Double sampling and semiparametric methods for informatively missing data

ALEXANDER W. LEVIS, RAJARSHI MUKHERJEE, RUI WANG, SEBASTIEN HANEUSE
Department of Biostatistics, Harvard T.H. Chan School of Public Health
Boston, Massachusetts 02115, U.S.A.

April 7, 2022

Abstract

Missing data arise almost ubiquitously in applied settings, and can pose a substantial threat to the validity of statistical analyses. In the context of comparative effectiveness research, such as in large observational databases (e.g., those derived from electronic health records), outcomes may be missing not at random with respect to measured covariates. In this setting, we propose a double sampling method, in which outcomes are obtained via intensive follow-up on a subsample of subjects for whom data were initially missing. We describe assumptions under which the joint distribution of confounders, treatment, and outcome is identified under this design, and derive efficient estimators of the average treatment effect under a nonparametric model, as well as a model assuming outcomes were initially missing at random. We compare these in simulations to an approach that adaptively selects an estimator based on evidence of violation of the missing at random assumption. We also show that the proposed double sampling design can be extended to handle arbitrary coarsening mechanisms, and derive consistent, asymptotically normal, and nonparametric efficient estimators of any smooth full data functional of interest, and prove that these estimators often are multiply robust.

Keywords: Missing data; Informative missingness; Semiparametric theory; Causal inference; Study design
1 Introduction

Missing data is a well-studied problem in the statistical literature, with researchers having a vast array of statistical methods at their disposal. Of these methods, the large majority adopt the missing at random (MAR) assumption (Rubin, 1976), in which the probability that data are missing is assumed to depend only on observed data. While methods have been proposed towards estimation of a range of parameters under alternative sets of assumptions (Miao and Tchetgen Tchetgen, 2016; Malinsky et al., 2020), in practice the most commonly used methods assume MAR, and implement inverse-probability weighting (IPW) (Seaman and White, 2013), multiple imputation (Rubin, 2004), or doubly-robust methods (Robins et al., 1994; Tsiatis, 2007). In settings where investigators believe that MAR may not plausibly hold, the usual recommended course of action is to conduct a sensitivity analysis (e.g., see Robins et al. (2000)), or to estimate bounds on the parameters of interest (e.g., see Manski (1990)). We refer to Daniels and Hogan (2008) and Molenberghs et al. (2014) for a review of parametric and semiparametric methods for missing data, including sensitivity analyses.

Common to all of these methods is that they approach the task of dealing with missing data as a post-hoc challenge, with an exclusive focus on analysis methods for the data at-hand. An alternative philosophy is to engage in additional data collection, to be referred to as double-sampling, specifically towards collecting data that can either inform the plausibility of assumptions or be used in an analysis to mitigate bias, or both. Such a philosophy is common when the task is to control confounding bias (e.g. case-control studies (Borgan et al., 2018)) or when concern relates to measurement error and/or miss-classification (Chatterjee and Wacholder, 2002; Carroll et al., 2006; Amorim et al., 2021). In the context of missing data, however, this philosophy seems to have been underexplored.
with exceptions including the use of additional data collection to: address non-response in the survey-sampling literature (Hansen and Hurwitz, 1946; Elliott et al., 2000; Guan et al., 2018; Miao et al., 2021); resolve potentially informative loss-to-follow-up in studies for which the endpoint is a time-to-event outcome (Frangakis and Rubin, 2001; An et al., 2009; Geng et al., 2010, 2012; Qian et al., 2019); and, improve nonparametric bounds on treatment effects when outcomes are missing in randomized trials (Aronow et al., 2015; Coppock et al., 2017). A general treatment of double sampling for missing data, or more generally coarsened data (Heitjan and Rubin, 1991), is lacking, as is the development of methods for settings where the MAR assumption is most tenuous. However, this perspective has not, to the best of our knowledge, been developed as of yet.

One important area of biomedical and public health research where missing data is common pertains to studies making use of electronic health records (EHR). With large sample sizes and rich covariate information over extended periods, EHR data represent a significant and cost-effective opportunity (Haneuse and Shortreed, 2017). Furthermore, these data present a key alternative when randomized clinical trials are not feasible or could not be conducted ethically. EHR systems, however, are typically designed to support clinical and/or billing activities, and not for any particular research agenda. As such, investigators who wish to use EHR data must deal with potential threats to validity (i.e., of a statistical analysis) including, as mentioned, missing data. Moreover, whether a particular data element is observed in an EHR is likely dependent on the complex interplay of numerous factors (Haneuse and Daniels, 2016), casting doubt on the plausibility of the MAR assumption. In such settings, augmentation of the EHR with additional information via double sampling may be especially helpful (Haneuse et al., 2016). Indeed, appealing to this philosophy, Koffman et al. (2021) report on a telephone-based survey used to obtain additional
information for use in an investigation of the association between bariatric surgery and 5-year weight outcomes using data from an EHR. Key to the latter was the fact that many subjects who had undergone bariatric surgery disenrolled from their health plan before their 5-year post-surgery date. Towards understanding the reasons for disenrollment and to evaluate the MAR assumption, the investigators conducted a telephone-based survey to obtain the otherwise missing weight information and other relevant factors. Although the report focuses on disenrollment in relation to missingness, the authors did stress the potential for using the augmented data to correct an otherwise invalid analysis (i.e., of the association between bariatric surgery and weight at 5 years), but identified that additional statistical methods need to be developed.

Responding to these gaps in missing data methods for EHR data, in this paper we explore double sampling as a means to deal with potentially informatively missing data. Specifically, in an observational setting with missing outcomes, we present novel identification results for the causal average treatment effect, and derive efficient and robust estimators. Further, we generalize many of these results to allow for arbitrary coarsening of the desired full data of interest, where the full data are recovered on a subsample via intensive follow-up.

2 A hypothetical EHR-based study

2.1 Data description, notation, and terminology

To help anchor the methods we propose, consider a hypothetical EHR-based study aiming to compare the treatment efficacy of two bariatric surgery procedures on 5-year weight outcomes (e.g., Arterburn et al., 2021). Formally, let $A \in \{0, 1\}$ denote the two levels of
treatment, \( Y \in \mathbb{R} \) the outcome. In our hypothetical study, \( A \) represents the type of surgery (e.g., \( A = 0 \) for Roux-en-Y gastric bypass, \( A = 1 \) for vertical sleeve gastrectomy), and \( Y \) the change in BMI or weight at 5 years post-surgery compared to baseline. Letting \( Y(a) \) denote the potential outcome had we fixed treatment level \( A = a \), for \( a \in \{0, 1\} \), the target parameter of interest is the average treatment effect: \( \mathbb{E}(Y(1)) - \mathbb{E}(Y(0)) \).

To control confounding bias, suppose that we measure a sufficient set of factors, \( L \), such that under the standard assumptions of consistency, positivity, and no unmeasured confounding, the counterfactual means are identified by the \( g \)-formula (Robins, 1986). While we assume \( L \) and \( A \) are measured on all patients, suppose that the outcome is subject to some missingness, for instance because some subjects disenrolled from their health plan prior to their 5-year post-surgery date. Let \( R \in \{0, 1\} \) be the indicator for observing \( Y \), i.e., \( R = 1(Y \text{ is observed}) \). Analogous to Koffman et al. (2021), we suppose follow-up on a subsample of patients for whom \( R = 0 \) was performed, specifically to determine the otherwise missing outcomes; let \( S = 1 \) if a given patient is in the subsample, and \( S = 0 \) otherwise. By design, we have \( S \equiv S(1 - R) \), i.e., \( S \) can only be 1 if \( R = 0 \).

Throughout, we refer to \( O = (L, A, R, S, (R + S)Y) \) as the observed data, meaning the set of information available on all subjects following the double sampling phase. Likewise, the initially observed data consists of \( (L, A, R, RY) \), which excludes the information obtained in the second phase. We refer to \( X = (L, A, Y) \) as the full data, meaning that the desired analysis could be performed had this tuple been completely observed on all subjects. Finally, the complete data refers to \( (L, A, Y, R, S) \), which combines the full and observed data. The complete data distribution will be denoted \( P^* \). We will denote the full data and the the observed data distribution, induced by \( P^* \), as \( P_X^* \) and \( P \) respectively. Despite \( P_X^* \) and \( P \) being determined by \( P^* \), we use different notation to help clarify iden-
tifying conditions. Throughout, except when considering only the initially observed data, we assume that we observe a random sample $O_1, \ldots, O_n \sim P$.

### 2.2 Assumptions and nonparametric identification

Under the standard causal assumptions alluded to above, the counterfactual mean would be identified from the full data distribution via the $g$-formula (Robins, 1986), $\chi_a(P_X^*) = \mathbb{E}_{P^*}(\mu^*_a(L))$, where $\mu^*_a(L) = \mathbb{E}_{P^*}(Y \mid L, A = a)$. Moreover, the typical assumption that the outcomes are MAR allows $\chi_a(P_X^*)$ to be identified from the initially observed data $(L, A, R, RY)$; this assumption in our example is as follows:

**Assumption 1** (Missing at random outcomes). $R \perp Y \mid L, A$.

In other words, Assumption 1 asserts that the probability that $Y$ is initially observed depends only on the measured variables ($L, A$). That said, one can conceptualize a number of missing not at random (MNAR) scenarios in which Assumption 1 may be violated. For example, it may be that patients with worse outcomes (in a manner beyond what can be predicted with $L$ and $A$) interact more often with the health care system, and thus have less missing data and/or are less likely to disenroll from their health plan. It is also possible that subjects with worse outcomes are more likely to drop out, perhaps to receive care outside of their original health plan.

A notable issue is that if Assumption 1 fails to hold, then $\chi_a(P_X^*)$ is not identified by the initially observed data distribution, since there is no way to capture the distribution of $Y$ given $L, A, R = 0$. This observation provides the key motivation for double sampling, specifically to collect data that will enable resolving this problem. With that in mind, suppose that the following assumptions hold:

**Assumption 2** (No informative second-stage selection). $S \perp Y \mid L, A, R = 0$. 

6
Assumption 3 (Positivity of second-stage selection probabilities). For some \( \epsilon > 0 \), it holds that \( P[S = 1 \mid L, A, R = 0] \geq \epsilon, P \)-almost surely.

These two conditions are sufficient to nonparametrically identify the complete data distribution \( P^* \), as we now establish. Below, let \( p = \frac{dP}{d\mu} \) and \( p^* = \frac{dP^*}{d\mu} \) denote the densities for \( P \) and \( P^* \), respectively, both with respect to some dominating measure \( \mu \).

Proposition 1. Under Assumptions 2 and 3, the complete data distribution \( P^* \) is identified from the observed data distribution \( P \).

Proof. By writing the the complete data density as a marginal times a conditional density,

\[
p^*(L, A, R, S, RY, (1 - R)Y) = p(L, A, R, RY, S)p^*(Y \mid L, A, R = 0, S)^{1 - R} \\
= p(L, A, R, RY, S)p(Y \mid L, A, R = 0, S = 1)^{1 - R},
\]

by Assumption 2. This expression only depends on \( p \), meaning that \( p \) identifies \( p^* \). Note, we may safely introduce \( S = 1 \) in the conditioning event by positivity (Assumption 3).

A few remarks are in order regarding the assumptions enabling identification in Proposition 1. Assumption 2 intuitively requires that membership in the second-stage sample be as good as randomized within levels of measured confounders \( L \) and treatment \( A \), among those with \( R = 0 \). For instance, this could be plausible if follow-up in the second stage was indeed (conditionally) randomized, and there was no failure to ascertain the desired outcomes. Alternatively, this assumption would hold if there were some missing data at the second stage, so long as the probability of missingness depended solely on \( L \) and \( A \). It is important to distinguish the reasons for incomplete data in the initial sample and potential unsuccessful follow-up at the second stage. In the motivating EHR example, the reasons that some outcomes are observed or not in the electronic record may be quite complex, and
are typically out of the hands of investigators using these data retrospectively. By con-
trast, the double sampling regime offers the investigators some control over the method of
subsampling during follow-up. That is, one may design the second stage of data collection
and allocate sufficient resources to maximize the potential for Assumption 2 to hold.

Meanwhile, Assumption 3 requires that within all levels of $L$ and $A$, the probability
of being sampled in the second stage is bounded away from zero. In particular, among
those with $R = 0$, no strata defined in terms of the observed data may be systematically
excluded from follow-up at the second stage.

By Proposition 1, under the two given assumptions, any functional depending on $P^*$ can
be targeted using the observed data. For instance, the mean counterfactual $\chi_a(P^*_X)$ is now
identified: the conditional mean of the outcome $\mu^*_a(L) = \mathbb{E}_{P^*}(Y \mid L, A = a)$ is represented
equally by $\mu_a(L) = \mu_{a,R}(L)\gamma_a(L) + \mu_{a,S}(L)(1 - \gamma_a(L))$, where $\gamma_a(L) = P[R = 1 \mid L, A = a]$,
$\mu_{a,R}(L) = \mathbb{E}_P(Y \mid L, A = a, R = 1)$, and $\mu_{a,S}(L) = \mathbb{E}_P(Y \mid L, A = a, S = 1)$, so that
$\chi_a(P^*_X)$ is given by $\tau_a(P) := \mathbb{E}_P(\mu_a(L))$. At this point, we can envision several approaches
to infer $\chi_a(P^*_X)$. We introduce these in the next section.

### 2.3 Overview of possible analytic approaches

One might appeal to several statistical methods for estimating the average treatment effect
$\chi_1(P^*_X) - \chi_0(P^*_X)$ in the hypothetical example presented above. In this subsection, we
describe five such families of analytic approaches, grouped together based on the data that
one has collected, and the assumptions that are deemed to be reasonable in a given study.
When possible, we highlight approaches that already exist within each family, and note
novel contributions of the present paper. For a summary of the estimands and relevant
nuisance functions, refer to Table 1.
Table 1: Summary of nuisance functions and estimands.

| Nuisance function | Definition                                                                 | Description                                                                 | Related estimand |
|-------------------|---------------------------------------------------------------------------|-----------------------------------------------------------------------------|------------------|
| \(\pi_a(L)\)     | \(P[A = a \mid L]\)                                                      | Treatment mechanism                                                        | \(\xi_a(P) = \mathbb{E}_P(\mu_{a,R}(L))\) |
| \(\gamma_a(L)\)  | \(P[R = 1 \mid L, A = a]\)                                              | Missingness mechanism                                                       |                 |
| \(\eta_{a,0}(L)\)| \(P[S = 1 \mid L, A = a, R = 0]\)                                       | Double sampling / second-stage selection probabilities                      |                 |
| \(\mu_{a,R}(L)\) | \(\mathbb{E}_P(Y \mid L, A = a, R = 1)\)                               | Outcome regression among initially observed                                 | \(\tau_a(P) = \mathbb{E}_P(\mu_a(L))\) |
| \(\mu_{a,S}(L)\) | \(\mathbb{E}_P(Y \mid L, A = a, S = 1)\)                               | Outcome regression among follow-up subsample                                | \(\tau^*_a(P) = \mathbb{E}_P(\mu_{a,MAR}(L))\) |
| \(\mu_a(L)\)     | \(\mu_{a,R}(L)\gamma_a(L) + \mu_{a,S}(L)(1 - \gamma_a(L))\)           | Outcome regression                                                         |                 |
| \(\mu_{a,MAR}(L)\)| \(\mathbb{E}_P(Y \mid L, A, R + S = 1)\)                               | Outcome regression among all observed                                       |                 |

First, suppose that one only has collected the initially observed data \((L, A, R, RY)\). In the absence of special circumstances, for example the presence of a fully observed “shadow variable” (Miao and Tchetgen Tchetgen, 2016), identification of the treatment effect of interest would typically rely on the MAR assumption (i.e., Assumption 1). In particular, MAR allows \(\mu^*_a(L)\) to be identified via \(\mu_{a,R}(L) = \mathbb{E}_P(Y \mid L, A = a, R = 1)\). On the basis of this fact, Ross et al. (2021) propose an inverse-probability weighted estimator in the presence of missing outcomes that, under Assumption 1, targets the average treatment effect \(\xi_1(P) - \xi_0(P)\), where \(\xi_a(P) := \mathbb{E}_P(\mu_{a,R}(L))\). This approach involves a model for the missingness mechanism \(\gamma_a(L)\), as well as for the treatment assignment probabilities, \(\pi_a(L) = P[A = a \mid L]\), which can be estimated as usual since \((L, A)\) is fully observed. Similarly, Davidian et al. (2005) and Williamson et al. (2012) developed doubly robust estimators of \(\xi_1(P) - \xi_0(P)\) that simultaneously estimate the outcome regression function \(\mu_{a,R}(L)\), and are consistent if models for either \(\mu_{a,R}\) or \((\gamma_a, \pi_a)\) are correctly specified.

Second, suppose that due to a belief that Assumption 1 failed to hold, one decided to collect outcomes via intensive follow-up, so that the observed data are \(O = (L, A, R, S, (R+S)Y)\). Under Assumptions 2 and 3, the counterfactual means are nonparametrically identi-
tified by $\tau_a(P)$, as explained in section 2.2. Thus, one can construct an estimator of the average treatment effect by targeting $\tau_a(P)$, for instance by modeling $\mu_a(L)$ through parametric models for $\mu_{a,R}(L)$, $\mu_{a,S}(L)$, and $\gamma_a(L)$ defined above. In section 3.1, we will derive a nonparametric efficient and multiply robust estimator of the average treatment effect that requires a model for the treatment probability $\pi_a(L)$, as well as for the double sampling probabilities $\eta_{a,0}(L) = P[S = 1 \mid L, A = a, R = 0]$. As we will prove, this approach has the advantage that under relatively mild conditions, $\sqrt{n}$-rate convergence can still be attained while using flexible machine learning-based models for each nuisance function.

Third, suppose that one has collected follow-up outcome data as above to obtain $O = (L, A, R, S, (R + S)Y)$, but is confident that the initial MAR assumption actually held. For example, investigators may have pursued such additional data collection in the hopes of gaining efficiency in estimating the parameter of interest. Combining Assumption 1 with $S \equiv S(1 - R)$ and Assumption 2, we would have $(R, S) \perp \perp Y \mid L, A$, so that $\mu_{a*}(L)$ can equally be written $\mu_{a,\text{MAR}}(L) = \mathbb{E}_P(Y \mid L, A, R + S = 1)$. Again, while we could construct estimating equations to estimate the mean counterfactuals $\tau_{a*}(P) := \mathbb{E}_P(\mu_{a,\text{MAR}})$ using a parametric model for $\mu_{a,\text{MAR}}(L)$, in section 3.1 we derive a semiparametric efficient estimator of $\tau_{a*}(P)$, which is of a doubly robust augmented IPW form.

Fourth, suppose one has access to $O = (L, A, R, S, (R + S)Y)$ as above, but is unwilling to make a judgment about whether Assumption 1 actually held. For example, one clinician collaborator believes firmly the outcomes were initially MAR given $(L, A)$, but another collaborator believes there were important missing confounders. In this case, it may be desirable to conduct a hypothesis test using the observed data to assess the plausibility of Assumption 1 — while this assumption is untestable from only the initially observed data, under Assumptions 2 and 3 the complete data distribution is identified and in principle
MAR can be tested. An intuitively appealing estimator would select, based on the results of a hypothesis test, between the nonparametric efficient estimator which does not rely on Assumption 1 and the semiparametric efficient estimator which assumes the latter. Note, however, that naively using the standard error for the chosen estimator would not account for the uncertainty in the estimator selection, and in general would not be valid. As a consequence, we do not derive any theory for this approach, but do consider it in our simulation study in section 4.1, as it may serve as a useful comparator.

Lastly, in the same setting as the previous paragraph, one may prefer a more explicit data-adaptive approach, one which selects between the two candidate estimators and provides valid confidence intervals that account for the selection process. The recently proposed method of Rothenhäusler (2020) is well-suited for this task. As an overview of their method in the one-parameter case, they consider estimation of a generic parameter \( \theta_0(P) \) when there are \( k + 1 \) asymptotically linear estimators \( \hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_k \), such that the base estimator \( \hat{\theta}_0 \) is \( \sqrt{n} \)-consistent for \( \theta_0(P) \) (i.e., \( \sqrt{n}(\hat{\theta}_0 - \theta_0(P)) = O_P(1) \)), but each other estimator is \( \sqrt{n} \)-consistent for \( \theta_j(P) \), for \( j = 1, \ldots, k \), where \( \theta_j(P) \) may or may not equal the parameter of interest \( \theta_0(P) \). Their proposed estimator selects among the \( k + 1 \) candidate estimators by minimizing an estimate of the mean squared error: \( \hat{\theta} = \hat{\theta}_j \), where \( j = \arg \min \overline{R}(j) \), and \( \overline{R}(j) = \max \{ 0, (\hat{\theta}_j - \hat{\theta}_0)^2 - \hat{\rho}_j^2 / n \} + \hat{\sigma}_j^2 / n \). Here, \( \hat{\sigma}_j^2 \) is an estimator of the asymptotic variance of \( \sqrt{n}(\hat{\theta}_j - \theta_j(P)) \), and \( \hat{\rho}_j^2 \) is an estimator of the asymptotic variance of \( \sqrt{n}(\hat{\theta}_j - \theta_j(P) - \hat{\theta}_0 + \theta_0(P)) \). Importantly, in their Theorem 4, Rothenhäusler (2020) derive asymptotically valid confidence intervals for \( \theta_0(P) \) based on \( \hat{\theta} \), by taking into account the uncertainty due to the selection procedure. In section 3.2, we apply this general methodology to construct an adaptive estimator \( \hat{\tau}_a^\dagger \) of \( \tau_a(P) \) (our \( \theta_0(P) \)), under Assumptions 2 and 3. The base estimator \( \hat{\theta}_0 \) will be the nonparametric efficient estimator \( \hat{\tau}_a \) (the second
approach described above) and $\hat{\theta}_1$ will be the semiparametric estimator $\hat{\tau}^{*_a}$ that additionally relies on Assumption 1 (the third approach described above).

3 Efficient treatment effect estimation

3.1 Estimators based on efficient influence functions

As outlined in the previous section, we aim to develop an efficient estimator $\hat{\tau}_a$ of $\tau_a(P)$, which equals $E(Y(a))$ under standard causal assumptions and under Assumptions 2 and 3, as well as an efficient estimator $\hat{\tau}^{*_a}$ of $\tau^{*_a}(P) = E_P(\mu_{a,\text{MAR}}(L))$, which corresponds to $E(Y(a))$ if Assumption 1 also holds. To proceed, we appeal to semiparametric theory, and derive the nonparametric influence function for $\tau_a(P)$, and the semiparametric efficient influence function for $\tau^{*_a}(P)$ (Bickel et al., 1993; Tsiatis, 2007). By modeling the necessary components to estimate these theoretical influence functions, and performing sample splitting, we present estimators that efficiently and robustly target $\tau_a(P)$ and $\tau^{*_a}(P)$, akin to many recently developed methods (Chernozhukov et al., 2018; Kennedy, 2019, 2020).

Theorem 1. The nonparametric influence function (with respect to the maximal tangent space) of $\tau_a(P)$ is

$$
\hat{\tau}_a(O; P) = \mu_a(L) - \tau_a(P) + \frac{1}{\pi_a(L)} \left\{ \left( R + \frac{S}{\eta_a,0(L)} \right) Y - \mu_a(L) \right. \\
\left. + (1 - R) \left( 1 - \frac{S}{\eta_a,0(L)} \right) \mu_{a,S}(L) \right\},
$$

and under Assumption 1, the semiparametric efficient influence function of $\tau^{*_a}(P) = \tau_a(P)$ is

$$
\hat{\tau}^{*_a}(O; P) = \mu_{a,\text{MAR}}(L) - \tau^{*_a}(P) + \frac{1}{\pi_a(L)[\gamma_a(L)+(1-\gamma_a(L))\eta_a,0(L)]} (Y - \mu_{a,\text{MAR}}(L)).
$$

Proof. The influence function $\hat{\tau}_a(O; P)$ is obtained as a special case of Proposition 3, noting
from Hahn (1998) that the full data influence function for the mean counterfactual $g$-formula functional $\chi_a(P_{X^*})$ is $\dot{\chi}_a(P_{X^*}) = \mu_a^*(L) - \chi_a(P_{X^*}) + \frac{1}{\pi_a(L)}(Y - \mu_a^*(L))$. The derivation of the tangent space of the model that asserts Assumption 1, and of the resulting efficient influence function $\dot{\tau}_a^*(O; P)$, is provided in Appendix B. \hfill $\Box$

It is interesting to note that $\tau_a^*(P)$ also equals $\mathbb{E}(Y(a))$ under the weaker condition that $R + S \perp \perp Y \mid L, A$, i.e., MAR holds for the combined sample. Moreover, the influence function $\dot{\tau}_a^*(O; P)$ remains valid and efficient under this modified assumption.

With these influence functions in hand, it is straightforward to construct estimators of $\tau_a(P)$ and $\tau_a^*(P)$: standard one-step estimators are given by adding the sample mean of the estimated influence functions to the plugin estimators $\tau_a(\hat{P})$ and $\tau_a^*(\hat{P})$. In order to ensure that the proposed estimators are consistent and asymptotically normal under relatively weak conditions (e.g., avoiding empirical process conditions that restrict the complexity of the classes of nuisance functions), we propose to use sample splitting and “cross-fitting” (Zheng and van der Laan, 2011; Chernozhukov et al., 2018). Concretely, suppose we randomly partition the sample of size $n$ into $K \geq 2$ equal subsets of size $n/K$, and let $I_k \subset \{1, \ldots, n\}$ denote the $k$-th subset, for $k = 1, \ldots, K$. Let $I_k^c = \{1, \ldots, n\} \setminus I_k$ be the set of sample points not belonging to $I_k$, suppose $\hat{P}_k$ is fit using only $I_k^c$, and define

$$\hat{\tau}_{a,k} = \tau_a(\hat{P}_k) + \frac{K}{n} \sum_{i \in I_k} \hat{\tau}_a(O_i; \hat{P}_k), \text{ for } k = 1, \ldots, K,$$

and define $\hat{\tau}_{a,k}^*$ analogously. Then we define

$$\hat{\tau}_a = \frac{1}{K} \sum_{k=1}^K \hat{\tau}_{a,k} \text{ and } \hat{\tau}_a^* = \frac{1}{K} \sum_{k=1}^K \hat{\tau}_{a,k}^*,$$

to be the final estimators. Note that we only need to fit $\hat{P}_k$ through the nuisance functions
appearing in a given influence function. That is, we need models for \((\mu_{a,R}, \mu_{a,S}, \gamma_a, \pi_a, \eta_{a,0})\) to construct \(\hat{\tau}_a\), and models for \((\mu_{a,\text{MAR}}, \gamma_a, \pi_a, \eta_{a,0})\) to construct \(\hat{\tau}_a^*\).

In the following result, we characterize the asymptotic consistency, normality, and efficiency of \(\hat{\tau}_a\) and \(\hat{\tau}_a^*\). Below, for any function \(f\) of observed data \(O\), we write \(\|f\| := \sqrt{\mathbb{E}_P(f(O)^2)}\) for its \(L_2(P)\)-norm.

**Theorem 2.** Suppose \(\|\hat{\tau}_a(\cdot; \hat{P}_k) - \tau_a(\cdot; P)\| = o_P(1)\) for \(k = 1, \ldots, K\). Then

\[
\hat{\tau}_a - \tau_a(P) = O_P \left( \frac{1}{\sqrt{n}} + \frac{1}{K} \sum_{k=1}^{K} \text{Bias}_{\tau_a}(\hat{P}_k; P) \right),
\]

where

\[
\text{Bias}_{\tau_a}(\hat{P}; P) = \mathbb{E}_P \left( \left( 1 - \frac{\pi_a(L)}{\bar{\pi}_a(L)} \right) (\bar{\mu}_a(L) - \mu_a(L)) \right) + \mathbb{E}_P \left( (1 - \gamma_a(L)) \frac{\pi_a(L)}{\bar{\pi}_a(L)} (1 - \frac{\eta_{a,0}(L)}{\bar{\eta}_{a,0}(L)}) (\bar{\mu}_{a,S}(L) - \mu_{a,S}(L)) \right) + \mathbb{E}_P \left( (1 - \gamma_a(L)) \frac{\pi_a(L)}{\bar{\pi}_a(L)} (1 - \frac{\eta_{a,0}(L)}{\bar{\eta}_{a,0}(L)}) (\bar{\mu}_{a,S}(L) - \mu_{a,S}(L)) \right).
\]

for any (fixed) \(\hat{P}\). Moreover, \(\text{Bias}_{\tau_a}(\hat{P}_k; P) = o_P(n^{-1/2})\), for \(k = 1, \ldots, K\), then \(\sqrt{n}(\hat{\tau}_a - \tau_a(P)) \xrightarrow{d} N(0, V)\), where \(V = \text{Var}_P(\hat{\tau}_a(O; P))\) is the nonparametric efficiency bound.

Similarly, assuming \(\|\hat{\tau}_a^*(\cdot; \hat{P}_k) - \tau_a^*(\cdot; P)\| = o_P(1)\) for \(k = 1, \ldots, K\),

\[
\hat{\tau}_a^* - \tau_a^*(P) = O_P \left( \frac{1}{\sqrt{n}} + \frac{1}{K} \sum_{k=1}^{K} \text{Bias}_{\tau_a^*}(\hat{P}_k; P) \right),
\]

where

\[
\text{Bias}_{\tau_a^*}(\hat{P}; P) = \mathbb{E}_P \left( \left( 1 - \frac{\pi_a(L)}{\bar{\pi}_a(L)} \{\gamma_a(L) + (1 - \gamma_a(L))\eta_{a,0}(L)\} \right) \left( \bar{\mu}_{a,\text{MAR}}(L) - \mu_{a,\text{MAR}}(L) \right) \right)
\]

for any \(\hat{P}\). If \(\text{Bias}_{\tau_a^*}(\hat{P}_k; P) = o_P(n^{-1/2})\), for \(k = 1, \ldots, K\), then \(\sqrt{n}(\hat{\tau}_a^* - \tau_a^*(P)) \xrightarrow{d} \).
$\mathcal{N}(0, V^*)$, where $V^* = \text{Var}_P(\hat{\tau}_a^*(O; P))$ is the semiparametric efficiency bound under Assumption 1.

**Proof.** The results for $\hat{\tau}_a$ are direct applications of Theorem 3 and Proposition A1, invoking the product-error form (e.g., see Chernozhukov et al. (2018)) of the asymptotic bias term of the full-data influence function:

$$
\text{Bias}_{\chi_a}(\tilde{P}_X^*; P_X^*) := \mathbb{E}_{P^*}(\chi_a(X; \tilde{P}_X^*)) + \chi_a(\tilde{P}_X^*) - \chi_a(P_X^*)
= \mathbb{E}_{P^*}\left(\left(1 - \frac{\pi_a(L)}{\tilde{\pi}_a(L)}\right)(\tilde{\mu}_a^*(L) - \mu_a^*(L))\right).
$$

The results for $\hat{\tau}_a^*$ follow from the exact same logic used to prove Theorem 3. The form of the asymptotic bias term can be seen by noting that

$$
\hat{\tau}_a^*(O; P) = \mu_{a, \text{MAR}}(L) - \tau_a^*(P) + \frac{T(Y - \mu_{a, \text{MAR}}(L))}{P[T = 1 | L]}
$$

where $T = (R + S)1(A = a)$ is a sort of modified treatment indicator. As $\mu_{a, \text{MAR}}(L) = \mathbb{E}_P(Y | L, T = 1)$, the form of $\tau_a^*(P)$ is the same as that for the standard $g$-formula, with $T$ in place of $1(A = a)$. Thus, the asymptotic bias term results from the same replacement in the full-data asymptotic bias term written above. □

In addition to consistency, asymptotic normality, and efficiency, Theorem 2 reveals robustness properties of both $\hat{\tau}_a$ and $\hat{\tau}_a^*$. Observe that $\text{Bias}_{\tau_a}(\tilde{P}; P) = 0$ if (i) $(\tilde{\mu}_{a,R}, \tilde{\mu}_{a,S}, \tilde{\gamma}_a) = (\mu_{a,R}, \mu_{a,S}, \gamma_a)$; (ii) $(\tilde{\mu}_{a,S}, \tilde{\pi}_a) = (\mu_{a,S}, \pi_a)$; or (iii) $(\tilde{\pi}_a, \tilde{\eta}_{a,0}) = (\pi_a, \eta_{a,0})$. In particular, if the double sampling probabilities $\eta_{a,0}$ are known by design, then $\text{Bias}_{\tau_a}(\tilde{P}; P) = 0$ if $(\tilde{\mu}_{a,R}, \tilde{\mu}_{a,S}, \tilde{\gamma}_a) = (\mu_{a,R}, \mu_{a,S}, \gamma_a)$ or $\tilde{\pi}_a = \pi_a$. Similarly, $\text{Bias}_{\tau_a^*}(\tilde{P}; P) = 0$ if $\tilde{\mu}_{a,\text{MAR}} = \mu_{a,\text{MAR}}$ or $(\tilde{\gamma}_a, \tilde{\pi}_a, \tilde{\eta}_{a,0}) = (\gamma_a, \pi_a, \eta_{a,0})$. This robustness extends to the rate of convergence of both estimators (as in Rotnitzky et al. (2020)), as the following corollary — an immediate con-
sequence of Theorem 2 — makes concrete.

**Corollary 1.** Suppose the assumptions of Theorem 2 hold, along with the positivity conditions \( \pi_a, \eta_{a,0} \geq \epsilon \) and \( \hat{\pi}_{a,k}, \hat{\eta}_{a,0,k} \geq \epsilon \), for all \( k \), \( P \)-almost surely, where the subscript \( k \) indicates a model fit using \( I_k \). If for each \( k \in \{1, \ldots, K\} \),

\[
\|\hat{\pi}_{a,k} - \pi_a\| \cdot \|\hat{\mu}_{a,k} - \mu_a\| + \|\hat{\eta}_{a,0,k} - \eta_{a,0}\| \cdot \|\hat{\mu}_{a,S,k} - \mu_{a,S}\| = o_P(n^{-1/2}),
\]

then \( \sqrt{n}(\hat{\tau}_a - \tau_a(P)) \overset{d}{\to} \mathcal{N}(0,V) \) i.e., \( \hat{\tau}_a \) attains the nonparametric efficiency bound. Likewise, if for each \( k \in \{1, \ldots, K\} \),

\[
\|\hat{\pi}_{a,k}(\hat{\gamma}_{a,k} + (1 - \hat{\gamma}_{a,k})\hat{\eta}_{a,0,k}) - \pi_a(\gamma_a + (1 - \gamma_a)\eta_{a,0})\| \cdot \|\hat{\mu}_{a,MAR,k} - \mu_{a,MAR}\| = o_P(n^{-1/2}),
\]

then \( \sqrt{n}(\hat{\tau}_a^* - \tau_a^*(P)) \overset{d}{\to} \mathcal{N}(0,V^*) \), i.e., \( \hat{\tau}_a^* \) attains the semiparametric efficiency bound under Assumption 1.

A particularly appealing feature of \( \hat{\tau}_a \) is that, by Corollary 1, we only require that the sum of a product of error terms converges to zero in \( L_2(P) \)-norm for \( \hat{\tau}_a \) to be \( n^{-1/2} \)-consistent. Therefore, we can afford to use more flexible methods for these possibly complicated nuisance functions. Another appealing consequence of the asymptotic normality proved in Theorem 2 is that simple Wald-type asymptotically valid confidence intervals are immediately available. For example, we can estimate \( \text{Var}(\hat{\tau}_a) \) with \( \hat{V} = n \cdot \hat{\text{Var}}(\hat{\tau}_a) \), where

\[
n \cdot \hat{\text{Var}}(\hat{\tau}_a) = \frac{1}{n} \sum_{k=1}^K \sum_{i \in I_k} \left( \hat{\tau}_a(O_i; \hat{P}_k) \right)^2,
\]

with corresponding Wald-type confidence interval given by \( \hat{\tau}_a \pm z_{1-\alpha/2} \sqrt{\hat{\text{Var}}(\hat{\tau}_a)} \), where \( z_\alpha \) denotes the \( \alpha \)-quantile of the standard normal distribution. The variance estimator \( \hat{V}^* = n \cdot \hat{\text{Var}}(\hat{\tau}_a^*) \) and corresponding confidence interval can be constructed entirely analogously.
3.2 Adaptive estimator

As described in section 2.3, in practice one may not be sure whether Assumption 1 is plausible. If in fact this assumption does not hold, then naively using $\hat{\tau}_a$ may result in substantial bias. In general, in Appendix B we show that

\[
\tau^*_a(P) - \tau_a(P) = \mathbb{E}_P \left( \gamma_a(L) \frac{(1 - \gamma_a(L))(1 - \eta_{a,0}(L))}{1 - (1 - \gamma_a(L))(1 - \eta_{a,0}(L))} (\mu_{a,R}(L) - \mu_{a,S}(L)) \right),
\]

which typically we would not expect to equal zero unless with probability 1, it holds that (i) $\mu_{a,R}(L) = \mu_{a,S}(L)$, (ii) $\gamma_a(L) = 0$, or (iii) $(1 - \gamma_a(L))(1 - \eta_{a,0}(L)) = 0$. In other words, we expect bias unless (i) MAR holds, (ii) the outcome is completely unobserved initially, i.e., $R + S = 1 \iff S = 1$ so that $\mu_a = \mu_{a,S} = \mu_{a,\text{MAR}}$, or (iii) we eventually observe the outcome on all individuals, i.e., $P[R = 0, S = 0 \mid L, A = a] = 0$ so that $P[R + S = 1 \mid L, A = a] = 1$.

Thus, as a middle ground between $\hat{\tau}_a$ and $\hat{\tau}^*_a$, we consider estimating $\tau_a(P)$ using a data-adaptive estimator, employing the method developed in Rothenhäusler (2020), described in section 2.3. Letting $\hat{Q} = \hat{V} + \hat{V}^* - \frac{2}{n} \sum_{k=1}^K \sum_{i \in I_k} \hat{\tau}_a(O_i; \hat{P}_k) \hat{\tau}^*_a(O_i; \hat{P}_k)$ be an estimator of $\text{Var}_P(\hat{\tau}_a(O; P) - \hat{\tau}^*_a(O; P))$, the adaptive estimator, which we will denote $\hat{\tau}^\dagger_a$, is defined as follows:

\[
\hat{\tau}^\dagger_a := \begin{cases} 
\hat{\tau}_a, & \text{if } \hat{V} < \max \left\{ n(\hat{\tau}_a - \hat{\tau}^*_a)^2 - \hat{Q}, 0 \right\} + \hat{V}^* \\
\hat{\tau}^*_a, & \text{otherwise}
\end{cases}
\]

We reiterate that Rothenhäusler (2020), in their Theorem 4, derives asymptotically valid confidence intervals for the parameter of interest, in our case $\tau_a(P)$, based on $\hat{\tau}^\dagger_a$. 

17
4 Simulation Study

In this section, we describe the results of a simulation study. In the first of two substudies, we demonstrate the validity of the double sampling approach for handling MNAR data, verify the the robustness properties of the proposed nonparametric influence function-based estimator $\hat{\tau}_a$, and compare, under differing degrees of violation of MAR, the bias and variance of $\hat{\tau}_a$ to $\hat{\tau}_a^*$ as well as approaches that only use the initially observed data. In the second simulation substudy, we compare in the absence of model misspecification the performance of $\hat{\tau}_a$, $\hat{\tau}_a^*$, $\hat{\tau}_a^\dagger$, and an ad hoc adaptive approach, over a grid of possible violations of MAR. We also assess the coverage and length of the proposed confidence intervals for these estimators.

4.1 Robustness, bias and variance

The framing of the simulation study is, following our motivating study in section 2, a hypothetical study comparing two bariatric surgery procedures on long-term weight outcomes (Arterburn et al., 2021). Specifically, we consider a binary point exposure $A$, taking on a value of 0 for Roux-en-Y gastric bypass (RYGB) and 1 for vertical sleeve gastrectomy (VSG), and continuous outcome $Y$ of the proportion weight change at 5 years post-surgery. For simplicity in this simulation, we consider only one confounder, that being gender, denoted $L_g$. The estimand of interest is taken to be the average treatment effect, $\tau_1(P)-\tau_0(P)$.

To help ground the simulation in a real-world setting, we used information on 5,693 patients who underwent either VSG or RYGB at Kaiser Permanente Washington between January 1, 2008, and December 31, 2010. For these patients, complete information was available on gender, bariatric surgery procedure, and weight outcomes, so that missingness in the outcome could then be induced by a known mechanism. We
then generated 1,000 simulated datasets of size \( n = 5,693 \) under each of three settings, where we varied the strength of the violation of MAR. Specifically, we proceeded by (i) sampling directly from the empirical distribution of \( L_g \); (ii) generating \( A \mid L_g \sim \text{Bernoulli}(p_1 L_g + p_0(1 - L_g)) \), where \( p_0 = 0.20 \) and \( p_1 = 0.34 \) were taken to approximately mirror their empirical values; (iii) generating \( R \mid L_g, A \sim \text{Bernoulli}(\expit(\delta_0 + \delta_L L_g + \delta_A A + \delta_{LA} L_g A)) \), where \((\delta_0, \delta_L, \delta_A, \delta_{LA}) = (-1.39, 0.09, -0.05, -0.35)\) and \( \expit(x) = \exp(x)/(1 + \exp(x)) \), inducing a marginal missingness probability \( P[R = 0] \approx 0.8 \); (iv) generating \( Y \mid L_g, A, R \sim \mathcal{N}(\beta_0 + \beta_L L_g + \beta_A A + \beta_{RA} RA, \sigma_Y^2) \), where \((\beta_0, \beta_L, \beta_A) = (-0.24, 0.023, 0.064)\), \( \beta_{RA} \in \{0, 0.016, 0.032\} \) and \( \sigma_Y = 0.11 \), approximately mirroring the marginal empirical distribution of \( Y \); and finally (v) generating the double sampling indicators \( S \mid L_g, A, R \sim \text{Bernoulli}((1 - R)\expit\{\zeta_0 + \zeta_L L_g + \zeta_A A + \zeta_{LA} L_g A\}) \), where \((\zeta_0, \zeta_L, \zeta_A, \zeta_{LA}) = (-2.2, 0.4, 0.3, 0.25)\), inducing a marginal double sampling probability \( P[S = 1 \mid R = 0] \approx 0.11 \). The parameter \( \beta_{RA} \) controls the degree to which the MAR assumption is violated: when \( \beta_{RA} = 0.032 \), there is a large violation, when \( \beta_{RA} = 0.016 \), there is a moderate violation, and when \( \beta_{RA} = 0 \), then MAR holds.

In all scenarios for \( \beta_{RA} \), we computed the nonparametric influence function-based estimator \( \hat{\tau}_1 - \hat{\tau}_0 \) described in section 3.1, where we plugged in the maximum likelihood estimators of the true generating models \( \pi_a, \gamma_a, \mu_{a,R}, \mu_{a,S} \) described above, and assumed the double sampling probabilities \( \eta_{a,0} \) were known. To verify the theoretical robustness of our influence function-based estimator, we considered misspecifying (i) models \((\hat{\mu}_{a,S}, \hat{\mu}_{a,R}, \hat{\gamma}_a)\), (ii) the model \( \hat{\pi}_a \), and (iii) both \((\hat{\mu}_{a,S}, \hat{\mu}_{a,R}, \hat{\gamma}_a) \) and \( \hat{\pi}_a \). In particular, \( \hat{\mu}_{a,S}, \hat{\mu}_{a,R} \) were misspecified by omitting the main effect of \( L_g \), \( \hat{\gamma}_a \) by omitting the main effect of \( A \) and its interaction with \( L_g \), and \( \hat{\pi}_a \), quite drastically, by estimating \( P[A = a \mid L_g] \) using \( \hat{P}[A = 1 - a \mid L_g] \).

For comparison, we also computed: (1) the estimator \( \hat{\tau}_1^* - \hat{\tau}_0^* \), based on the influence
functions $\hat{\tau}_a^*$ described in section 3.1 that are efficient under MAR; and (2) estimators that did not make use of the second-stage outcomes. We acknowledge that there are very many approaches one might consider for analyzing the data using only the initially observed data, but decided that a reasonable analyst might assume MAR, and proceed by targeting $\xi_1(P) - \xi_0(P)$ (introduced in section 2.3) with an outcome regression based estimator based on the $g$-formula (i.e., averaging $\mu_{a,R}$ over the empirical distribution of $L_g$), an inverse-probability weighted (IPW) estimator with missingness-treatment weights $\pi_{a}\gamma_{a}$, as described in Ross et al. (2021), or an augmented-IPW estimator combining both approaches as in Davidian et al. (2005) and Williamson et al. (2012). We pitted each of these estimators (using the correct models for $\mu_{a,\text{MAR}}, \mu_{a,R}, \pi_{a},$ and $\gamma_{a}$) against our influence function-based estimator in all three scenarios.

The results of the simulation study are presented in Figure 1. The robustness of the influence-function based estimator $\hat{\tau}_1 - \hat{\tau}_0$ is clearly seen, as unbiased inference was obtained in all scenarios when all models were correctly specified, or either ($\hat{\mu}_{a,S}, \hat{\mu}_{a,R}, \hat{\gamma}_{a}$) or $\hat{\pi}_{a}$ was misspecified. When both were misspecified, some bias was observed in all three MAR violation scenarios. The initial-sample-only MAR-based estimators had slightly lower variance, but were substantially biased when there was even a moderate violation of MAR. The MAR-efficient estimator $\hat{\tau}_1^* - \hat{\tau}_0^*$, as expected, had the lowest variance of all estimators and was unbiased in the MAR scenario. When there was a moderate or large violation of MAR, $\hat{\tau}_1^* - \hat{\tau}_0^*$ was biased, though less so than the initial-sample-only MAR-based estimators.

4.2 Inference and assessment of adaptive estimator

Within the same simulation framework, we also assessed the performance of the adaptive estimator described in section 3.2, and evaluate proposed confidence intervals of all the
estimators considered. For each value in a grid of MAR violation parameters $\beta_{RA} \in [0, 0.04]$, we simulated 5,000 datasets exactly as in the previous section. In each case, we computed both $\hat{\tau}_1 - \hat{\tau}_0$ and $\hat{\tau}_1^* - \hat{\tau}_0^*$, where all underlying nuisance models were correctly specified. Based on these, we then also computed $\hat{\tau}_1^\dagger - \hat{\tau}_0^\dagger$ described in section 3.2.

Lastly, to show that care is required when using the data to decide between $\hat{\tau}_1 - \hat{\tau}_0$ and $\hat{\tau}_1^* - \hat{\tau}_0^*$, we contrasted $\hat{\tau}_1^\dagger - \hat{\tau}_0^\dagger$ to an ad hoc adaptive estimator. For this, we first test the hypothesis that $\tau_a(P) = \tau_a^*(P)$ by assessing the magnitude of the difference $\hat{\tau}_a - \hat{\tau}_a^*$. Formally, under appropriate conditions, Theorem 3 implies that
\[
\sqrt{n}(\hat{\tau}_a - \hat{\tau}_a^*) \xrightarrow{d} \text{MAR} N(0, Q),
\]
where $Q = \text{Var}_P(\hat{\tau}_a(O; P)) + \text{Var}_P(\hat{\tau}_a^*(O; P)) - 2 \cdot \text{Cov}_P(\hat{\tau}_a(O; P), \hat{\tau}_a^*(O; P))$. An ad hoc adaptive estimator is simply to choose $\hat{\tau}_a$ if we reject a test of MAR based on this result, i.e., if $\sqrt{n}|\hat{\tau}_a - \hat{\tau}_a^*|/\sqrt{Q} > z_{1-\alpha/2}$, and otherwise choose $\hat{\tau}_a^*$ if we fail to reject. We computed this estimator across all simulation settings.

To evaluate confidence intervals, we again focused on the three parameter values $\beta_{RA} \in \{0, 0.016, 0.032\}$. In the 5,000 simulated datasets for each value, we constructed confidence intervals for the four estimators described above. For $\hat{\tau}_1 - \hat{\tau}_0$ and $\hat{\tau}_1^* - \hat{\tau}_0^*$, we used the influence function-based Wald-type confidence interval outlined at the end of section 3.1. For $\hat{\tau}_1^\dagger - \hat{\tau}_0^\dagger$, we constructed confidence intervals based on Theorem 4 of Rothenhäusler (2020). For the ad hoc adaptive estimator, we used the Wald-type interval corresponding to the baseline estimator chosen according to the hypothesis test — a naive approach which we expect will lead to undercoverage.

The results on the grid of $\beta_{RA}$ values are shown in Figure 2. When $\beta_{RA} = 0$ (i.e., MAR holds), all estimators are unbiased, $\hat{\tau}_1^* - \hat{\tau}_0^*$ is most efficient, and the two adaptive estimators have variance somewhere between that of $\hat{\tau}_1^* - \hat{\tau}_0^*$ and $\hat{\tau}_1 - \hat{\tau}_0$. As $\beta_{RA}$ increases, the bias of $\hat{\tau}_1^* - \hat{\tau}_0^*$, which wrongly assumes MAR, increases roughly linearly. The two adaptive
estimators also inherit some bias due to being pulled away by \( \hat{\tau}^*_1 - \hat{\tau}^*_0 \). Interestingly, when \( \beta_{RA} \) becomes really large, indicating quite a substantial violation of MAR, the bias of the two adaptive estimators returns back towards zero, as it becomes increasingly rare for either of these to select the estimator which assumes MAR.

The results on the focused set of values \( \beta_{RA} \in \{0, 0.016, 0.032\} \) are arranged in Table A1 (in Appendix C). The confidence interval for \( \hat{\tau}_1 - \hat{\tau}_0 \) has the appropriate coverage in all scenarios, as does the interval for \( \hat{\tau}^*_1 - \hat{\tau}^*_0 \) when MAR holds. In the two MNAR settings, however, \( \hat{\tau}^*_1 - \hat{\tau}^*_0 \) is biased and its confidence interval is off target. The confidence interval for the adaptive estimator \( \hat{\tau}^\dagger_1 - \hat{\tau}^\dagger_0 \) also appears to be valid, with perhaps a bit of undercoverage in finite samples for moderately large values of \( \beta_{RA} \). Finally, the naive confidence intervals of the ad hoc adaptive estimator tend to be overly narrow, as expected.

5 Double sampling for arbitrary coarsening

Up until now, we have focused on the specific causal problem outlined in section 2. It turns out that the nonparametric identification and estimation results are entirely generic, and do not depend on either the data structure of the given problem or the specific mean counterfactual estimand of interest. In this section, we lay out the notation for arbitrary coarsening of a given full data structure and show that under a generalization of Assumptions 2 and 3, double sampling identifies the complete data distribution; derive a transformation of the full data nonparametric influence function of an arbitrary smooth functional that yields the observed data nonparametric influence function; construct influence function-based estimators using sample splitting; and characterize the asymptotic behavior of these estimators, including multiple robustness properties.
5.1 Coarsened data and nonparametric identification

Suppose the desired full data for a given problem are the random vector \( X \sim P_X^* \). That is, with \( X \) observed on every subject in a random sample, a parameter of interest, say \( \chi(P_X^*) \), could be estimated consistently. Suppose, however, that the initially observed data consists only of \((C, \sigma_C(X))\), where \( C \in \mathbb{N} \) is a coarsening random variable, and \( \sigma_C(X) \) is a coarsened version of the full data: \( \sigma_k \) is some (typically many-to-one) function for every possible value \( k \) of \( C \). As in Tsiatis (2007), we will write \( C = \infty \) to denote that the full data are observed, i.e., there is no coarsening. We will further assume that there exist functions \( \sigma_k \) for every \( k \), such that \((\sigma_k, \sigma_k)\) is injective; that is, there exist functions \( h_k \) with \( h_k(\sigma_k(X), \sigma_k(X)) = X \). In our previous observational example from section 2.1, \( C = R \cdot \infty \), where \( 0 \cdot \infty = 0 \), and \( \sigma_0(X) = (L, A), \sigma_0(X) = Y \).

We now suppose that a subsample is intensively followed up, and the initially unobserved data \( \sigma_C(X) \) are obtained on some subjects. Let \( S \in \{0, 1\} \) indicate successful follow-up in the subsample when \( S = 1 \). The complete data are \((C, \sigma_C(X), S, \sigma_C(X)) \sim P^* \), and the observed data are independent and identically distributed copies of \( O = (C, \sigma_C(X), S, \sigma_C(X)) \sim P \). Here, as before, the observed data probability distribution \( P \) and the full data distribution \( P_X^* \) are induced by \( P^* \). Henceforth, we suppose that the data at hand are a random sample \( O_1, \ldots, O_n \overset{\text{iid}}{\sim} P \).

Let \( p = \frac{dP}{d\mu} \) and \( p^* = \frac{dP^*}{d\mu} \) denote the densities for \( P \) and \( P^* \), respectively, both with respect to some dominating measure \( \mu \). The density of the complete data distribution can be factored via \( p^*(C, \sigma_C(X), S, \sigma_C(X)) = p(C, \sigma_C(X), S) \times \prod_k p^*(\sigma_k(X) \mid C = k, \sigma_k(X), S) \mathbb{1}(C = k) \), whereas the density of the observed data can be factored via \( p(O) = p(C, \sigma_C(X), S) \times \prod_k p(\sigma_k(X) \mid C = k, \sigma_k(X), S = 1) \mathbb{1}(S = 1, C = k) \). As the conditioning event \( S = 0 \) is possible in the complete data but not in the observed data (i.e.,
appears in expression for $p^*$ but not for $p$), $P^*$ will not be identified from the observed data distribution $P$ unless further assumptions are made. That said, analysis of the components of the complete data density $p^*$ that are not present in $p$ motivates the following conditions, which generalize Assumptions 2 and 3.

**Assumption 4** (No informative second-stage missingness, general version). For all $k \neq \infty$, $S \perp \sigma_k(X) \mid C = k, \sigma_k(X)$.

**Assumption 5** (Positivity of second-stage sampling probabilities, general version). For some $\epsilon > 0$, $P[P[S = 1 \mid C = k, \sigma_k(X)] \geq \epsilon] = 1$, for all $k \neq \infty$.

By the following result, a generalization of Proposition 1, Assumptions 4 and 5 are sufficient to identify the complete data distribution $P^*$.

**Proposition 2.** Assumptions 4 and 5 are sufficient to identify the complete data distribution $P^*$ from the observed data distribution $P$.

*Proof.* By Assumption 4, the complete data distribution may be factorized via

$$p^*(C, \sigma_C(X), S, \overline{\sigma}_C(X)) = p(C, \sigma_C(X), S)p^*(\overline{\sigma}_C(X) \mid C, \sigma_C(X), S)$$

$$= p(C, \sigma_C(X), S)p(\overline{\sigma}_C(X) \mid C, \sigma_C(X), S = 1) \quad (1)$$

which only depends on the observed data distribution $p$. Note that we may safely introduce $S = 1$ in the conditioning event due to Assumption 5, and we may ignore the vacuous case $C = \infty$ as $\sigma_\infty(X) = X$, given which $\overline{\sigma}_\infty(X)$ is degenerate. \(\square\)

The identifying Assumption 4 may be interpreted as asserting that whether or not a subject is successfully double sampled is independent of all the initially unobserved data, conditional on all the initially observed data. While we have circumvented the need for
the usual coarsening at random assumption for the initial sample, it is worth noting that
given only observed data \( O \), Assumption 4 is untestable. In practice, however, these can
be ensured by certain study designs and the successful follow-up of the chosen subsample:
(i) subsample selection completely at random prior to the study; (ii) subsample selected at
random among those with any initially missing information; and (iii) subsample selected
with investigator-defined probabilities depending only on \( (C, \sigma_C(X)) \). Of course, successful
follow-up of the entire intended subsample may not be possible, and in these cases it must
be that initially observed data is sufficient to predict successful follow-up. In general, if
the same method of contacting subjects is used at the first stage and second stage of data
collection, then it may be unreasonable to assume that coarsening at random fails to hold
but Assumption 4 is valid. Thus, the double sampling approach may be most justifiable
when the method of data collection at the second stage differs from the first, e.g., in the
EHR example, where the initial sample is the data that happened to be recorded in the
electronic record, and the subsample is followed up via telephone or in-depth chart review.

On the other hand, Assumption 5 asserts that there are no subpopulations, defined by
observed data patterns, that are systematically excluded from the double sampling strategy
(other than those with initially complete data) — if this were not the case, one could not
learn about the subpopulations with initial missing information that were not followed up.

5.2 Estimation of full data parameters

Suppose interest lies in estimating the full data parameter \( \chi(P^*_X) \in \mathbb{R} \), viewed as a func-
tional from a model space of probability distributions on \( X \) — to which \( P^*_X \) belongs —
to the real line. By Proposition 2, under Assumptions 4 and 5, any full data functional
\( \chi(P^*_X) \) has a corresponding observed data functional representation \( \tau(P) \). For example, if
\[ \chi(P_X) = \mathbb{E}_{P}(g(X)), \]  
for some function \( g \), then \( \chi(P_X) = \mathbb{E}_P(\mathbb{E}_{P^*}(g(X) \mid C, \sigma_C(X))) = \mathbb{E}_P(\mathbb{E}_P(g(X) \mid C, \sigma_C(X), S = 1)) =: \tau(P). \) It will often be the case that if the full data \( X \) were completely observed, one would have in mind a valid estimator of \( \chi(P_X) \). A natural goal is thus to develop a general procedure that can in a certain sense transform a full-data estimator into one that uses only the observed data \( O_1, \ldots, O_n \). The following proposition, proved in Appendix A (and a special case of the general theory developed in Robins et al. (1994)), is the key semiparametric-theoretical result that will facilitate such a procedure.

**Proposition 3.** Suppose \( \chi \) is pathwise differentiable \(^1\) with respect to the full data model at \( P_X^* \), with influence function \( \dot{\chi}(X; P_X^*) \) (with respect to maximal tangent space), and that Assumptions 4 and 5 hold. Then \( \tau(P) = \chi(P_X) \) is pathwise differentiable with influence function \( \dot{\tau}(O; P) = \nu_C(\sigma_C(X)) + \frac{S}{\eta(C, \sigma_C(X))} \{ \dot{\chi}(X; P_X^*) - \nu_C(\sigma_C(X)) \} \) at \( P \), where \( \nu_C(\sigma_C(X)) = \mathbb{E}_P(\dot{\chi}(X; P_X^*) \mid C, \sigma_C(X), S = 1) \), and \( \eta(C, \sigma_C(X)) = P[S = 1 \mid C, \sigma_C(X)] \).

Recalling that \( \sigma_\infty(X) = X \), we allow for the possibility that \( \eta(\infty, X) = 0 \), and define \( \dot{\tau}(O; P) \) to equal \( \nu_\infty(X) = \dot{\chi}(X; P_X^*) \) when \( C = \infty \).

We remark that another way to interpret the term \( \frac{S}{\eta(C, \sigma_C(X))} \) in the case that \( \eta(\infty, X) = 0 \) is to use the convention \( 0 \cdot \infty = 0 \), so that the second term drops out.

We are now equipped to define a one-step estimator of the general functional \( \tau(P) \) that uses its estimated influence function to correct the bias of a plugin estimator \( \tau(\hat{P}) \). Depending on the form of the parameter and its influence function, certain components of the observed data distribution may not need to be estimated (e.g., see the causal example in section 3.1). In general, though, an estimate of \( P_X^* \) can be reconstructed by marginalizing an estimated version of the identified complete data density (1) over \((C,S)\). Letting \( \lambda_C(\sigma_C(X); \sigma_C(X)) \) be the distribution function of \( \sigma_C(X) \) given \( C, \sigma_C(X), S = 1 \), we can

---

\(^1\)see e.g. Bickel et al. (1993, Chapter 3) for precise definitions.
use $\hat{P}_X^s, \hat{\eta}, \hat{\lambda}$ to obtain $\hat{\tau}(O; \hat{P}) = \hat{\nu}_C(\sigma_C(X)) + \frac{S}{\eta_C, \sigma_C(X)} \left\{ \hat{\chi}(X; \hat{P}_X^s) - \hat{\nu}_C(\sigma_C(X)) \right\}$, where $\hat{\nu}_C(\sigma_C(X)) = \int \hat{\chi}(h_C(\sigma_C(X), t; \hat{P}_X^s)) d\hat{\lambda}_C(t; \sigma_C(X))$.

As in section 3.1, we propose to use sample splitting and cross-fitting (Chernozhukov et al., 2018), and fit $\hat{P}_k$ using data $I^*_k$, for $k = 1, \ldots, K$. We then define $\hat{\tau}_k = \tau(\hat{P}_k) + \frac{K}{n} \sum_{i \in I^*_k} \hat{\tau}(O_i; \hat{P}_k)$, for $k = 1, \ldots, K$, so that the sample-split influence function-based estimator is given by $\hat{\tau} = \frac{1}{K} \sum_{k=1}^{K} \hat{\tau}_k$.

### 5.3 Consistency, asymptotic normality, and robustness

The following result, proved in Appendix A, is the basis for consistency, asymptotic normality, nonparametric efficiency, and multiple robustness of the proposed nonparametric influence function-based estimator $\hat{\tau}$.

**Theorem 3.** Suppose $\left\| \hat{\tau}(.; \hat{P}_k) - \hat{\tau}(.; P) \right\| = o_P(1)$ for $k = 1, \ldots, K$. Then

$$\hat{\tau} - \tau(P) = O_P \left( \frac{1}{\sqrt{n}} + \frac{1}{K} \sum_{k=1}^{K} \text{Bias}_r(\hat{P}_k; P) \right),$$

where $\text{Bias}_r(\hat{P}; P) = \mathbb{E}_P(\hat{\tau}(O; \hat{P})) + \tau(\hat{P}) - \tau(P)$ for any $\hat{P}$. Moreover, if $\text{Bias}_r(\hat{P}_k; P) = o_P(n^{-1/2})$, for $k = 1, \ldots, K$, then $\sqrt{n}(\hat{\tau} - \tau(P)) \xrightarrow{d} \mathcal{N}(0, V)$, where $V = \text{Var}_P(\hat{\tau}(O; P))$ is the nonparametric efficiency bound.

It is important to note that we have only established that $\hat{\tau}$ is fully (semiparametric) efficient in a nonparametric model, i.e., when the model tangent space (see Bickel et al. (1993), Tsiatis (2007)) is $L_2^0(P)$, consisting of all mean-zero functions of $O$ with finite variance. Nonparametric efficiency of $\hat{\tau}$ is guaranteed because there is a unique (thus efficient) influence function in a nonparametric model, its variance equal to the nonparametric efficiency bound (Bickel et al., 1993). Derivation of the semiparametric efficient observed
data influence function in proper semiparametric models where restrictions are placed on
$P$ will depend on the form of those restrictions (e.g., see MAR example in section 3.1), so
we leave characterizations of efficiency over classes of restrictions on $P$ for future research.

In Appendix A, we present two further propositions, A1 and A2, and provide a simple
asymptotic variance estimator for $\hat{\tau}$. The propositions relate the asymptotic variance and
bias of $\hat{\tau}$ to the corresponding quantities of a full-data influence function-based estimator.
We also discuss how the latter provides the basis for the multiple robustness properties of
the proposed estimators.

6 Discussion

In this paper, we consider causal effect estimation in the context of an observational study
subject to missing outcomes, in which double sampling recovers some missing outcome in-
formation. We construct a nonparametric efficient estimator, and a semiparametric efficient
estimator under the assumption that the outcomes are initially MAR, and compare these
to an adaptive approach that selects between these two estimators. Simulations evaluate
the bias and variance of these estimators under differing degrees of violations of MAR,
verify the robustness of the proposed nonparametric efficient estimator, and validate in-
tended coverage of proposed confidence intervals. Moreover, we propose the first general
framework for identification, estimation, and inference in settings where double sampling is
used for handling data subject to potentially informative coarsening. Finally, we develop a
general method for constructing consistent, asymptotically normal, nonparametric efficient
and often multiply robust estimators of arbitrary smooth functionals.

Based on our theoretical and simulation-based results, we believe that double sampling
is a very promising approach for handling informatively missing data in a variety of settings,
but especially in EHR-based studies. In practice, it has been successfully carried out via telephone-based surveys (Haneuse et al., 2016), but alternatively, augmented data obtained via chart review, an increasingly popular approach (Zhang et al., 2019; Weiskopf et al., 2019; Yin et al., 2022), could also be considered within this paradigm. Due to the reliance of the identification results and ensuing estimators on non-informative selection at the second stage (i.e., Assumption 4), it is critical to try and ensure this assumption holds when implementing these designs. In light of this, it may be worthwhile to spend more resources on following up on a smaller subsample but with greater probability of success, than to try and obtain a larger follow-up subsample. A cost-benefit / bias-variance tradeoff analysis for allocating resources to the first stage versus the second stage of sampling will be important for investigators wishing to apply double sampling.

Beyond exact identification, in the situation where Assumption 4 may not hold, it may be possible to use the second-stage data to improve nonparametric bounds on parameters of interest. Indeed, Aronow et al. (2015) and Coppock et al. (2017) show that this is the case in estimating treatment effects in simple randomized trials. Extending these ideas to more general estimands and coarsening patterns is an important problem that we are pursuing in upcoming work.

Finally, in the general coarsening setting of section 5, there is potential for improved efficiency if we believe restrictions hold on $P_X$. For example, if certain conditional independence relations hold among the variables $X$, then this could potentially be exploited to gain efficiency. This is a challenging problem with many implications for study design within this paradigm, and we will pursue this in future research.
References

Amorim, G., R. Tao, S. Lotspeich, P. A. Shaw, T. Lumley, and B. E. Shepherd (2021). Two-phase sampling designs for data validation in settings with covariate measurement error and continuous outcome. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184(4), 1368–1389.

An, M.-W., C. E. Frangakis, B. S. Musick, and C. T. Yiannoutsos (2009). The need for double-sampling designs in survival studies: an application to monitor pepfar. *Biometrics* 65(1), 301–306.

Aronow, P. M., A. S. Gerber, D. P. Green, H. Kern, and M. J. LaCour (2015). Double sampling for nonignorable missing outcome data in randomized experiments.

Arterburn, D. E., E. Johnson, K. J. Coleman, L. J. Herrinton, A. P. Courcoulas, D. Fisher, R. A. Li, M. K. Theis, L. Liu, J. R. Fraser, et al. (2021). Weight outcomes of sleeve gastrectomy and gastric bypass compared to nonsurgical treatment. *Annals of Surgery* 274(6), e1269–e1276.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press Baltimore.

Borgan, Ø., N. Breslow, N. Chatterjee, M. H. Gail, A. Scott, and C. J. Wild (2018). *Handbook of statistical methods for case-control studies*. CRC Press.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Chatterjee, N. and S. Wacholder (2002). Validation studies: bias, efficiency, and exposure assessment. *Epidemiology* 13(5), 503–506.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1), C1–C68.

Coppock, A., A. S. Gerber, D. P. Green, and H. L. Kern (2017). Combining double sampling and bounds to address nonignorable missing outcomes in randomized experiments. *Political Analysis* 25(2), 188–206.

Daniels, M. J. and J. W. Hogan (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. chapman and hall/CRC.
Davidian, M., A. A. Tsiatis, and S. Leon (2005). Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical science: a review journal of the Institute of Mathematical Statistics* 20(3), 261.

Elliott, M. R., R. J. Little, and S. Lewitzky (2000). Subsampling callbacks to improve survey efficiency. *Journal of the American Statistical Association* 95(451), 730–738.

Frangakis, C. E. and D. B. Rubin (2001). Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring. *Biometrics* 57(2), 333–342.

Geng, E. H., D. R. Bangsberg, N. Musinguzi, N. Emenyonu, M. B. Bwana, C. T. Yiannoutsos, D. V. Glidden, S. G. Deeks, and J. N. Martin (2010). Understanding reasons for and outcomes of patients lost to follow-up in antiretroviral therapy programs in Africa through a sampling-based approach. *Journal of acquired immune deficiency syndromes (1999)* 53(3), 405.

Geng, E. H., D. V. Glidden, D. R. Bangsberg, M. B. Bwana, N. Musinguzi, D. Nash, J. Z. Metcalfe, C. T. Yiannoutsos, J. N. Martin, and M. L. Petersen (2012). A causal framework for understanding the effect of losses to follow-up on epidemiologic analyses in clinic-based cohorts: the case of HIV-infected patients on antiretroviral therapy in Africa. *American journal of epidemiology* 175(10), 1080–1087.

Guan, Z., D. H. Leung, and J. Qin (2018). Semiparametric maximum likelihood inference for nonignorable nonresponse with callbacks. *Scandinavian Journal of Statistics* 45(4), 962–984.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66(2), 315–331.

Haneuse, S., A. Bogart, I. Jazic, E. O. Westbrook, D. Boudreau, M. K. Theis, G. E. Simon, and D. Arterburn (2016). Learning about missing data mechanisms in electronic health records-based research: a survey-based approach. *Epidemiology* 27(1), 82.

Haneuse, S. and M. Daniels (2016). A general framework for considering selection bias in EHR-based studies: what data are observed and why? *eGEMs* 4(1), 1203.

Haneuse, S. J. A. and S. M. Shortreed (2017). On the use of electronic health records. In *Methods in Comparative Effectiveness Research*, pp. 469–502. Chapman and Hall/CRC.

Hansen, M. H. and W. N. Hurwitz (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association* 41(236), 517–529.
Heitjan, D. F. and D. B. Rubin (1991). Ignorability and coarse data. *The Annals of Statistics* 19(4), 2244–2253.

Kennedy, E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association* 114(526), 645–656.

Kennedy, E. H. (2020). Efficient nonparametric causal inference with missing exposure information. *The International Journal of Biostatistics* 16(1), 20190087.

Koffman, L., A. W. Levis, D. Arterburn, K. J. Coleman, L. J. Herrinton, J. Cooper, J. Ewing, H. Fischer, J. R. Fraser, E. Johnson, et al. (2021). Investigating bias from missing data in an electronic health records-based study of weight loss after bariatric surgery. *Obesity Surgery* 31(5), 2125–2135.

Malinsky, D., I. Shpitser, and E. J. Tchetgen Tchetgen (2020). Semiparametric inference for non-monotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association* 0(0), 1–22.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review* 80(2), 319–323.

Miao, W., X. Li, and B. Sun (2021). A stableness of resistance model for nonresponse adjustment with callback data.

Miao, W. and E. J. Tchetgen Tchetgen (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika* 103(2), 475–482.

Molenberghs, G., G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke (2014). *Handbook of Missing Data Methodology*. CRC Press.

Qian, T., C. Frangakis, and C. Yiannoutsos (2019). Deductive semiparametric estimation in double-sampling designs with application to PEPFAR. *Statistics in Biosciences* 12(3), 1–29.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* 7(9-12), 1393–1512.

Robins, J. M., A. Rotnitzky, and D. O. Scharfstein (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 1–94. Springer.
Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89(427), 846–866.

Ross, R., A. Breskin, T. Breger, and D. Westreich (2021). Reflection on modern methods: combining weights for confounding and missing data. *International Journal of Epidemiology* 0(0), dyab205.

Rothenhäusler, D. (2020). Model selection for estimation of causal parameters.

Rotnitzky, A., E. Smucler, and J. Robins (2020). Characterization of parameters with a mixed bias property. *Biometrika* 106(4), 875–888.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63(3), 581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, Volume 81. John Wiley & Sons.

Seaman, S. R. and I. R. White (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research* 22(3), 278–295.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

Weiskopf, N. G., A. M. Cohen, J. Hannan, T. Jarmon, and D. A. Dorr (2019). Towards augmenting structured ehr data: a comparison of manual chart review and patient self-report. In *AMIA Annual Symposium Proceedings*, Volume 2019, pp. 903. American Medical Informatics Association.

Williamson, E., A. Forbes, and R. Wolfe (2012). Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder. *Statistics in Medicine* 31(30), 4382–4400.

Yin, Z., J. Tong, Y. Chen, R. A. Hubbard, and C. Y. Tang (2022). A cost-effective chart review sampling design to account for phenotyping error in electronic health records (ehr) data. *Journal of the American Medical Informatics Association* 29(1), 52–61.

Zhang, Y., T. Cai, S. Yu, K. Cho, C. Hong, J. Sun, J. Huang, Y.-L. Ho, A. N. Ananthakrishnan, Z. Xia, et al. (2019). High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (phecap). *Nature protocols* 14(12), 3426–3444.

Zheng, W. and M. J. van der Laan (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pp. 459–474. Springer.
Figure 1: Simulation results for experiments of section 4.1, with a) no violation of MAR ($\beta_{RA} = 0$); b) a moderate violation of MAR ($\beta_{RA} = 0.016$); and c) a large violation of MAR ($\beta_{RA} = 0.032$). IF-DS = influence-function based estimator using double sampling, $\hat{\tau}_1 - \hat{\tau}_0$; IF-DS-PW = IF-DS but with $\hat{\pi}_a$ misspecified; IF-DS-MW = IF-DS but with $(\hat{\mu}_{a,S}, \hat{\mu}_{a,R}, \hat{\gamma}_a)$ misspecified; IF-DS-BW = IF-DS but with both $(\hat{\mu}_{a,S}, \hat{\mu}_{a,R}, \hat{\gamma}_a)$ and $\hat{\pi}_a$ misspecified; IF-MAR = augmented IPW-based estimator assuming MAR, $\hat{\xi}_1 - \hat{\xi}_0$; OR-MAR = outcome regression-based estimator assuming MAR; IPW-MAR = IPW-based estimator assuming MAR; IF-MAR-EFF = semiparametric efficient estimator under MAR, $\hat{\tau}^*_1 - \hat{\tau}^*_0$. ATE = average treatment effect; the red dashed line indicates the true ATE.
Figure 2: Simulation results for experiments of section 4.2, regarding $\hat{\tau}_1 - \hat{\tau}_0$, MAR semiparametric efficient estimator $\hat{\tau}_1^* - \hat{\tau}_0^*$, the adaptive estimator of Rothenhäusler (2020), $\hat{\tau}_1^\dagger - \hat{\tau}_0^\dagger$, and the ad hoc approach described in section 4.1. Subplot a) shows the empirical mean squared error (MSE) of each approach, divided by the MSE of the nonparametric efficient estimator. Subplot b) shows the empirical bias of each approach. Subplot c) shows the empirical variance of each approach, divided by the variance of the nonparametric efficient estimator.