Causal Inference with Confounders MNAR under Treatment-independent Missingness Assumption

Jian Sun and Bo Fu
School of Data Science, Fudan University, Shanghai 200433, China
November 30, 2022

Abstract

Causal inference in observational studies can be challenging when confounders are subject to missingness. Generally, the identification of causal effects is not guaranteed even under restrictive parametric model assumptions when confounders are missing not at random. To address this, we propose a general framework to establish the identification of causal effects when confounders are subject to treatment-independent missingness, which means that the missing data mechanism is independent of the treatment, given the outcome and possibly missing confounders. We give special consideration to commonly-used models for continuous and binary outcomes and provide counterexamples when identification fails. For estimation, we provide a weighted estimation equation estimating method for model parameters and propose three estimators for the average causal effect based on the estimated models. We evaluate the finite-sample performance of the estimators via simulations. We further illustrate the proposed method with real data sets from the National Health and Nutrition Examination Survey.

Keywords Causal inference; Doubly robust; Identification; Missing not at random; Treatment-independent missingness

1 Introduction

Observational studies play an important role to evaluate the causal effect of a treatment or an exposure on outcomes when a randomized experiment is infeasible. The estimation of causal effects from observational data is challenging as it requires adequate controlling potential confounding of the treatment-outcome relationship. Standard methods such as propensity score weighting (Rosenbaum, 1987), subclassification (Rosenbaum and Rubin, 1984) and matching (Rosenbaum and Rubin, 1983) have been proposed to adjust for confounding from baseline characteristics between treated and untreated individuals when all the confounders are fully observed.
However, confounders are often subject to missingness in practice. When confounders are completely unobserved for all the individuals, different approaches such as instrumental variable (Angrist et al., 1996; Baiocchi et al., 2014), propensity score calibration (Stumer et al., 2005; Hjellvik et al., 2019) and sensitivity analysis (VanderWeele and Arah, 2011) have been proposed to handle unmeasured confounding for causal inference. In other cases, some confounders are partially observed and missing values in confounders may occur under different missing data mechanisms defined by (Little and Rubin, 2014) including missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). When the probability that a confounder is missing does not depend on any observed and unobserved information, the confounder is MCAR and complete case analysis yields an unbiased estimator of the average causal effect in this scenario (Imai and Van Dyk, 2004). When the probability of being missing depends only on observed data values but not on missing information given observed information, the confounder is missing at random. Multiple imputation (Rubin, 1987; Qu and Lipkovich, 2009; Crowe et al., 2010; Mitra and Reiter, 2011; leyrat et al., 2019; Shan et al., 2021), fractional imputation (Corder and Yang, 2019), inverse missing probability weighting (Moodie et al., 2008; Seaman and White, 2014; Leyrat et al., 2021), Bayesian nonparametric generative models (Roy et al., 2017) and doubly robust (Williamson Forbe Wolfe, 2012; Bagmar and Shen, 2022) methods have been proposed to process MAR confounders in causal inference. When confounders are MNAR, the missing-data mechanism depends on unobserved information. Under this scenario, causal effects may also be nonparametrically identified. Mohan and Pearl (2021) described a situation where there are a number of confounders subject to missingness and the probability that a particular confounder is missing only depends on other partially observed confounders. They illustrated that causal effects can be consistently estimated under this constrained situation.

However, causal effects are often non-identifiable when the the probability that a confounder is missing depends on unobserved values of the confounder itself (Frangakis et al., 2007; Egleston et al., 2009). Identifiable means that the causal effect of interest can be uniquely determined by the observed data. To establish identifiability, D’Agostino Jr and Rubin (2000); Blake et al. (2020) described a modified unconfoundness assumption that treatment and counterfactual outcome are conditionally independent given the missing pattern and observed values of confounders. Ding and Geng (2014) studied the identifiability of causal effects under different missing assumptions for discrete confounders and outcomes, and their results were generalized by Yang et al. (2019) to continuous variables under outcome-independent missingness and bounded completeness assumptions. Based on the established identification, EM algorithm (Yang et al., 2014), nonparametric two-stage least squares (Yang et al., 2019), inverse probability weighting (Sun and Liu, 2021) and doubly robust (Mayer et al., 2020) methods have been proposed to estimate causal effects. When the assumptions for identification do not hold, Lu and Ashmead (2018) discussed bounds for the causal effect with sensitivity analysis.

This paper is motivated by a study of the bias caused by missing confounder
values when estimating the impact of marital status on depression (Knol et al., 2010). They simulated missing values in the confounder "income" under MCAR and MAR mechanisms. Multiple imputation, missing indicator method, and complete case analysis are compared in these scenarios. Recommendations for handling missing values have been proposed. However, Davern et al. (2005) suggested that the probability that income is missing depends on the values of income in surveys. We also notice that the current mental condition may be related to the absence of income information, and marital status seems to be independent of missingness given other confounders. To illustrate this problem, we propose a treatment-independent missingness assumption that the missingness of a confounder depends on confounders and outcome but, given these, is independent of treatment.

In this paper, we discuss the identification and inference of causal effects when confounders are missing not at random under the treatment-independent missingness assumption. Note that the treatment-independent missingness assumption is closely related to the shadow variable assumption (Miao and Tchetgen Tchetgen, 2018) where a continuous shadow variable was used to help identification of parametric models when covariates are missing not at random. There are three main differences: First, The treatment which acts as a shadow variable in our work is a binary instead of a continuous variable and contains much less information. So it is more difficult to establish identification. Second, we focus on causal inference and the partially observed confounder is a cause of treatment in our model, while the partially observed covariate is a child node of the shadow variable in the DAG of Miao and Tchetgen Tchetgen (2018), which leads to different modeling on the conditional distribution between the shadow variable and the partially observed variable. Third, based on the estimation of parametric outcome models, we build a doubly robust estimator for average causal effect which can be consistent when one or both of the treatment propensity score model and the outcome model are correctly specified.

The rest of this article is organized as follows. Section 2 describes a general identification strategy and discusses the identification of causal effects in commonly used outcome, treatment propensity score, and missing probability models. In Section 3, we propose a doubly robust estimator of average causal effect which can be consistent when one or both of the treatment propensity score model is correctly specified. Section 4 shows the simulation and real data analysis performance of the proposed method. A discussion of the proposed method is included in Section 5.

2 Identification

2.1 Notation and Assumptions

We follow the counterfactual framework proposed by Rubin (1974). Consider a sample or finite population of n units, let $A_i$ be the binary variable that $A_i = 1$ if individual $i$ received treatment and $A_i = 0$ otherwise. For each unit,
let \( C_i = (C_{i1}, C_{i2}, \ldots, C_{im})^T \) be a vector of the \( m \)-dimensional confounders and \( Y_i \) be an outcome of interest. With causal consistency assumption we have the observed outcome \( Y = AY(1) + (1 - A)Y(0) \), where \( Y(a) \) denotes the potential outcome when \( A = a \). The average treatment effect is \( \tau = E[Y(1) - Y(0)] \). The exchangeability assumption: \( \{Y(1), Y(0)\} \perp A \mid C \) is required to remove the confounding bias. Combining it with the positivity assumption \( 0 < P(A = 1 \mid C) < 1 \), we can use either propensity score matching or weighting methods to estimate causal effects when the confounders are fully observed (Hernán and Robins, 2020).

When one of the confounders has missing values, without loss of generality, we assume the partially observed confounder is \( C_1 \) and let \( R_C \) denote the missing indicator of \( C_1 \). \( R_C = 1 \) if \( C_1 \) is observed and \( R_C = 0 \) if \( C_1 \) is missing. Note that we focus on the situation where the missing values occur in only one confounder. So we observe \((A, Y, R, C_1, \ldots, C_m)^T\) for all samples and \( C_1 \) only for those with \( R_C = 1 \). During this article, we use capital letters to denote random variables and corresponding lowercase letters to denote particular realizations of those random variables.

To establish identification, we introduce the following assumption for missingness:

**Treatment-independent missingness:** Given the outcome \( Y \) and confounder \( C \), the missing indicator \( R_C \) is conditional independent of the treatment; that is, \( R_C \perp A \mid C, Y \).

Combining this assumption with consistency, exchangeability, and positivity assumptions mentioned above and some parametric assumptions, it is easier to identify the causal effects. Figure 1 is a DAG illustrating the causal relationships of these variables.

![Figure 1: A DAG illustrating causal relationships](image)

Note that the treatment \( A \) can act as a shadow variable mentioned by Miao and Tchetgen Tchetgen (2016, 2018) and Miao et al. (2019). But, in this paper, the treatment \( A \) is binary and contains less information than continuous shadow variables. So we need stronger assumptions to guarantee the identification of causal effects.
2.2 A General Identification Strategy

Let \( \Pr(A = a, C = c, Y = y; \theta) \) be the full data distribution model, where \( c = (c_1, c_2, \ldots, c_m) \) is the vector of all confounders and \( \theta = (\theta_1, \theta_2, \ldots, \theta_k) \) is the vector of all parameters. We say that a parameter \( \theta_i \) is identifiable if and only if the mapping from the parameter space \( \Theta_i = \{\theta_i\} \) to the space of observed data distribution \( \{\Pr(a, c, y, R_C = 1; \theta), \Pr(a, c_2, \ldots, c_m, y, R_C = 0; \theta) : \theta \in \Theta\} \) is injective. Note that there is no constraint on other parameters except for \( \theta_i \). It is possible that for two parameters \( \theta_i, \theta_j \in \Theta \), \( \theta_i \) is identifiable while \( \theta_j \) is not.

**Proposition 1.** Consider two parameters \( \theta = (\theta_1, \theta_2, \ldots, \theta_k)^T \), \( \theta' = (\theta_1', \theta_2', \ldots, \theta_k')^T \) with \( \theta_i \neq \theta_i' \) \( (i \in \{1, 2, \ldots, k\}) \). If the parameter \( \theta_i \) is identifiable, then for any \( \Pr(a, y | c; \theta) \neq \Pr(a, y | c; \theta') \).

**Proof:**

If \( \theta_i \) is identifiable, for \( \theta = (\theta_1, \theta_2, \ldots, \theta_k)^T \) and \( \theta' = (\theta_1', \theta_2', \ldots, \theta_k')^T \) with \( \theta_i \neq \theta_i' \), we have:

\[
\Pr(a, c, y, R_C = 1; \theta) \neq \Pr(a, c, y, R_C = 1; \theta')
\]

\[
\Pr(a, y | c; \theta) \neq \Pr(a, y | c; \theta')
\]

Under the treatment-independent missingness assumption \( A \indep R_c | C, Y \), we have:

\[
\frac{\Pr(c; \theta') \cdot \Pr(R_C = 1 | a, c, y; \theta')}{\Pr(c; \theta) \cdot \Pr(R_C = 1 | a, c, y; \theta)} = \frac{\Pr(c; \theta') \cdot \Pr(R_C = 1 | c, y; \theta')}{\Pr(c; \theta) \cdot \Pr(R_C = 1 | c, y; \theta)} = \Pr(h(c, y))
\]

Proposition 1 shows that if the ratio of two different \( \Pr(a, y | c; \theta) \) models indexed by different parameter values \( \theta_i \) and \( \theta_i' \) is not a function of \( a \), then \( \theta_i \) is identifiable. Neither the missing probability model nor the distribution of \( C \) is restricted in this method. The conditional distribution \( \Pr(a, y | c) \) can be determined by treatment propensity score model \( \Pr(a | c) \) and outcome model \( \Pr(y | a, c) \). So when only the treatment propensity score model and the outcome model are specified, we may check the identification of some parameters in these models under the treatment-independent missingness assumption.

**Proposition 2.** Consider models \( \Pr(a | c; \theta) \) and \( \Pr(y | a, c; \theta) \). If \( \theta_i = \theta_i' \) is a necessary condition for

\[
\frac{\Pr(y | A = 1, c; \theta) \Pr(A = 1 | c; \theta)}{\Pr(y | A = 0, c; \theta) \Pr(A = 0 | c; \theta)} = \frac{\Pr(y | A = 1, c; \theta') \Pr(A = 1 | c; \theta')}{\Pr(y | A = 0, c; \theta') \Pr(A = 0 | c; \theta')}, \forall y, c
\]

Then \( \theta_i \) is identifiable.

**Proof:**

According to Proposition 1, If \( \Pr(a, y | c; \theta) / \Pr(a, y | c; \theta') \) varies with \( a \) for any \( \theta_i \neq \theta_i' \), then the parameter \( \theta_i \) is identifiable. Because \( A \) is a binary variable, if

\[
\frac{\Pr(A = 0, y | c; \theta)}{\Pr(A = 0, y | c; \theta')} \neq \frac{\Pr(A = 1, y | c; \theta)}{\Pr(A = 1, y | c; \theta')}
\]

with a positive probability, then the parameter
\( \theta_i \) is identifiable. Also we have \( \text{pr}(a, y | c) = \text{pr}(y | a, c) \cdot \text{pr}(a | c) \), by taking contraposition we can derive the Proposition 2.

The Propositions provide general identification conditions for model parameters, and they can be used to check identification with specified models. In the following literature, we will show that the identification of causal effects also holds under some commonly used model assumptions when confounders are MNAR.

### 2.3 Identification with Parametric Assumptions

**Theorem 1** For the outcome model \( Y \mid a, c \sim N(g(a, c; \beta_c), \phi) \), if the treatment-independent missingness assumption holds, we have:

(a) The conditional average treatment effect \( \mathbb{E}\left[E[Y(1) - Y(0) \mid c]\right] \) is identifiable if

\[
\lim_{y \to \pm \infty} \text{pr}(R_C = 1 | c, y) \xrightarrow{\phi} 1 \quad \text{or} \quad \lim_{y \to \pm \infty} \text{pr}(R_C = 1 | c, y) \xrightarrow{\phi} 1.
\]

(b) If \( \logit \{ \text{pr}(R_C = 1 \mid c, y) \} = G_0(c; \alpha_c) + \alpha_2 y, \alpha_2 \neq 0 \), and the sign of \( \alpha_2 \) is known, the full data distribution \( \text{pr}(a, c, y; \theta) \) is identifiable.

(c) If \( \logit \{ \text{pr}(R_C = 1 \mid c, y) \} = G_0(c; \alpha_c) + \alpha_2 y, \alpha_2 \neq 0 \) and \( Y \mid a, c \sim N(P_0(c; \beta_0) + aG_1(c; \beta_1), \phi) \), the full data distribution is identifiable if \( \exists i \in \{1, 2, \ldots, m\} \), \( c_0 \), such that \( \frac{\partial G_i(c; \alpha_c)}{\partial c_i} \bigg|_{c = c_0} 
eq 0 \) and the sign of \( \frac{\partial G_i(c; \alpha_c)}{\partial c_i} \bigg|_{c = c_0} \) is known,

where \( g(a, c; \beta_c), G_0(c; \alpha_c) \) and \( G_1(c; \beta_1) \) can be any known functions with unknown parameters, \( P_0(c; \beta_0) \) is a polynomial of \( c \) with finite order.

The proof is in the Web Appendix A. Note that the restrictions in Theorem 1(a) are satisfied in Theorem 1(b) and Theorem 1(c). The identification of the conditional average treatment effect is easy to establish if the error term in the outcome model follows i.i.d. normal distribution. While the identification of the average treatment effect needs stronger assumptions on the missing probability model. In real data analysis, it would be helpful to establish identification if we have some domain knowledge on whether the missing probability would become larger or smaller when the value of \( y \) or \( c_i \) changes.

**Remark 1.** The commonly used linear outcome model \( Y \mid a, c \sim N(\beta_0 + \beta_1 a + \beta_2^T c; \beta_c), \phi \) and logistic missing probability model \( \logit \{ \text{pr}(R_C = 1 \mid c, y) \} = \alpha_0 + \alpha_1^T c + \alpha_2 y \) satisfy the restrictions in this Theorem 1. So the average treatment effect is identifiable if these models are correctly specified in empirical studies. Especially, when \( Y \mid a, c \sim N(\beta_0 + \beta_1 a + \beta_2^T c; \beta_c), \phi \) is correctly specified, \( \beta_1 \) equals to the average causal effect.

**Example 1.** Consider the following models: \( C_1 \sim N(\theta_1, 1) \), \( \text{pr}(A = 1 \mid c_1) = \expit(c_1^2 - 1) \), \( Y \mid a, c_1 \sim N(a | c_1), 1 \), and \( \text{pr}(R_C = 1 \mid c_1) = \expit(\theta_2 c_1) \). For two
parameter sets \((\theta_1, \theta_2) = (1, 2)\) and \((\theta'_1, \theta'_2) = (-1, -2)\), we have:

\[
\text{pr}(c_1, R_C = 1; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{(c_1 - 1)^2}{2}\right) \frac{\exp(2c_1)}{1 + \exp(2c_1)}
\]
\[
= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{(c_1 + 1)^2}{2}\right) \frac{\exp(-2c_1)}{1 + \exp(-2c_1)}
\]
\[
= \text{pr}(c_1, R_C = 1; \theta'_1, \theta'_2)
\]

The observed data distributions \(\text{pr}(a, c_1, y, R_C = 1)\) and \(\text{pr}(a, y, R_C = 1)\) are the same under these two parameter sets, but the full data distributions are different. This example shows that even if not all parameters can be identified in the parametric models, the conditional average treatment effect is identifiable, which is consistent with Theorem 1 (a).

Although the normally distributed error term is widely used when the error distribution is assumed to be symmetric, there are other scenarios where the error terms are believed to be skewed-distributed. For example, in the Weibull regression model (Skinner and Humphreys, 1999), the error terms are often assumed to follow the Weibull distribution. We investigate the identification under such scenario in the following theorem:

**Theorem 2** If the treatment-independent missingness assumption holds, and the outcome \(Y = g(a, c; \beta_c) + \sigma W\), where \(W\) follows the standard extreme value distribution, then:

(a) The conditional average causal effect is identifiable.

(b) If logit\(\{\text{pr}(R_C = 1 \mid c, y)\} = G_0(c; \alpha_c) + \alpha_2 y\), the full data distribution is identifiable.

The proof of this theorem is given in Web Appendix B.

Except for continuous outcomes, binary outcomes are also commonly measured in health research. In the following theorem, we state the identification when the logistic outcome model is employed.

**Theorem 3** If the binary outcome follows a Bernoulli distribution and logit\(\{\text{pr}(Y = 1 \mid a, c)\} = g(a, c; \beta_c)\), then:

(a) The causal odds ratio \(\frac{\text{pr}(Y=1\mid a=1, c)}{\text{pr}(Y=0\mid a=0, c)}\) is identifiable.

(b) If logit\(\{\text{pr}(Y = 1 \mid a, c)\} = P_0(c; \beta_0) + aG_1(c; \beta_1)\), logit\(\{\text{pr}(R_C = 1 \mid c, y)\} = \alpha_0 + \alpha_c c + \alpha_2 y\), the full data distribution \(\text{pr}(a, y, c; \theta)\) is identifiable if \(\alpha_2 \neq 0\).

(c) If logit\(\{\text{pr}(Y = 1 \mid a, c)\} = \beta_0 + \beta_1 a + \beta_c c \) (\(\beta_1 \neq 0\)), logit\(\{\text{pr}(A = 1 \mid c)\} = P_1(c; \gamma_c)\), the outcome model and the treatment propensity score model are identifiable,

where \(P_0(c; \beta_0), P_1(c; \gamma_c)\) are polynomials of \(c\) with finite order, \(c\) are continuous variables.

7
The proof of this Theorem is given in Web Appendix C. When confounders are discrete variables, the identification of causal effects has been established nonparametrically in Ding and Geng (2014). Here with continuous confounders and parametric assumptions, we can prove that the full data distribution is identifiable.

Remark 2. In empirical studies, the outcome model logit \( \{ \text{pr}(Y = 1 \mid a, c) \} = \beta_0 + \beta_1 a + \beta_T c \) and the treatment propensity score model logit \( \{ \text{pr}(A = 1 \mid c) \} = \gamma c \) are often employed, where they are assumed to be correctly specified. These two models satisfy the assumptions in Theorem 2(b) and Theorem 2(c), respectively. \( \beta_1 \) is equal to the causal odds ratio in this scenario.

3 Estimation

In this section, we take a two-stage strategy to obtain estimators of causal effects. Firstly we estimate the parameters in missing probability, treatment propensity score, and outcome models. Based on these parametric models, we establish outcome regression (OR), inverse propensity score (IPW), and doubly-robust (DR) estimators for average causal effects.

3.1 Weighted Estimating Equation for Parametric Models

The inverse missing probability weighted estimation equation is an influential approach for missing data analysis when covariates are missing at random. Under MNAR, Miao and Tchetgen Tchetgen (2018) and Sun and Liu (2021) adopted an analogous approach when identification conditions hold. With fully observed data, the standard estimating equation approach employs estimating functions of the form \( \tilde{E} \left[ \left( \frac{r_c}{\text{pr}(RC = 1 \mid c, y; \hat{\alpha})} - 1 \right) G(c, y) \right] = 0 \), where \( G(c, y) \) is a user-specified vector function whose dimension is equal to that of \( \alpha \). It is an unbiased estimating equation when the missing probability model is correctly specified. If \( E \left[ \frac{\partial \left( \frac{r_c}{\text{pr}(RC = 1 \mid c, y; \alpha)} \right)}{\partial \alpha} G(c, y) \right] \) is nonsingular for all \( \alpha \), one can obtain consistent and asymptotically normal estimators. However, when confounders are missing not at random, this method is not feasible because \( C_1 \) is only observed when \( R_C = 1 \). Instead, one can replace \( G(c, y) \) by \( G(c_r, a, y) \):

\[
\tilde{E} \left[ \left( \frac{r_c}{\text{pr}(RC = 1 \mid c, y; \hat{\alpha})} - 1 \right) G(c_r, a, y) \right] = 0, \tag{3.1}
\]

where \( c_r \) denotes the fully-observed confounders which are directly correlated with the partially observed confounder \( C_1 \), \( G(c_r, a, y) \) is a user-specified differentiable vector function of dimension equal to that of \( \alpha \), for example, if \( \text{pr}(R_C = 1 \mid c, y; \hat{\alpha}) = \logit(\alpha_0 + \alpha_c c + \alpha_2 y) \), then we can set \( G(c_r, a, y) = (1, a, y)^T \).

After obtaining \( \text{pr}(R_C = 1 \mid c, y) \), we can combine it with the score functions of the treatment propensity score model and the outcome model to estimate the
parameters in these models:

\[ \mathbb{E} \left[ \frac{r_c}{\text{pr}(R_C = 1 \mid c, y; \hat{\alpha})} \frac{\partial \log(\text{pr}(a \mid c; \gamma))}{\partial \gamma} \right] = 0, \quad (3.2) \]

\[ \mathbb{E} \left[ \frac{r_c}{\text{pr}(R_C = 1 \mid c, y; \hat{\alpha})} \frac{\partial \log(\text{pr}(y \mid a, c; \beta))}{\partial \beta} \right] = 0. \quad (3.3) \]

**Theorem 4** If the missing probability \( \text{pr}(R_C = 1 \mid c, y; \alpha) \) is correctly specified, then (3.1) is an unbiased estimation equation for \( \alpha \). If further \( \text{pr}(a \mid c; \gamma) \) and \( \text{pr}(y \mid a, c; \beta) \) are correctly specified, then (3.2) and (3.3) are unbiased estimation equations for \( \beta \) and \( \gamma \), respectively. The estimated parameters \( \alpha, \beta, \gamma \) are consistent and asymptotically normal under suitable regularity conditions. As \( n \to \infty \),

\[
\sqrt{n}(\hat{\alpha} - \alpha^*) \to N(0, V_{G;\alpha}) \\
\sqrt{n}(\hat{\beta} - \beta^*) \to N(0, V_{G;\beta}) \\
\sqrt{n}(\hat{\gamma} - \gamma^*) \to N(0, V_{G;\gamma})
\]

, where \( (\alpha^*, \beta^*, \gamma^*) \) denote the true values of \( (\alpha, \beta, \gamma) \). \( (V_{G;\alpha}, V_{G;\beta}, V_{G;\gamma}) \) are given in Web Appendix D with proof of this Theorem.

Regularity conditions for asymptotic normality in Theorem 3 can be formulated by applying the general theory of estimating equations such as Theorem 3.4 in Newey and McFadden (1994). Note that the choice of \( G(c_r, a, y) \) can affect the efficiency of estimator. We derive the optimal choice of \( G(c_r, a, y) \) within our class of estimating equations in Web Appendix E.

### 3.2 OR, IPW and DR Estimators

The purposed average treatment effect estimators are based on the model parameter estimators in Section 3.1. Given that we have consistent outcome model and missing probability model estimators, we can get the conditional average treatment effect from the difference between the estimated conditional outcomes in the treated and the placebo groups. Then by averaging the conditional average treatment effect over the weighted empirical distribution of the observed confounders, we can consistently estimate the average treatment effect,

\[ \tau_{OR} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{r_{c_i}}{M(c_i; y_i; \hat{\alpha})} \left[ O_1(c_i; \hat{\beta}) - O_0(c_i; \hat{\beta}) \right] \right\}, \]

where \( O_a(c_i; \hat{\beta}) \) and \( M(c_i, y_i; \hat{\alpha}) \) are the outcome model and the missing probability model estimated by the weighted estimating equations (WEE) approach purposed in Section 3.1, respectively. So we call it the WEE-OR estimator.

**Theorem 5** If the missing probability model and the outcome model are correct, then \( \tau_{OR} \) is a consistent estimator for the average treatment effect under suitable regularity conditions.
The proof of this Theorem is given in Web Appendix F.

When estimating the average causal effect, inverse treatment propensity score weighting is an influential approach. It creates a pseudo population by weighting each sample with the inverse of its condition probability of receiving the treatment given confounders, (i.e., \( \text{pr}(A \mid C) \)). In this population, the treatment \( A \) can be viewed as completely randomized so the difference between the average outcomes in the treatment group and that in the control group is an unbiased estimator of average causal effect. The consistency of the inverse treatment propensity score weighing (IPW) estimator depends on the correct specification of the treatment propensity score model instead of the outcome model. So we can employ a correctly specified propensity score model to obtain an inverse propensity score estimator for the counterfactual outcome and the average treatment effect when the outcome model may be wrongly specified:

\[
\hat{\gamma}^{\text{IPW}}(0) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{r_{c_i}}{M(c_i, y_i; \hat{\alpha})} \frac{(1-a_i)y_i}{1-PS(c_i; \hat{\gamma})} \right\},
\]

\[
\hat{\gamma}^{\text{IPW}}(1) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{r_{c_i}}{M(c_i, y_i; \hat{\alpha})} \frac{a_i y_i}{PS(c_i; \hat{\gamma})} \right\},
\]

\[
\hat{\gamma}_{\text{DR}} = \hat{\gamma}^{\text{IPW}}(1) - \hat{\gamma}^{\text{IPW}}(0),
\]

where \( PS(c; \hat{\gamma}) \) is the treatment propensity score model estimated by the WEE method. So we denote the estimating method as WEE-IPW.

**Theorem 6** If the missing probability model and the treatment propensity score model are correct, then \( \hat{\gamma}^{\text{IPW}}(a) \) is a consistent estimator for \( E[Y(a)] \) under suitable regularity conditions.

The proof of this Theorem is provided in the Web Appendix G. The consistency of \( \hat{\gamma}_{\text{IPW}} \) can be directly derived from the consistency of \( \hat{\gamma}^{\text{IPW}}(a) \), \( a = 0, 1 \).

The doubly robust estimator was proposed as an augmented inverse probability weighting estimator by Robins et al. (1994). In causal inference, this method employs both the treatment propensity score model and the outcome model to derive an estimator that remains consistent if either of the two models is correctly specified. Another advantage of the doubly robust estimator is that it can achieve the semiparametric efficiency bound when both the outcome model and the propensity model are correctly specified. Based on the models estimated with the WEE method, we can build doubly robust (WEE-DR) estimators in the following form:

\[
\hat{\gamma}^{1\text{DR}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{r_{c_i}}{M(c_i, y_i; \hat{\alpha})} \times \left[ \frac{a_i y_i}{PS(c_i; \hat{\gamma})} - \frac{a_i - PS(c_i; \hat{\gamma})}{PS(c_i; \hat{\gamma})} O_i(c_i; \hat{\beta}) \right] \right\},
\]

\[
\hat{\gamma}^{0\text{DR}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{r_{c_i}}{M(c_i, y_i; \hat{\alpha})} \times \left[ \frac{(1-a_i)y_i}{1-PS(c_i; \hat{\gamma})} - \frac{a_i - PS(c_i; \hat{\gamma})}{1-PS(c_i; \hat{\gamma})} O_0(c_i; \hat{\beta}) \right] \right\},
\]

\[
\hat{\gamma}_{DR} = \hat{\gamma}^{1\text{DR}} - \hat{\gamma}^{0\text{DR}},
\]
Theorem 7 When the missing probability model is correct and suitable regularity conditions hold, \( \hat{Y}(a)_{DR} \) is a consistent estimator for \( E[Y(a)] \) if one or both of the treatment propensity score model and the outcome model is correctly specified.

The proof of this Theorem is given in Web Appendix H.

4 Simulation Studies

In this section, We conducted simulation experiments to investigate the finite-sample performance of the proposed estimators in several scenarios. Firstly, we evaluated the weighted estimating equation estimator for model parameters when all models are correctly specified. Then we illustrated the performance of OR, IPW, and DR estimators when models are subject to misspecification. For comparison purposes, we also consider two alternative methods including multiple imputation (Little and Rubin, 2014) and complete-case analysis to address the missingness in the confounder. We generate 1000 independent data sets with sample size \( n = 500 \) and 2000.

4.1 Estimators for Model Parameters

We consider both binary and continuous outcome scenarios when evaluating the weight estimating equation estimators.

In both experiments, we generate \( C_1 \) from \( N(-0.5, 1) \), \( \logit \{ \Pr(A = 1 \mid c_1) \} = \gamma_0 + \gamma_1 c_1 \). In the binary outcome scenario, we generate the outcome from \( \logit \{ \Pr(Y = 1 \mid a, c_1) \} = \beta_0 + \beta_1 a + \beta_2 c_1 \). The missing indicator of \( C_1 \) is generated from \( \text{Ber}(\pi_{\text{bin}}) \), where \( \logit(\pi_{\text{bin}}) = 0.5 - c_1 + 2y \). While in the continuous outcome scenario, \( Y \mid a, c_1 \sim N(\beta_0 + \beta_1 a + \beta_2 c_1, 1) \) and the missing indicator of \( C_1 \) follows \( \text{Ber}(\pi_{\text{con}}) \), where \( \logit(\pi_{\text{con}}) = -1 + c_1 + y \). The parameters in the treatment propensity score model and the outcome model, \( (\gamma_0, \gamma_1, \beta_0, \beta_1, \beta_2) \), are taken to be \( (0.5, 0.5, 0.5, 1.5, -0.5) \) in both binary and continuous outcome experiments. Under such settings, \( A \) and \( Y \) are fully-observed while \( C_1 \) has missing values. The missing data proportion is around 12% in the binary outcome experiment and 48% in the continuous outcome experiment. We apply the proposed weighted estimating equation (WEE), complete-case analysis (CC), and multiple imputation (MI) methods. When conducting multiple imputation, we adopt the predictive mean matching method to impute the missing values (Little and Rubin, 2014). We calculate the bias, the estimated asymptotic standard error (\( \hat{\text{Std}} \)), the sample standard error (\( \text{Std} \)) and the empirical coverage probability of the estimated 95% confidence interval (CI) based on \( \hat{\text{Std}} \). Table 1 summarizes the simulation results.

We can see that the proposed weighted estimating equation estimators have negligible biases, and the 95% coverage probabilities are close to the nominal value. The sample standard errors are close to the estimated asymptotic standard errors, indicating the large-sample estimate of the variance is satisfactory.
Table 1: Simulation results for the estimators of parameters in the treatment propensity score model and the outcome model under different sample sizes

|                  | γ₀  | γ₁  | β₀  | β₁  | β₂  |
|------------------|-----|-----|-----|-----|-----|
|                  | 500 | 2000| 500 | 2000| 500 | 2000|
| **Binary Y**     |     |     |     |     |     |     |
| WEE              | Bias| -0.001| -0.005| 0.021| 0.006| 0.042| 0.015| 0.048| 0.000|
|                  | Std | 0.137| 0.068| 0.141| 0.070| 0.192| 0.096| 0.308| 0.154| 0.239| 0.120|
|                  | Std | 0.136| 0.071| 0.127| 0.073| 0.202| 0.096| 0.325| 0.156| 0.200| 0.112|
|                  | CI  | 0.945| 0.945| 0.973| 0.937| 0.931| 0.950| 0.931| 0.945| 0.975| 0.967|
| MI               | Bias| 0.070| 0.071| 0.050| 0.060| 0.162| 0.149| -0.151| -0.148| 0.285| 0.264|
|                  | Std | 0.117| 0.060| 0.111| 0.057| 0.190| 0.096| 0.257| 0.129| 0.155| 0.082|
|                  | Std | 0.122| 0.060| 0.115| 0.057| 0.188| 0.097| 0.256| 0.129| 0.138| 0.076|
|                  | CI  | 0.905| 0.782| 0.917| 0.818| 0.866| 0.680| 0.909| 0.783| 0.542| 0.093|
| CC               | Bias| 0.117| 0.116| 0.064| 0.067| 0.494| 0.497| 0.039| 0.017| 0.260| 0.258|
|                  | Std | 0.120| 0.060| 0.108| 0.054| 0.205| 0.102| 0.314| 0.155| 0.151| 0.075|
|                  | Std | 0.125| 0.061| 0.113| 0.055| 0.208| 0.102| 0.318| 0.155| 0.143| 0.070|
|                  | CI  | 0.823| 0.514| 0.898| 0.744| 0.346| 0.000| 0.948| 0.954| 0.587| 0.058|
| **Continuous Y** |     |     |     |     |     |     |
| WEE              | Bias| 0.004| 0.001| -0.010| -0.003| 0.017| 0.004| -0.023| -0.008| -0.008| -0.004|
|                  | Std | 0.162| 0.081| 0.176| 0.088| 0.129| 0.064| 0.179| 0.090| 0.127| 0.064|
|                  | Std | 0.162| 0.078| 0.185| 0.088| 0.127| 0.065| 0.175| 0.091| 0.106| 0.058|
|                  | CI  | 0.949| 0.962| 0.935| 0.956| 0.951| 0.949| 0.955| 0.943| 0.979| 0.965|
| MI               | Bias| -0.205| -0.209| -0.232| -0.231| 0.191| 0.189| -0.091| -0.088| -0.093| -0.104|
|                  | Std | 0.100| 0.049| 0.149| 0.073| 0.085| 0.042| 0.105| 0.052| 0.060| 0.029|
|                  | Std | 0.099| 0.049| 0.149| 0.071| 0.085| 0.041| 0.105| 0.052| 0.061| 0.030|
|                  | CI  | 0.452| 0.016| 0.663| 0.113| 0.391| 0.006| 0.861| 0.612| 0.651| 0.060|
| CC               | Bias| 0.570| 0.565| -0.122| -0.122| 0.504| 0.502| -0.247| -0.245| -0.072| -0.076|
|                  | Std | 0.154| 0.077| 0.151| 0.075| 0.113| 0.056| 0.130| 0.065| 0.060| 0.030|
|                  | Std | 0.157| 0.076| 0.153| 0.071| 0.111| 0.057| 0.130| 0.066| 0.060| 0.029|
|                  | CI  | 0.038| 0.000| 0.874| 0.633| 0.007| 0.000| 0.525| 0.038| 0.782| 0.268|

Note: WEE, the purposed weighted estimating equation method; MI, the multiple imputation method using predictive mean matching; CC, the complete-case analysis method; Std, the estimated asymptotic standard error; Std, the sample standard error; CI, the empirical coverage probability of the estimated 95% confidence interval.
As the sample size increases, both bias and variance become smaller. Note that in the binary outcome scenario, the complete-case analysis estimator of \( \beta_1 \) performs well, this can be explained by Bartlett et al. (2015). Except for this, the estimators from complete-case analysis and multiple imputation have relatively large bias, which is not alleviated as the sample size increases.

4.2 Estimators for the Average Treatment Effect

We further compare the estimators for the average treatment effect when there are multiple confounders and model misspecification.

We evaluate the three purposed estimators (WEE-DR, WEE-IPW, WEE-OR) in the following studies. For comparison purposes, we include the following competing methods. (1) CC-OR, MI-OR: The outcome regression method (Bang and Robins, 2005) in which the outcome models are estimated by CC and MI methods mentioned in Section 4.1, respectively. (2) CC-IPW, MI-IPW: The IPW method (Rosenbaum, 1987) in which the treatment propensity score models are estimated by CC and MI methods mentioned in Section 4.1, respectively. (3) CC-AIPW, MI-AIPW: The augment inverse propensity score weighting method (Bang and Robins, 2005) in which the outcome models and the treatment propensity score models are estimated by CC and MI methods mentioned in Section 4.1, respectively.

We consider the condition that confounders are missing not at random. Specifically, we generate the confounders \( C_1 \) from \( N(0, 1) \) and \( C_2 \) from \( \text{Ber}(0.5) \). The missing indicator of \( C_1 \) is \( R_C \), which follows a Bernoulli distribution with

\[
\logit \{ \Pr(R_C = 1 \mid c_1, c_2, y) \} = 1 - 2c_1 + c_2 + 3y.
\]

We treat \( C_1 \) of samples with \( R_C = 0 \) as missing values. Note that \( \Pr(R_C = 1) \) directly depends on the value of \( C_1 \), so the confounders are missing not at random.

We assume the missing probability model is correctly specified and consider the following four scenarios: (a) where both the outcome model and the treatment propensity score model are correctly specified (OCPC); (b) where the outcome model is correctly specified, but the treatment propensity score model is misspecified (OCPM); (c) where the outcome model is misspecified, but the treatment propensity score model is correctly specified (OMPC); (d) where neither the outcome model nor the treatment propensity score model is correctly specified (OMPM). The data-generating mechanisms of the treatment \( A \) and the outcome \( Y \) in the four scenarios are described in Table 2. In all scenarios, we specify the treatment propensity score model as

\[
\logit \{ \Pr(A = 1 \mid c_1, c_2) \} = \gamma_0 + \gamma_1 c_1 + \gamma_2 c_2.
\]

And we specify the outcome model as

\[
Y \sim N(\beta_0 + \beta_1 a + \beta_2 c_1 + \beta_3 c_2, 1).
\]

Figure 2 presents boxplots of WEE-DR, WEE-IPW, WEE-OR, CC-OR, and MI-OR estimators of average causal effects. The results of other estimators are summarized in the Web Appendix I.

The CC-OR and the MI-OR estimators are always biased in all proposed scenarios. The WEE-OR method leads to biased estimators when the outcome model is misspecified while the WEE-IPW method gives biased estimators when the treatment propensity score model is wrong. The WEE-DR estimators perform well when one or both of the outcome and the treatment propensity score
Table 2: A Description of Data-generating Mechanism of $A$ and $Y$

| Scenario | Data-generating Mechanism |
|----------|---------------------------|
| (a). OCPC | $A \sim \text{Ber}(\expit(-0.5 + c_1 + c_2))$  <br> $Y \sim N(1 + 3a + c_1 - c_2, 1)$ |
| (b). OCPM | $A \sim \text{Ber}(\expit(-3 + 3c_1c_2 + 3\exp(c_1c_2)))$  <br> $Y \sim N(1 + 3a + c_1 - c_2, 1)$ |
| (c). OMPC | $A \sim \text{Ber}(\expit(-0.5 + c_1 + c_2))$  <br> $Y \sim N(-1 + 3a + 0.5\exp(c_1 + c_2), 1)$ |
| (d). OMPM | $A \sim \text{Ber}(\expit(-3 + 3c_1c_2 + 3\exp(c_1c_2)))$  <br> $Y \sim N(-1 + 3a + 0.5\exp(c_1 + c_2), 1)$ |

Figure 2: Boxplots of average causal effect estimators. Note: In each boxplot, white boxes are for sample size 500 and gray ones for 2,000. The horizontal line marks the true value of the average causal effect.
models are correctly specified.

5 Real Data Application

We now apply the proposed methods to the study on the average causal effect of marital status on depression in 45+ years old adults (Knol et al., 2010) using a dataset from the 2017-2018 U.S. National Health and Nutrition Examination Survey. There are 2918 subjects in this dataset, among them, 1771 are married or living with a partner, denoted by $A = 1$, and 1147 are single, including divorced and widowed, denoted by $A = 0$. Depression was assessed using the Patient Health Questionnaire-9 (PHQ-9) score (Kroenke et al., 2001). The PHQ-9 is the nine-item depression scale of the patient health questionnaire. Response categories "not at all," "several days," "more than half the days" and "nearly every day" were given a score ranging from 0 to 3. And the total score was calculated ranging from 0 to 27. Following Knol et al. (2010), we include age, sex and monthly income-to-poverty ratio as confounders. We also include education level as a confounder because it may causally influence both the marital status (Cherlin, 2010) and the mental health (Assari, 2020). Among all these confounders, only the monthly income-to-poverty ratio has missing values. The missingness of income to poverty ratio is likely related to the value of income (Davern et al., 2005) and mental status of participants (Torvik et al., 2012). The missing rate is 19.3% for married people and 18.0% for single people. It is plausible that this missingness is independent of marital status after controlling for confounders and outcome. We analyze the dataset with a linear outcome model and logistic missing probability and treatment propensity score models. A standard bootstrap approach has been used for inference. We employ 1000 bootstrap resamples with the replacement of the adult IDs with a resample size of 2918. The bootstrap confidence intervals were calculated using the 2.5th and 97.5th percentiles of the resulting estimates.

Table 3 contains the estimation of the average causal effect in our data analysis. The point estimate of the corresponding parameter of income to poverty ratio in the missing probability model is $-1.211$ with bootstrap standard error 0.179, which is significantly different from 0. This indicates that families with high income tend not to disclose their income information, which is consistent with the MNAR assumption. We can observe differences between the point estimates using the WEE method with those using CC and MI methods, which illustrates the impact of the missing data assumption on causal inference with missing confounders. In the three proposed methods, the WEE-OR estimator of the average causal effect is slightly different from the WEE-DR and the WEE-IPW estimators, which may be caused by the wrongly specified outcome model. According to the WEE-DR estimator, being single can increase the PHQ-9 score by 0.441 on average.
Table 3: Result for the real data analysis

| Method       | $\alpha_{\text{income}}$ | BS | Std | BS 95CI       |
|--------------|--------------------------|----|-----|---------------|
|              | Est                       |    |     |               |
| WEE-DR       | -1.211                    | 0.179 |     | (-1.252 , -1.003) |
| WEE-IPW      | -0.441                    | 0.121 |     | (-0.686 , -0.239) |
| WEE-OR       | -0.432                    | 0.120 |     | (-0.678 , -0.235) |
| CC-AIPW      | -0.885                    | 0.205 |     | (-1.230 , -0.472) |
| CC-IPW       | -0.864                    | 0.203 |     | (-1.252 , -0.504) |
| CC-OR        | -0.930                    | 0.209 |     | (-1.319 , -0.521) |
| MI-AIPW      | -0.912                    | 0.190 |     | (-1.241 , -0.541) |
| MI-IPW       | -0.896                    | 0.187 |     | (-1.252 , -0.555) |
| MI-OR        | -0.966                    | 0.192 |     | (-1.329 , -0.555) |

$\alpha_{\text{income}}$, the parameter for income to poverty ratio in the missing probability model; Est, point estimates; BS, Std, the sample standard error of 1000 points estimates of bootstrap samples; BS 95CI, CI constructed by bootstrap quantiles; WEE-DR, WEE-IPW, WEE-OR, CC-AIPW, CC-IPW, CC-OR, MI-AIPW, MI-IPW, MI-OR, average causal effect estimates of corresponding methods.

6 Discussion

We have discussed identification and inference of causal effects in observational studies when confounders are missing not at random under the treatment-independent missingness assumption. We establish the identification of causal effects in commonly-used outcome, treatment propensity score, and missing probability models and propose a doubly robust estimator of the average causal effect. Simulations show that the doubly robust estimator remains unbiased even when the outcome model is misspecified while the complete case analysis and the multiple imputation method usually lead to biased estimators when confounders are missing not at random.

The treatment-independent missingness assumption is also reasonable in other scenarios. For example, in randomized experiments, if the missing confounders are part of the pre-treatment covariates which are collected before the treatment assignment, their missingness cannot be affected by the treatment (Zhao and Ding, 2021).

There are several possible avenues of future research related to this work. For instance, the proposed method focus on the condition that missingness appears in only one confounder. It would be of interest to extend to more complicated missing pattern scenarios. Also, MNAR data is frequently encountered in longitudinal studies, where extensions of the proposed method can obtain unbiased estimations of causal effects.
References

Angrist, J. D., Imbens, G.W. and Rubin, D. B. (1996) Identification of Causal Effects Using Instrumental Variables (with discussion). *Journal of the American Statistical Association* 91, 444–472.

Assari, S. (2020). Combined effects of ethnicity and education on burden of depressive symptoms over 24 years in middle-aged and older adults in the United States. *Brain Sciences* 10, 4–209.

Baiocchi, M., Cheng, J. and Small, DS. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine* 33, 2297–2340.

Bagmar, M. S. H., and Shen, H. (2022). Causal inference with missingness in confounder. *Journal of Statistical Computation and Simulation* 1-14.

Bang, H. and Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* 61, 962–973.

Bartlett, J. W., Harel, O., and Carpenter, J. R. (2015), Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American Journal of Epidemiology* 182, 730–736.

Blake, H. A., Leyrat, C., Mansfield, K. E., Seaman, S., Tomlinson, L. A., Carpenter, J., and Williamson, E. J. (2020). Propensity scores using missingness pattern information: a practical guide. *Statistics in medicine* 39, 1641-1657.

Cherlin A. (2010) Demographic trends in the United States: A review of research in the 2000s. *Journal of Marriage and the Family* 72, 403-419.

Corder, N. and Yang, S. (2020). Estimating average treatment effects utilizing fractional imputation when confounders are subject to missingness. *Journal of Causal Inference* 8, 249-271.

Crowe, B. J., Lipkovich, I. A. and Wang, O. (2010). Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharmaceutical statistics* 9, 269–79.

D’Agostino Jr, R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* 95, 749–759.

Davern, M., Rodin, H., Beebe, T. J. and Call, K. T. (2005). The effect of income question design in health surveys on family income, poverty and eligibility estimates. *Health Services Research* 40, 1534–1552.

Egleston BL, Scharfstein DO, MacKenzie E. (2009). On Estimation of the Survivor Average Causal Effect in Observational Studies When Important Confounders Are Missing Due to Death. *Biometrics* 65, 497–504.
Ding, P. and Geng, Z. (2014). Identifiability of subgroup causal effects in randomized experiments with nonignorable missing covariates. *Statistics in Medicine* **33**, 1121–1133.

Frangakis, C. E., Rubin, D. B., An, M. W. and MacKenzie, E. (2007). Principal stratification designs to estimate input data missing due to death. *Biometrics* **63**, 641-649.

Hernán MA, Robins J.M. (2020). *Causal Inference: What If*. Boca Raton: Chapman and Hall.

Hjellvik, V., De Bruin, ML., Samuelsen, SO., et al. (2019) Adjusting for unmeasured confounding using validation data: simplified two-stage calibration for survival and dichotomous outcomes. *Statistics in Medicine* **38**, 2719–2734.

Imai, K., Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99**, 854-866.

Knol MJ, Janssen KJ, Donders AR, et al. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of Clinical Epidemiology* **63**, 728–736.

Kroenke, K., Spitzer, R.L., Williams, J.B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine* **16**, 606—613.

Leyrat, C., Seaman, S. R., White, I. R., et al. (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used?. *Statistical methods in medical research* **28**, 3-19.

Leyrat, C., Carpenter, J. R., Bailly, S., and Williamson, E. J. (2021). Common methods for handling missing data in marginal structural models: what works and why. *American Journal of Epidemiology* **190**, 663–672.

Little, R. J., and Rubin, D. (2014) *Statistical Analysis With Missing Data*. New York: Wiley

Lu, B. and Ashmead, R. (2018). Propensity score matching analysis for causal effects with MNAR covariates. *Statistica Sinica* **28**, 2005–2025.

Mayer, I., E. Sverdrup, T. Gauss, J.-D. Moyer, S. Wager, and J. Josse (2020) Doubly robust treatment effect estimation with missing attributes. *The Annals of Applied Statistics* **14**, 1409–1431

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. *Technical report*
Miao, W. and Tchetgen Tchetgen E. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. Biometrika 103, 475–482.

Miao, W. and Tchetgen Tchetgen, E. (2018). Identification and inference with nonignorable missing covariate data. Statistica Sinica 28, 2049–2067.

Mitra, R. and Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. Statistics in Medicine 30, 627–41.

Mohan K, Pearl J. Graphical models for processing missing data.(2021). Journal of the American Statistical Association 116 1023–1037.

Moodie, E. E., Delaney, J. A., Lefebvre, G., and Platt, R. W. (2008) Missing confounding data in marginal structural models: a comparison of inverse probability weighting and multiple imputation. International Journal of Biostatistics 4.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. Handbook of Econometrics 4, 2111–2245.

Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. Statistics in Medicine 28, 1402–1414.

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

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika 74, 13–26.

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effect. Biometrika 70, 41–55.

Rosenbaum, P.R. and Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 387, 516–524.

Roy, J., Lum, KJ., Daniels, MJ. (2017). A Bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. Biostatistics 18, 32-47

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.
Seaman, S. and White, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods* 43, 3499–515.

Shan M., Thomas K.S. and Gutman R. (2021) A multiple imputation procedure for record linkage and causal inference to estimate the effects of home-delivered meals *The Annals of Applied Statistics* 15, 412–436.

Skinner, C. J. and Humphreys, K. (1999) Weibull regression for lifetimes measured with error. *Lifetime Data Analysis* 5, 23–37.

Stumer, T., Schneeweiss, S., Avorn, J. and Glynn, R.J. (2005) Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration *American Journal of Epidemiology* 162, 279–289.

Sun, Z. and Liu, L. (2021) Semiparametric inference of causal effect with non-ignorable missing confounders *Statistica Sinica* 31.

Torvik F.A., Rognmo K., Tambs K. (2012) Alcohol use and mental distress as predictors of non-response in a general population health survey: the HUNT study. *Social Psychiatry and Psychiatric Epidemiology* 47, 805–816.

VanderWeele, T. J., Arah, O.A. (2011) Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* 22, 42–52.

Williamson, E. J., Forbes, A. and Wolfe, R. (2012). Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder. *Statistics in medicine* 31, 4382-4400.

Yang, F., Lorch, S. A., and Small, D. S. (2014). Estimation of causal effects using instrumental variables with nonignorable missing covariates: application to effect of type of delivery NICU on premature infants. *The Annals of Applied Statistics* 48–73.

Yang, S., Wang, L., and Ding, P. (2019). Causal inference with confounders missing not at random. *Biometrika* 106, 875–888.

Zhao, A. and Ding, P. (2022). To adjust or not to adjust? Estimating the average treatment effect in randomized experiments with missing covariates. *Journal of the American Statistical Association*, 1-11.