*, under Assumption 1 and 2, we have*

$$\begin{aligned}
0&=\mathbb{E}\left[\left(w(X,Y^{(1)})R-1\right)h(X,M^{(1)})|G=1\right]\\
&=\mathbb{E}\left[\left(w(X,Y^{(1)})R\cdot h(X,M^{(1)})\right)|G=1\right]-\mathbb{E}\left[h(X,M^{(1)})|G=1\right]\\
&=\mathbb{E}\left[\left(w(X,Y^{(1)})R\cdot h(X,M^{(1)})\right)|G=1\right]-\sum_{X,M^{(1)}}h(X,M^{(1)})p(M^{(1)}\mid X,G=1)p(X\mid G=1)\\
&=\mathbb{E}\left[\left(w(X,Y^{(1)})R\cdot h(X,M^{(1)})\right)|G=1\right]-\sum_{X,M^{(1)}}h(X,M^{(1)})p(M^{(1)}\mid X,G=2)p(X\mid G=1)\\
&=\mathbb{E}\left[\left(w(X,Y^{(1)})R\cdot h(X,M^{(1)})\right)|G=1\right]-\sum_{X,M^{(1)}}h(X,M^{(1)})p(M^{(1)}\mid X,R=1,G=2)p(X\mid G=1)\\
&=\mathbb{E}\left[\left(w(X,Y^{(1)})R\cdot h(X,M^{(1)})\right)-\mathbb{E}\left[h(X,M)\mid X,R=1,G=2\right]\mid G=1\right]
\end{aligned}$$

B. DATA GENERATION AND ESTIMATION

B.1. Model 1

We create a binary grouping variable, denoted as G and taking values from the set $\{1,2\}$, for a dataset with a total sample size of n, assuming that each G follows a Bernoulli distribution with parameters 0.5.

For the primary domain, we assume the data generated as follows satisfying Assumption 3:

$$\begin{aligned}
&X\mid G=1\sim\mathcal{N}(1,1)\\
&M^{(1)}\mid X,G=1\sim\mathcal{N}\left(\beta_{m0}+\beta_{m1}X+\beta_{m2}X^2,1\right)\\
&Y^{(1)}\mid X,M^{(1)},G=1\sim\mathcal{N}\left(\beta_{y0}+\beta_{y1}X+\beta_{y2}X^2+\beta_{y3}M^{(1)},1\right)\\
&p(R=1\mid X,M^{(1)},G=1)=logis(a_0+a_1X+a_2M^{(1)}+a_3[M^{(1)}]^2)
\end{aligned}$$

For the correctly specified model (T), we assume $a_3=0$, and for the misspecified setting, we have $a_3=-1$ in particular.

The auxiliary domain is generated so as to satisfy Assumption 1 and 2:

$$\begin{aligned}
&X\mid G=2\sim\mathcal{N}(0,1)\\
&M^{(1)}\mid X,G=2\sim\mathcal{N}\left(\beta_{m0}+\beta_{m1}X+\beta_{m2}X^2,1\right)\\
&p(R=1\mid X,M^{(1)},G=2)=logis(b_0+b_1X)
\end{aligned}$$

The parameters used for presenting results have been discussed in the paper. Details of the simulation setting can be found in the code scripts.

B.2. Model 2

For a dataset with total sample size n, we first generate the binary grouping variable G, $G \in \{1, 2\}$,by assuming each $G \sim Bernouli(0.5)$

In the primary domain ($G = 1$), the data is generated by assuming:

$$
\begin{aligned}
&X \mid G = 1 \sim \mathcal{N}(0, 1) \\
&M^{(1)} \mid R = 1, X, G = 1 \sim \mathcal{N}\left(\beta_{m0} + \beta_{m1}X + \beta_{m2}X^2, 1\right) \\
&Y^{(1)} \mid R = 1, X, M^{(1)}, G = 1 \sim \mathcal{N}\left(\beta_{y0} + \beta_{y1}X + \beta_{y2}X^2 + \beta_{y3}M^{(1)}, 1\right) \\
&\text{logit}\, p(R = 1 \mid X, Y^{(1)} = 0, G = 1) = \alpha_0 + \alpha_1 X + \alpha_2 X^2 \\
&\text{OR}(X, Y^{(1)}) = \exp(-\gamma Y^{(1)})
\end{aligned}
$$

Overall speaking, we generate the dataset using following the variable order: $X \to R \to M \to Y$. To achieve such a data generation order, we leverage the proposed properties in Proposition 1 under Assumption 3.

We first calculated some functions that will be used in the generating process:

For $\mathbb{E}[\text{OR}(X, Y^{(1)}) \mid R = 1, X, M]$, using the above proposed models, let $\mu_Y = \beta_{y0} + \beta_{y1}X + \beta_{y2}X^2 + \beta_{y3}M^{(1)}$, then we have

$$
\begin{aligned}
\mathbb{E}[\text{OR}(X, Y^{(1)}) \mid R = 1, X, M, G = 1] &= \frac{1}{\sqrt{2\pi}} \int \exp\left(-\gamma y - \frac{1}{2}(y - \mu_Y)^2\right) dy \\
&= \frac{1}{\sqrt{2\pi}} \int \exp\left(-\frac{1}{2}(y - \mu_Y + \gamma)^2 - \frac{1}{2}\left(2\mu_Y\gamma - \gamma^2\right)\right) dy \\
&= \exp\left(-\mu_Y\gamma + \frac{1}{2}\gamma^2\right)
\end{aligned}
$$

In a similar way, with regard to the conditional expectation $\mathbb{E}[\text{OR}(X, Y) \mid R = 1, X, G = 1]$, let $\tilde{\mu}Y = \beta_{y0} + \beta_{y1}X + \beta_{y2}X^2$ and $\mu_M = \beta_{z0} + \beta_{z1}X + \beta_{z2}X^2$, we can establish the following functional relationships utilizing the previously mentioned equation:

$$
\begin{aligned}
\mathbb{E}[\text{OR}(X, Y^{(1)}) \mid R = 1, X, G = 1] &= \mathbb{E}[\mathbb{E}[\text{OR}(X, Y^{(1)}) \mid R = 1, X, M, G = 1] \mid R = 1, X, G = 1] \\
&= \frac{1}{\sqrt{2\pi}} \int \exp\left(-\mu_Y\gamma + \frac{1}{2}\gamma^2 - \frac{1}{2}(m - \mu_M)^2\right) dm \\
&= \exp\left(-\tilde{\mu}_Y\gamma + \frac{1}{2}\gamma^2\right) \times \frac{1}{\sqrt{2}} \int \exp\left(-\gamma\beta_{y3}m - \frac{1}{2}(m - \mu_M)^2\right) dm \\
&= \exp\left(-\tilde{\mu}_Y\gamma + \frac{1}{2}\gamma^2 - \mu_M\beta_{y3}\gamma + \frac{1}{2}\beta_{y3}^2\gamma^2\right)
\end{aligned}
$$

Having assembled the essential components as outlined in the equation (11) we can obtain the distribution of $R = 1 \mid X, G = 1$. Let $\mu_R = \alpha_0 + \alpha_1 X + \alpha_2 X^2$, we rewrite equation (11) to be

$$
p(R = 1 \mid X, Y^{(1)} = 0, G = 1) = \frac{\mathbb{E}[\text{OR}(X, Y^{(1)}) \mid R = 1, X, G = 1]}{p(R = 0 \mid X, G = 1)/p(R = 1 \mid X, G = 1) + \mathbb{E}[\text{OR}(X, Y^{(1)}) \mid R = 1, X, G = 1]}
$$

Therefore,

$$\begin{aligned}\frac{p(R=0 \mid X, G=1)}{p(R=1 \mid X, G=1)} &= \mathbb{E}[\mathrm{OR}(X, Y^{(1)}) \mid R=1, X, G=1] \times \frac{1-p(R=1 \mid X, Y^{(1)}=0, G=1)}{p(R=1 \mid X, Y^{(1)}=0, G=1)} \\ &= \exp\left(-\tilde{\mu}_Y\gamma + \frac{1}{2}\gamma^2 - \mu_M\beta_{y3}\gamma + \frac{1}{2}\beta_{y3}^2\gamma^2\right) \times \exp\left(-\alpha_0 - \alpha_1 X - \alpha_2 X^2\right) \\ &= \exp\left(-\tilde{\mu}_Y\gamma + \frac{1}{2}\gamma^2 - \mu_M\beta_{y3}\gamma + \frac{1}{2}\beta_{y3}^2\gamma^2 - \mu_R\right)\end{aligned}$$

so that $\operatorname{logit} p(R=1 \mid X, G=1) = \tilde{\mu}_Y\gamma - \frac{1}{2}\gamma^2 + \mu_M\beta_{y3}\gamma - \frac{1}{2}\beta_{y3}^2\gamma^2 + \mu_R$

Then we calculate for $M^{(1)}|R=0, X, G=1$ using the combination of equations (12) and (13),

$$\begin{aligned}&p(M^{(1)} \mid R=0, X, G=1) \\ &= p(M^{(1)} \mid R=1, X, G=1) \times \frac{\mathbb{E}(\mathrm{OR}(X, Y^{(1)}) \mid R=1, X, M^{(1)}, G=1)}{\mathbb{E}(\mathrm{OR}(X, Y^{(1)}) \mid R=1, X, G=1)} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(M^{(1)} - \mu_M\right)^2\right) \exp\left(-\mu_Y\gamma + \frac{1}{2}\gamma^2 - \left(-\tilde{\mu}_Y\gamma + \frac{1}{2}\gamma^2 - \mu_M\beta_{y3}\gamma + \frac{1}{2}\beta_{y3}^2\gamma^2\right)\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(M^{(1)} - \mu_M\right)^2\right) \times \exp\left(-\beta_{y3}\gamma Z + \mu_M\beta_{y3}\gamma - \frac{1}{2}\beta_{y3}^2\gamma^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(M^{(1)} - \mu_M\right)^2 - \beta_{y3}\gamma\left(M^{(1)} - \mu_M\right) - \frac{1}{2}\beta_{y3}^2\gamma^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(M^{(1)} - \mu_M + \beta_{y3}\gamma\right)^2\right)\end{aligned}$$

So $M^{(1)} \mid R=0, X, G=1$ is a normal distribution with mean $\mu_M - \beta_{y3}\gamma$ and variance 1 .

Upon acquiring the probability distributions for both the random variables, R and $M^{(1)}$, we employ equation (9) to derive the conditional probability distribution of $Y^{(1)}$ under the given conditions: $R=0, X, M^{(1)}$, and $G=1$ as follows:

$$\begin{aligned}p(Y^{(1)} \mid R=0, X, M^{(1)}, G=1) \frac{p\left(Y^{(1)} \mid R=1, X, M^{(1)}, G=1\right) OR\left(X, Y^{(1)}\right)}{\mathbb{E}\left[OR\left(X, Y^{(1)}\right) \mid R=1, X, M^{(1)}, G=1\right]} \\ = \frac{1}{\sqrt{2\pi}} \exp\left(-\gamma y - \frac{1}{2}\left(y - \mu_Y\right)^2 + \mu_Y\gamma - \frac{1}{2}\gamma^2\right) \\ = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\left(y - \mu_Y\right)^2 + 2\gamma(y - \mu_Y) + \gamma^2\right)\right) \\ = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(y - \mu_Y + \gamma\right)^2\right)\end{aligned}$$

As a result, $Y^{(1)} \mid R=0, X, M(1), G=1 \sim \mathcal{N}\left(\mu_Y - \gamma, 1\right)$

In conclusion, we are prepared to articulate the data generation process as follows:

Denoting

$$\begin{aligned}\mu_Y &= \beta_{y0} + \beta_{y1}X + \beta_{y2}X^2 + \beta_{y3}M^{(1)} \\ \tilde{\mu}_Y &= \beta_{y0} + \beta_{y1}X + \beta_{y2}X^2 \\ \mu_M &= \beta_{m0} + \beta_{m1}X + \beta_{m2}X^2 \\ \mu_R &= \alpha_0 + \alpha_1 X + \alpha_2 X^2\end{aligned}$$

The dataset of Domain 1 is generated in a sequential order as follows:

$$
\begin{aligned}
&X \mid G=1 \sim \mathcal{N}(0,1)\\
&\operatorname{logit} p(R=1 \mid X, G=1)=\tilde{\mu}_Y\gamma-\frac{1}{2}\gamma^2+\mu_M\beta_{y3}\gamma-\frac{1}{2}\beta_{y3}^2\gamma^2+\mu_R\\
&M \mid R=1, X, G=1 \sim \mathcal{N}\left(\underbrace{\beta_{m0}+\beta_{m1}X+\beta_{m2}X^2}_{\mu_M}, 1\right)\\
&M \mid R=0, X, G=1 \sim \mathcal{N}\left(\mu_M-\beta_{y3}\gamma, 1\right)\\
&Y \mid R=1, X, M^{(1)}, G=1 \sim \mathcal{N}\left(\underbrace{\beta_{y0}+\beta_{y1}X+\beta_{y2}X^2+\beta_{y3}M}_{\mu_Y}, 1\right)\\
&Y \mid R=0, X, M^{(1)}, G=1 \sim \mathcal{N}\left(\mu_Y-\gamma, 1\right)
\end{aligned}
$$

After the data generation of the primary domain, we generate the data of the auxiliary domain. The key constrain of this auxiliary domain in simulation is Assumption 2. To satisfy that, we first generate a temporary variable R_{tmp} in order to obtain the distribution for $M^{(1)}|X, G=2$ as follows:

$$
\begin{aligned}
&X \mid G=2 \sim \mathcal{N}(1,1)\\
&\operatorname{logit} p(R_{tmp}=1 \mid X, G=2)=\tilde{\mu}_Y\gamma-\frac{1}{2}\gamma^2+\mu_M\beta_{y3}\gamma-\frac{1}{2}\beta_{y3}^2\gamma^2+\mu_R\\
&M \mid R_{tmp}=1, X, G=2 \sim \mathcal{N}\left(\beta_{m0}+\beta_{m1}X+\beta_{m2}X^2, 1\right)\\
&M \mid R_{tmp}=0, X, G=2 \sim \mathcal{N}\left(\mu_M-\beta_{y3}\gamma, 1\right)
\end{aligned}
$$

The missing mechanism only depends on X and we further generate the true R in Domain 2 by assuming

$$p(R=1 \mid X, G=2)=logis(c_0+c_1x)$$

where $logis(x)=(1+\exp(-x))^{-1}$.

For the inverse probability weighting estimation method, we need to estimate $p(R=1|X, Y, G=1; \alpha, \gamma)$, as in equation (10). Recall that $\operatorname{logit} p(R=1 \mid X, Y=0, G=1)=\alpha_0+\alpha_1X+\alpha_2X^2$, $\mathrm{OR}(X, Y^{(1)})=\exp(-\gamma Y^{(1)})$, therefore we have

$$
\begin{aligned}
&\frac{1}{p(R=1|X, Y^{(1)}, G=1)}\\
&=\frac{p(R=1|X, Y^{(1)}=0, G=1; \alpha)+OR(X, Y^{(1)}; \gamma)p(R=0|X, Y^{(1)}=0, G=1; \alpha)}{f(R=1|X, Y^{(1)}=0, G=1; \alpha)}\\
&=1+OR(X, Y^{(1)}; \gamma)\frac{p(R=0|X, Y^{(1)}=0, G=1; \alpha)}{p(R=1|X, Y^{(1)}=0, G=1; \alpha)}\\
&=1+\exp(-\gamma Y^{(1)})\exp\left(-(\alpha_0+\alpha_1X+\alpha_2X^2)\right)\\
&=1+\exp\left(-\gamma Y^{(1)}-(\alpha_0+\alpha_1X+\alpha_2X^2)\right)
\end{aligned}
$$

As a result, $p(R=1|X, Y^{(1)}, G=1)=\frac{1}{1+\exp\left(-\gamma Y^{(1)}-(\alpha_0+\alpha_1X+\alpha_2X^2)\right)}$. The parameters used for presenting results have been discussed in the paper. Details of the simulation setting can be found in the code scripts.

B.3. Additional Simulation Results

In this section, we discussed simulation details on true value, and statistics of results for Model 1 displayed in Fig. 2a and Fig. 2b, and for Model 2 displayed in Fig. 2c and Fig. 2d. Also, we presented the full table of bootstrapping results in Table 5 and Table 6.

The true value of model 1 is calculated theoretically, where $\beta = 1.8$ for both model 1 (T) and model 1 (F). The true value of model 2 is calculated by taking the sample mean of $Y^{(1)}$ in the primary domain in 1000 trials of sample size $n = 20000$. True value of $\beta = -0.659$ for T and $\beta = -0.615$ for F settings.

Table 3. Results for Model 1 displayed in Fig. 2a and Fig. 2b. (True value: $\beta = 1.8$ for both T and F settings)

Setting	**Bias**	**%Bias**	**MSE**	**Var**
IPW (n=500,T)	0.006	0.003	0.062	0.034
IPW (n=1000,T)	0.005	0.003	0.007	0.017
IPW (n=2000,T)	-0.002	-0.001	0.024	0.009
MAR (n=500,T)	0.269	0.149	0.127	0.025
MAR (n=1000,T)	0.270	0.150	0.095	0.011
MAR (n=2000,T)	0.270	0.150	0.010	0.006
MICE (n=500,T)	-0.359	-0.200	0.238	0.046
MICE (n=1000,T)	-0.378	-0.210	0.137	0.023
MICE (n=2000,T)	-0.376	-0.209	0.162	0.012
IPW (n=500,F)	-0.040	-0.022	0.026	0.053
IPW (n=1000,F)	-0.031	-0.017	0.121	0.027
IPW (n=2000,F)	-0.017	-0.010	0.064	0.014
MAR (n=500,F)	-0.544	-0.302	0.244	0.027
MAR (n=1000,F)	-0.541	-0.300	0.486	0.013
MAR (n=2000,F)	-0.544	-0.302	0.403	0.006
MICE (n=500,F)	-0.459	-0.255	0.094	0.040
MICE (n=1000,F)	-0.420	-0.234	0.298	0.022
MICE (n=2000,F)	-0.413	-0.229	0.274	0.011

In Section 5, we presented the bootstrapping result of correctly specified models. Here we have the bootstrapping result of Model 1 and Model 2 (See Table 5 and Table 6, respectively) for settings including the result of both MAR and IPW, and for correctly specified model (T) and misspecified model (F).

B.4. Application to COVID-19 case data

In this study, we conducted meticulous data preprocessing on the COVID-19 Case Surveillance Public Use Data obtained from the Centers for Disease Control and Prevention (CDC), specifically focusing on New York state.

We selected two time periods, March 2020 as the primary MNAR domain, and March 2023 as the auxillary MAR conditions. The original data in March 2020 comprised $129,795$ records, while the March 2023 subset contained $43,210$ records. The same preprocessing steps were applied to both datasets to filter out the low-quality entries: we excluded rows with missing values in critical columns such as the county, age group, sex, case-positive specimen interval, and exposure status. Additionally, cases were filtered to include only those with a "Laboratory-confirmed" status. Categorical variables were transformed into numerical indices, and a county score was recalculated into $[-1, 1]$ to capture the relative positioning of each county. After these steps, the sample size of datasets in March 2020 and March 2023 are $125,737$ and $38,237$ respectively. Since we assumed the missingness mechanism is shared by race and hospitalization in primary domain, we further filtered the dataset in March 2020 accordingly and the final dataset consisted of $78,119$ patients.

Table 4. Results for Model 2 displayed in Fig. 2c and Fig. 2d. (True value: $\beta = -0.659$ (T) , $\beta = -0.615$ (F)

Setting	Bias	%Bias	MSE	Var
IPW (n=500,T)	0.029	-0.044	0.010	0.022
IPW (n=1000,T)	0.018	-0.027	0.023	0.010
IPW (n=2000,T)	0.015	-0.023	0.00007	0.005
MAR (n=500,T)	0.363	-0.551	0.032	0.020
MAR (n=1000,T)	0.363	-0.551	0.241	0.011
MAR (n=2000,T)	0.358	-0.543	0.112	0.005
MICE (n=500,T)	0.113	-0.171	0.000	0.042
MICE (n=1000,T)	0.091	-0.138	0.027	0.021
MICE (n=2000,T)	0.079	-0.120	0.007	0.011
IPW (n=500,F)	0.087	-0.142	0.001	0.018
IPW (n=1000,F)	0.086	-0.139	0.010	0.009
IPW (n=2000,F)	0.084	-0.136	0.028	0.005
MAR (n=500,F)	0.195	-0.317	0.023	0.021
MAR (n=1000,F)	0.193	-0.314	0.044	0.010
MAR (n=2000,F)	0.195	-0.317	0.128	0.005
MICE (n=500,F)	-0.094	0.152	0.384	0.044
MICE (n=1000,F)	-0.106	0.172	0.002	0.025
MICE (n=2000,F)	-0.105	0.171	0.000	0.012

Table 5. Bootstrap confidence intervals for Model 1. (True value of $\beta = 1.8$ for both T and F settings).

Setting	Est.	95% CI	Width	Bias
IPW(n=500,T)	1.941	[1.594, 2.275]	0.681	0.141
IPW(n=1000,T)	1.744	[1.508, 1.970]	0.462	-0.056
IPW(n=2000,T)	1.767	[1.598, 1.930]	0.332	-0.033
MAR(n=500,T)	2.350	[2.052, 2.636]	0.584	0.550
MAR(n=1000,T)	1.957	[1.743, 2.184]	0.441	0.157
MAR(n=2000,T)	2.007	[1.869, 2.139]	0.269	0.207
MICE (n=500,T)	1.592	[1.088 ,2.075]	0.988	-0.208
MICE (n=1000,T)	1.305	[0.994 ,1.624]	0.630	-0.495
MICE (n=2000,T)	1.505	[1.285 ,1.725]	0.440	-0.295
IPW(n=500,F)	1.630	[1.279, 2.012]	0.733	-0.170
IPW(n=1000,F)	1.842	[1.510, 2.136]	0.627	0.042
IPW(n=2000,F)	1.727	[1.442, 2.011]	0.569	-0.073
MAR(n=500,F)	1.337	[1.013, 1.662]	0.649	-0.463
MAR(n=1000,F)	1.331	[1.095, 1.580]	0.484	-0.469
MAR(n=2000,F)	1.213	[1.053, 1.368]	0.315	-0.587
MICE (n=500,F)	1.379	[0.998 ,1.764]	0.766	-0.421
MICE (n=1000,F)	1.575	[1.256 ,1.885]	0.629	-0.225
MICE (n=2000,F)	1.331	[1.114 ,1.531]	0.393	-0.469

Table 6. Bootstrap confidence intervals for Model 2. (True value of $\beta = -0.659$ for T and $\beta = -0.615$ for F settings).

Setting	**Est.**	**95% CI**	**Width**	**Bias**
IPW(n=500,T)	-0.557	[-0.815, -0.303]	0.513	0.102
IPW(n=1000,T)	-0.760	[-0.949, -0.569]	0.380	-0.101
IPW(n=2000,T)	-0.707	[-0.841, -0.561]	0.281	-0.048
MAR(n=500,T)	-0.309	[-0.571, -0.044]	0.527	0.350
MAR(n=1000,T)	-0.380	[-0.585, -0.177]	0.409	0.279
MAR(n=2000,T)	-0.295	[-0.447, -0.150]	0.297	0.364
MICE (n=500,T)	-0.453	[-0.792,-0.085]	0.707	0.206
MICE (n=1000,T)	-0.743	[-1.090,-0.421]	0.669	-0.084
MICE (n=2000,T)	-0.613	[-0.863,-0.355]	0.508	0.046
IPW(n=500,F)	-0.502	[-0.756, -0.275]	0.481	0.113
IPW(n=1000,F)	-0.621	[-0.806, -0.435]	0.371	-0.006
IPW(n=2000,F)	-0.625	[-0.747, -0.505]	0.242	-0.010
MAR(n=500,F)	-0.449	[-0.739, -0.153]	0.585	0.166
MAR(n=1000,F)	-0.543	[-0.738, -0.354]	0.384	0.072
MAR(n=2000,F)	-0.512	[-0.644, -0.368]	0.276	0.103
MICE (n=500,F)	0.672	[-1.078,-0.279]	0.799	-0.057
MICE (n=1000,F)	0.890	[-1.190,-0.559]	0.630	-0.274
MICE (n=2000,F)	-0.944	[-1.166, -0.718]	0.449	-0.329

Table 7. Bootstrap confidence intervals for application to COVID-19 dataset

n	**Est.**	**95% CI**	**Width**
Model 1	0.7396	[0.7190,0.7570]	0.0380
Model 2	0.7836	[0.7558, 0.8071]	0.0513
MAR	0.7337	[0.7277, 0.7395]	0.0117
MCAR	0.7533	[0.7477, 0.7590]	0.0112
MICE	0.6564	[0.6041 ,0.7085]	-0.0969