Statistica Sinica Preprint No: SS-2023-0373

Title	Robust and Efficient Case-Control Studies With Contaminated Case Pools: A Unified M-Estimation Framework
Manuscript ID	SS-2023-0373
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0373
Complete List of Authors	Guorong Dai and Jinbo Chen
Corresponding Authors	Guorong Dai
E-mails	guorongdai@fudan.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0373

Statistica Sinica

ROBUST AND EFFICIENT CASE-CONTROL STUDIES WITH CONTAMINATED CASE POOLS: A UNIFIED M-ESTIMATION FRAMEWORK

Guorong Dai[a] and Jinbo Chen[b]

[a]*Fudan University and* [b]*University of Pennsylvania*

Abstract: We consider a general M-estimation problem based on contaminated case-control data, including the primary and secondary analyses of case-control studies as special examples. The case pool contains ineligible patients who should be excluded from the study if known, but the true status of an individual in the case pool is unclear except in a small subset. Through imputing the possibly unobserved status variable with a function of all available relevant predictors, followed by an appropriate debiasing procedure, we exploit the whole sample to develop a family of robust and efficient estimators, eliminating bias from the case contamination. With the help of cross-fitting, the imputation function can be constructed using any reasonable regression or machine learning approaches. Our estimators are always root-n-consistent and asymptotically normal regardless of the imputation function's limit. Further, we explore relaxation of requirements on the imputation function. We show even without any assumption on its convergence properties, our estimators are still root-n-consistent while asymptotic normality can be achieved by a sample-splitting variant. We also demonstrate results of this type, which are entirely free of convergence assumptions on the nuisance estimators, can be extended to other problems involving nuisance functions. The finite-sample superiority of our method is demonstrated by comprehensive simulation studies. We also apply our method to analyze sepsis-related death based on a real data set from electronic health records.

Key words and phrases: Contaminated case pool; Estimating equation; Nuisance estimation; Primary and secondary analyses of case-control data; Robustness and efficiency.

1. Introduction

In epidemiology and many other biomedical fields, case-control designs have been serving as flexible and cost-effective tools for investigating risk factors for conditions of interest, e.g., the occurrence of rare diseases and disease-related mortality. In stark contrast with prospective cohort designs, a case-control sample is assembled by combining two independent subsamples drawn separately from two groups: individuals with (cases) and without (controls) the condition of interest. A detailed overview of case-control methods can be found in Breslow (1996). In biomedical research, case-control data are popularly used for two purposes:

(a) *Primary analysis* that aims to understand how the primary outcome defining the case-control status is associated with a set of covariates. The most frequently used approach is fitting a prospective logistic regression model (Prentice and Pyke, 1979).

(b) *Secondary analysis* that focuses on the relationship between the covariates and a secondary outcome, whose data are also available in the case-control sample defined in terms of the primary outcome. A variety of strategies have been proposed to adapt prospective regression methods to accommodate the case-control sampling scheme; see Tchetgen Tchetgen (2014) for a thorough review of the secondary analysis literature.

1.1 Case contamination in electronic health records

Recent applications of case-control methods to electronic health record (EHR) data have produced many promising results; see, for example, Palen et al. (2012). Containing a wealth of patients' health information, EHRs provide rich resources for clinical and translational studies (Casey et al., 2016). Nonetheless, standard analytical techniques are often *not* suitable for analyzing EHR data because they were collected mainly for purposes other than research (Pathak et al., 2013). Specific to EHR-based case-control studies, a widely recognized challenge is *case contamination*, that is, inclusion of ineligibles in the case pool who cannot be treated as either cases or controls but should be *excluded* from the study. This is an essential difference of our problem from two relevant traditional ones: the outcome misclassification framework and the exponential tilt mixture model; see the clarification in Section S1 of the Supplementary Material. On the other hand, phenotyping information that can help validate eligibility of cases is often known only in a random subset of the whole pool (Klarin et al., 2019). For example, when phenotyping information is not directly available from the records, a common strategy for researchers is randomly drawing a small portion of all cases and validating their eligibility by approaches such as medical chart reviews, which cannot be conducted for the whole case pool due to logistic constraints (Wang et al., 2021). In contrast, a sufficient number of controls can usually be collected with accurate phenotyping information, since there are typically many more controls than cases given, for example, the condition of interest is the occurrence of a rare disease or disease-related mortality. Actually, because the definitions of cases and controls are interchangeable in case-control samples, the method developed in this work can be directly applied to studies with ineligible controls. It can also be easily generalized to handle problems where contamination exists in both the case and control pools. We focus on case-control studies with only case contamination, which is the most common situation in practice, to illustrate our main ideas and simplify notation.

The phenotyping challenge described above can be clearly illustrated by a case-control study on sepsis-related death in Dai et al. (2023), which is based on a data set extracted from the Medical Information Mart for Intensive Care (MIMIC) III (Johnson et al., 2016). The control pool of the study consists of survivors with sepsis who are eligible for the study. But the case pool, which is supposed to contain only patients who died of sepsis-related causes, is actually contaminated by ineligibles, i.e., deceased individuals who died of reasons unrelated to sepsis. As the definition of controls is "having sepsis and surviving", those ineligibles cannot be treated

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0373

as controls but should be excluded from the study. Since the phenotyping information for discerning cases and ineligibles is available only in the "*validation set*", which was verified as a random subset of the case pool, it is infeasible to identify all the sepsis patients followed by applying standard case-control methods. Detailed descriptions and analyses of this data set can be found in Section 6.

1.2 Existing methods and motivations of our work

The case contamination problem illustrated in Section 1.1 is ubiquitous in EHR data, necessitating novel strategies for conducting valid case-control studies using the contaminated samples. In such settings, some recent progress has been made for estimating odds ratio parameters of the primary analysis in light of a feature of the data structure: due to being a random subset of the case pool, the validation set provides information that can be transferred to the nonvalidated *candidate cases*. Here the "candidate cases" is a collective name of cases and ineligibles. Along this line, Wang et al. (2021) pointed out the key role of the *phenotyping model*, which predicts based on the covariates the likelihood of a candidate case being a case, in correcting bias caused by case contamination and making use of the nonvalidated candidate cases to improve estimation efficiency. Instead of attempting to exclude ineligibles, Wang et al. (2021) constructed a weighted estimating equation involving all available data, where the weight of a candidate case is determined by a phenotyping model learned from the validation set. Despite outperforming the naive approach that ignores the contamination, their method undesirably relies on an assumption that the phenotyping model has a logistic form, violation of which can result in inconsistent estimation. This stringent condition was loosened by a recent unbiased estimating equation approach developed in Dai et al. (2023), which possesses full robustness against model misspecification and achieves semiparametric efficiency when the phenotyping model is indeed logistic.

The above-mentioned methods from Wang et al. (2021) and Dai et al. (2023) share a common limitation: they both build a low-dimensional parametric (working) phenotyping model where the number of predictors is much smaller than the validation set size. In EHR data sets, the collection of predictors for discerning cases and ineligibles can be very rich, ranging from demographics, healthcare utilization, labs and prescriptions to co-morbidity statuses, etc. To fully exploit the *predictiveness* of these variables whose number may well exceed the validation set size, it is necessary to allow for high dimensional phenotyping models of flexible forms.

In addition to investigating risk factors for the case-control status as in

Wang et al. (2021) and Dai et al. (2023), another important task in case-control studies is to perform association analysis for co-morbid statuses or other secondary outcomes given the tremendous effort required to prepare an EHR data set for a research study. None of the existing methods is directly applicable when case contamination exists. The paucity of relevant research motivates us to develop a unified theory that can facilitate both the primary and secondary analyses based on the contaminated data.

1.3 Our contributions

We aim to provide a thorough and comprehensive understanding of case-control studies with contaminated case pools. Our theoretical analysis is conducted under a general M-estimation framework, where the target parameters are defined as the solution to a set of estimating equations involving cases and controls only, *without* assumptions on the underlying relation between the covariates and primary or secondary outcome. This highly flexible model-free framework renders our method applicable to inference for (i) *the odds ratio parameters in the primary analysis* and (ii) *the generalized linear model parameters in the secondary analysis*, among others.

For these important parameters in the primary or secondary analysis of case-control studies, we devise a family of estimators that make use of all available data, indexed by a function employed to impute the possibly unobserved status variables of candidate cases. Constructed using the one-step update strategy (Van der Vaart, 2000), our estimators have a simple *closed-form* expression allowing for easy implementation. Under the *high-level* assumptions listed in Section 3.2, we establish in Theorem 1 the $n^{1/2}$-consistency and asymptotic normality of our estimators with n being the number of validated individuals. Thanks to an appropriate debiasing procedure and the use of cross-fitting (Newey and Robins, 2018) in calculating the imputation function as an estimator for the (possibly misspecified) phenotyping model, these properties rely on *neither* (i) correct specification of the phenotyping model *nor* (ii) first-order asymptotics or stochastic equicontinuity conditions of the imputation function. This feature enables us to exploit a wide range of regression and machine learning algorithms for constructing the possibly high dimensional imputation function *without* any knowledge of its asymptotic expansion or specific convergence rate. Moreover, when the phenotyping model is correctly specified by the imputation function, our estimators enjoy *semiparametric efficiency* (Tsiatis, 2007) under appropriate semiparametric models; see Remark 3.

Another notable contribution of this work is rigorous development of inferential theory under minimum assumptions on the nuisance estimator in our method, i.e., the imputation function. We demonstrate in Remark 2 that for deriving the $n^{1/2}$-consistency and asymptotic normality of our estimators, the L_2 convergence of the nuisance estimator is generally sufficient, where the limit does not have to equal the true phenotyping model. This is a fairly mild requirement that holds for a variety of regression and machine learning approaches. Further relaxation of this assumption is considered in Theorem 2, which reveals that given we are *entirely agnostic to asymptotic behaviors of the nuisance estimator whose limit may not even exist*, our estimators are still guaranteed to be $n^{1/2}$-consistent for the target parameters while asymptotic normality can be achieved by a variant resorting to the sample-splitting technique (Cox, 1975). For other estimation problems involving nuisance functions, e.g., semi-supervised inference (Zhang and Bradic, 2022) and treatment effect estimation in randomized experiments (Wager et al., 2016), results of this type, which are *completely free of convergence assumptions on the nuisance estimators*, can also be pursued in a similar way while having not been studied in the existing literature; see Remark 4 and Section S4 in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0373

2. Problem setup

Notations Throughout, the lower case letter c stands for a generic positive constant, including c_1, c_2, etc, which may vary from place to place. For a vector $\mathbf{u}$, we use $\mathbf{u}_{[j]}$ to represent its jth component. The symbols $\|\cdot\|$ and $\lambda_{\min}(\cdot)$ respectively refer to the maximum and minimum singular value of a matrix, while $\mathbb{1}(\cdot)$ is the indicator function. The bold number $\mathbf{0}$ is a zero vector of an appropriate length and $\mathbf{I}$ an identity matrix of an appropriate size. For any random function $\widehat{\mathbf{g}}(\cdot)$ and random vector $\mathbf{U}$ with copies $\{\mathbf{U}_i : i = 1, \ldots, n\}$, denote $\mathbb{E}_{\mathbf{U}|S=s}\{\widehat{\mathbf{g}}(\mathbf{U})\} := \int \widehat{\mathbf{g}}(\mathbf{u})dF_{\mathbf{U}|S=s}(\mathbf{u})$ as the integral of $\widehat{\mathbf{g}}(\cdot)$ with respect to the conditional distribution function $F_{\mathbf{U}|S=s}(\cdot)$ of $\mathbf{U}$ given a binary variable $S = s \in \{0, 1\}$. The symbol $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Data structure To formulate the case contamination problem, we introduce a three-valued variable $D \in \{0, 1, 2\}$ to indicate whether an individual in the sample is (i) an ineligible ($D = 0$), (ii) a case ($D = 1$) or (iii) a control ($D = 2$). Cases and ineligibles are collectively referred to as "*candidate cases*", while the latter should be excluded from the study since our target population consists of cases and controls only. We also let a binary variable $S := \mathbb{1}(D \neq 2)$ represent the status of being a candidate case

($S = 1$) or a control ($S = 0$). To reflect the difficulty in obtaining phenotyping information of candidate cases, we assume S is always observed but the value of D (0 or 1) is available only in a small subset of the candidate case pool, i.e., the *validation set*. In other words, it is clear whether an individual in the sample is a candidate case ($S = 1$) or a control ($S = 0$), but whether a candidate case is a case ($D = 0$) or an ineligible ($D = 1$) is unknown except in the validation set of size n. In a typical case-control study among cases ($D = 1$) and controls ($D = 2$), the primary goal is to study relationship between D and a set of covariates $\mathbf{X} \in \mathbb{R}^p$ where $p \geq 1$ is a fixed integer. We are also interested in how the covariates in $\mathbf{X}$ affect a secondary outcome $Y \in \mathbb{R}$ which is potentially associated with D. Besides $\mathbf{X}$ and Y, there are also records of predictors $\mathbf{X}_*$ available in the data, which are informative for the true status of a candidate case. We allow the dimension d of $\mathbf{Z} := (Y, \mathbf{X}^{\mathrm{T}}, \mathbf{X}_*^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{Z} \subset \mathbb{R}^d$ to diverge and exceed the validation set size n. Our study sample can be written as the union of three mutually independent subsets: (i) the *validation set* $\mathcal{V} := \{(D_i, S_i \equiv 1, \mathbf{Z}_i) : 1 \leq i \leq n\}$ of size n, (ii) the *nonvalidated candidate case pool* $\mathcal{N} := \{(S_i \equiv 1, \mathbf{Z}_i) : n < i \leq N_1\}$ of size $N_1 - n$ and (iii) the *control pool* $\mathcal{C} := \{(S_i \equiv 0, \mathbf{Z}_i) : N_1 < i \leq N\}$ of size $N_0 := N - N_1$, which contain independent copies of base observations $(D, S = 1, \mathbf{Z})$, $(S = 1, \mathbf{Z})$

and $(S = 0, \mathbf{Z})$. In retrospective sampling, the numbers of candidate cases and controls were determined in advance, so that N_1, N_0 and S_i are nonrandom while the proportion of candidate cases, $\tau := N^{-1}\sum_{i=1}^{N} S_i \equiv N_1/N$, may not equal the population counterpart $\eta := \mathbb{E}(S)$.

Remark 1 (Difference from missing data problems)**.** Since typically the validation set is randomly drawn from the candidate case pool (see the first paragraph of Section 1.1), one can roughly treat D as "missing by design" in $\mathcal{V} \cup \mathcal{N}$ from the missing data perspective. The exposition in Chen (2000) indicates the known "missingness" mechanism of D guarantees the validity of transferring information from the validation set to the nonvalidated candidate case pool. Nonetheless, as clarified in Dai et al. (2023), there is a subtle and rather important difference between the traditional missing data framework and our data structure: in our setting, the validation set can be arbitrarily small relative to the whole candidate case pool. That is, for $\delta_n := n/N_1$, we allow $\delta := \lim_{n\to\infty} \delta_n \in [0, 1)$. Since $n < N_1$, the sequence $N_1 \to \infty$ whenever $n \to \infty$. We thus suppress the subscript N_1 in δ_n for brevity. The special case $\delta = 0$ holds when, for example, $N_1 = n^2$. It apparently violates the "positivity assumption" that the proportion of complete observations in the sample is bounded away from zero, which is typically considered inevitable in the missing data literature (Tsiatis, 2007).

In EHR-based case-control studies, the whole candidate case pool size N_1 can be very large, say 10^5, while the typical validation set size n is just several hundred. Then $\delta = 0$ approximately holds.

Target parameters Our target parameter vector $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^p$ is defined as the solution to the following equation:

$$\alpha\,\mathbb{E}\{D\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0) \mid S=1\} + (1-\alpha)\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0) \mid S=0\} \;=\; \mathbf{0} \qquad (2.1)$$

with $\mathbf{W} := (S, Y, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$, $\alpha \in \{\tau \equiv N_1/N, \eta \equiv \mathbb{E}(S)\}$ and $\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}) \in \mathbb{R}^p$ an estimating function that is differentiable with respect to $\boldsymbol{\theta}$. Recalling $S \equiv \mathbb{1}(D \neq 2)$, equation (2.1) actually defines $\boldsymbol{\theta}_0$ *within the population of cases* ($D = 1$) *and controls* ($D = 2$) *only*, since ineligibles ($D = 0$) should be excluded from the study. Except for the basic smoothness and moment conditions on $\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta})$ specified in Section 3.2, we do not impose extra requirements on the distribution of $(D, \mathbf{W})$. Hence, our framework is entirely *model-free*. In Section S2 of the Supplementary Material, practical relevance of the general M-estimation problem (2.1) is illustrated by two important special examples, which correspond to the primary and secondary analyses in case-control studies. We also address the existence, uniqueness and identifiability of $\boldsymbol{\theta}_0$ in the remarks therein.

3. Estimation and inference

Since the status variable D is unknown in the nonvalidated candidate case pool $\mathcal{N}$, we cannot estimate $\boldsymbol{\theta}_0$ by directly constructing the empirical version of (2.1) with the whole study sample. Therefore, we consider replacing D with a function of $\mathbf{Z}$ in equation (2.1) so that $\mathcal{N}$ can be used for estimation. Denote the *phenotyping model* $\mu(\mathbf{Z}) := \mathbb{E}(D \mid \mathbf{Z}, S = 1)$. Recalling $\mathbf{W} \equiv (S, Y, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$ and $(Y, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$ are subvectors of $\mathbf{Z}$, we have $\mathbb{E}[\{D - \mu(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid \mathbf{Z}, S = 1] = \mathbf{0}$, which suggests substituting $\mu(\mathbf{Z})$ for D in (2.1) does not cause bias at the population level. We can thus estimate $\boldsymbol{\theta}_0$ based on the equation $\alpha\,\mathbb{E}\{\mu(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1\} + (1-\alpha)\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 0\} = \mathbf{0}$, which does not involve D. Nevertheless, fitting $\mu(\mathbf{Z})$ fully nonparametrically is usually infeasible considering the dimension d of $\mathbf{Z}$ can be greater than the validation set size n. On the other hand, positing parametric assumptions on the form of $\mu(\cdot)$ will lead to potential model misspecification and estimation bias. These issues highlight the necessity of appropriate debiasing strategies in the construction of estimators involving the nonvalidated candidate cases.

For an arbitrary $\mu^* : \mathbb{R}^d \mapsto \mathbb{R}$ that may *not* equal $\mu(\cdot)$, the following identity always holds: $\mathbf{0} = \boldsymbol{\Phi}(\boldsymbol{\theta}_0) :=$

$$\alpha\,\mathbb{E}\{D\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1\} + (1-\alpha)\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 0\} \qquad (3.1)$$

$$= \alpha\,\mathbb{E}\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0) \mid S=1\} + (1-\alpha)\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0) \mid S=0\} +$$
$$\alpha\,\mathbb{E}[\{D-\mu^*(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0) \mid S=1]. \tag{3.2}$$

The above (3.2) provides an unbiased estimating equation for $\boldsymbol{\theta}_0$ that is robust against misspecification of the phenotyping model $\mu(\cdot)$. Noticing the three expectations $\mathbb{E}\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0) \mid S=1\}$, $\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0) \mid S=0\}$ and $\mathbb{E}[\{D-\mu^*(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0) \mid S=1]$ in (3.2) can respectively be approximated based on the whole candidate case pool $\mathcal{V}\cup\mathcal{N}$, the control pool $\mathcal{C}$ and the validation set $\mathcal{V}$, it is natural to expect an estimator of $\boldsymbol{\theta}_0$ can be obtained from solving, with respect to $\boldsymbol{\theta}$, the empirical version of (3.2), that is,

$$\mathbf{0} = \alpha N_1^{-1}\textstyle\sum_{i=1}^{N_1}\widehat{\mu}(\mathbf{Z}_i)\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}) + (1-\alpha)N_0^{-1}\textstyle\sum_{i=N_1+1}^{N}\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}) +$$
$$\alpha\, n^{-1}\textstyle\sum_{i=1}^{n}\{D_i-\widehat{\mu}(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}), \tag{3.3}$$

where $\mathbf{W}_i := (S_i, Y_i, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}$ and $\widehat{\mu}(\cdot)$ is an estimator of $\mu^*(\cdot)$ based on the validation set $\mathcal{V}$. Whereas, this is quite a challenging task: rewrite (3.3) as $\sum_{i=1}^{N}\widehat{\xi}_i\,\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}) = \mathbf{0}$ with
$\widehat{\xi}_i := \alpha[\mathbb{1}(i\le n)\{D_i-\widehat{\mu}(\mathbf{Z}_i)\}/n + \mathbb{1}(i\le N_1)\widehat{\mu}(\mathbf{Z}_i)/N_1] + (1-\alpha)\mathbb{1}(i>N_1)$. Since $n < N_1$, it is possible that $\widehat{\xi}_i \equiv \widehat{\mu}(\mathbf{Z}_i)(N_1^{-1}-n^{-1}) < 0$ for $i \in \{i : 1\le i\le n, D_i = 0\}$. If we view (3.3) as a weighted estimating equation, the weight $\widehat{\xi}_i$ can be negative for some summands in $\sum_{i=1}^{N}\widehat{\xi}_i\,\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta})$. Therefore, solving (3.3) may correspond to some nonconvex optimization problem even when $\boldsymbol{\psi}'(\mathbf{W},\boldsymbol{\theta}) := \partial\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is positive definite for any $\boldsymbol{\theta}\in\Theta$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0373

3.1 Construction of debiased estimators using one-step update

To avoid complicated algorithms and simplify the implementation, we adopt the one-step update strategy (Van der Vaart, 2000). Notice the derivative of function $\boldsymbol{\Phi}(\cdot)$ in (3.1) is $\boldsymbol{\Phi}'(\boldsymbol{\theta}) := d\boldsymbol{\Phi}(\boldsymbol{\theta})/d\boldsymbol{\theta} \equiv$

$$\alpha\,\mathbb{E}\{D\boldsymbol{\psi}'(\mathbf{W},\boldsymbol{\theta}) \mid S=1\} + (1-\alpha)\mathbb{E}\{\boldsymbol{\psi}'(\mathbf{W},\boldsymbol{\theta}) \mid S=0\}. \tag{3.4}$$

Let $\boldsymbol{\Omega}(\boldsymbol{\theta}) := -\{\boldsymbol{\Phi}'(\boldsymbol{\theta})\}^{-1}$. Then, at the population level, we can refine an (initial) solution $\boldsymbol{\theta}_{\text{IN}}$ to (3.2) by a one-step update $\boldsymbol{\theta}_{\text{IN}} - \{\boldsymbol{\Phi}'(\boldsymbol{\theta}_{\text{IN}})\}^{-1}\boldsymbol{\Phi}(\boldsymbol{\theta}_{\text{IN}}) \equiv$

$$\boldsymbol{\theta}_{\text{IN}} + \boldsymbol{\Omega}(\boldsymbol{\theta}_{\text{IN}})(\alpha\,\mathbb{E}\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_{\text{IN}}) \mid S=1\} + (1-\alpha)\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_{\text{IN}}) \mid S=0\}$$
$$+\,\alpha\,\mathbb{E}[\{D-\mu^*(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_{\text{IN}}) \mid S=1]\,),$$

whose empirical version provides a family of estimators for $\boldsymbol{\theta}_0$:

$$\widehat{\boldsymbol{\theta}} \;:=\; \widehat{\boldsymbol{\theta}}_{\text{IN}} + \widehat{\boldsymbol{\Omega}}[\alpha N_1^{-1}\textstyle\sum_{i=1}^{N_1}\widehat{\mu}(\mathbf{Z}_i)\boldsymbol{\psi}(\mathbf{W}_i,\widehat{\boldsymbol{\theta}}_{\text{IN}}) + (1-\alpha)N_0^{-1}\sum_{i=N_1+1}^{N}\boldsymbol{\psi}(\mathbf{W}_i,\widehat{\boldsymbol{\theta}}_{\text{IN}})$$
$$+\,\alpha\,n^{-1}\textstyle\sum_{i=1}^{n}\{D_i-\widehat{\mu}(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i,\widehat{\boldsymbol{\theta}}_{\text{IN}})], \tag{3.5}$$

indexed by (initial) estimators $\{\widehat{\boldsymbol{\theta}}_{\text{IN}},\, \widehat{\boldsymbol{\Omega}},\, \widehat{\mu}(\cdot)\}$ for $\{\boldsymbol{\theta}_0,\, \boldsymbol{\Omega} \equiv \boldsymbol{\Omega}(\boldsymbol{\theta}_0),\, \mu^*(\cdot)\}$ that are calculated from the complete observations in $\mathcal{V}\cup\mathcal{C}$. We do not specify their forms in our theoretical analysis. Imposing the high-level conditions in Assumptions 2–3 on $\{\widehat{\boldsymbol{\theta}}_{\text{IN}},\, \widehat{\boldsymbol{\Omega}},\, \widehat{\mu}(\cdot)\}$ is sufficient for establishing the results in Sections 3.2 and 5. Some reasonable choices of $\{\widehat{\boldsymbol{\theta}}_{\text{IN}}, \widehat{\boldsymbol{\Omega}}\}$ will be provided in Section 3.3, while thorough discussions of the phenotyping model estimator $\widehat{\mu}(\cdot)$ can be found in Section 5.

To facilitate derivations of our method's properties, we adopt the *cross-*

fitting technique (Newey and Robins, 2018) for the nuisance estimator $\widehat{\mu}(\cdot)$: without loss of generality, divide the index set $\mathcal{I} := \{1, \ldots, n\}$ into M *disjoint* subsets $\{\mathcal{I}_1, \ldots, \mathcal{I}_M\}$ of size n/M for some *fixed* integer $M \geq 2$. Let $\widehat{\mu}_m(\cdot)$ be an estimator for $\mu^*(\cdot)$ based on the data set $\mathcal{V}_m^- := \{(D_i, S_i \equiv 1, \mathbf{Z}_i) : i \in \mathcal{I} \backslash \mathcal{I}_m\}$ $(m = 1, \ldots, M)$. Then we calculate $\{\widehat{\mu}(\mathbf{Z}_i) : i = 1, \ldots, N_1\}$ in (3.5) as follows:

$$\widehat{\mu}(\mathbf{Z}_i) \equiv \textstyle\sum_{m=1}^M \{\mathbb{1}(i \in \mathcal{I}_m)\widehat{\mu}_m(\mathbf{Z}_i) + \mathbb{1}(i > n)\widehat{\mu}_m(\mathbf{Z}_i)/M\}. \tag{3.6}$$

The cross-fitting procedure removes the dependence between $\widehat{\mu}(\cdot)$ and $\mathbf{Z}_i$ in terms $\{\widehat{\mu}(\mathbf{Z}_i) : i = 1, \ldots, n\}$, making remainders involving them in the expansion of $\widehat{\boldsymbol{\theta}}$ easier to control without changing the influence function thereof. Consequently, when establishing asymptotic properties of $\widehat{\boldsymbol{\theta}}$, we avoid some stringent conditions on $\widehat{\mu}(\cdot)$, which are similar to the stochastic equicontinuity assumptions in empirical processes theory (Van der Vaart, 2000). As shown in Theorem 1 and Remark 3, employing $\widehat{\mu}(\cdot)$ in (3.6) does not reduce asymptotic efficiency of our method though $\widehat{\mu}_m(\cdot)$ involves only a part of the validation set. The use of cross-fitting does not require specific forms of $\widehat{\mu}(\cdot)$, so the flexibility of our method is not degraded.

3.2 Asymptotic properties of $\widehat{\boldsymbol{\theta}}$

We thoroughly study in Theorem 1 asymptotic properties of our estimators $\widehat{\boldsymbol{\theta}}$, presenting their expansion with explicit remainder rates, followed by their limiting distribution. The semiparametric efficiency of $\widehat{\boldsymbol{\theta}}$ is discussed in Remark 3. We first specify assumptions needed for these results.

Assumption 1. Let $\boldsymbol{\psi}'_{[j]}(\mathbf{W},\boldsymbol{\theta})$ and $\boldsymbol{\psi}''_{[j]}(\mathbf{W},\boldsymbol{\theta})$ be the first- and second-order derivatives of $\boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Then there exists $\mathcal{B}_0 := \{\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| < c\}$ with some $c > 0$ such that $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\boldsymbol{\psi}'_{[j]}(\mathbf{W},\boldsymbol{\theta})\|^2\} < \infty$ and $\sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\mathbb{E}\{\boldsymbol{\psi}''_{[j]}(\mathbf{W},\boldsymbol{\theta})\}\| < \infty$ for $j = 1,\ldots,p$.

Assumption 2. For some positive sequences $u_n = o(1)$ and $v_n = o(1)$, the estimators $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ and $\widehat{\boldsymbol{\Omega}}$ satisfy $\|\widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0\| = O_p(u_n)$ and $\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\| = O_p(v_n)$.

Assumption 3. The function $\mu^*(\cdot)$ is bounded. Its estimator $\widehat{\mu}_m(\cdot)$ satisfies for some positive sequences $\{a_{n,\infty}, a_{n,2}\}$ and $m = 1,\ldots,M$ that

$$\sup_{\mathbf{z}\in\mathcal{Z}}|\widehat{\mu}_m(\mathbf{z}) - \mu^*(\mathbf{z})| = O_p(a_{n,\infty}) \quad \text{and} \tag{3.7}$$

$$\mathbb{E}^{1/2}_{\mathbf{Z},\mathbf{W}|S=1}[\|\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\|^2] = O_p(a_{n,2}). \tag{3.8}$$

Assumption 1 on the estimating function $\boldsymbol{\psi}(\cdot,\cdot)$ is usually equivalent to some mild moment conditions on the covariates $\mathbf{X}$. For the primary and secondary analyses specified in Section S2 of the Supplementary Material, Assumption 1 holds as long as $\mathbb{E}(\|\mathbf{X}\|^4) < \infty$. Assumption 2 requires consistency of the (initial) estimators $\{\widehat{\boldsymbol{\theta}}_{\text{IN}}, \widehat{\boldsymbol{\Omega}}\}$ for $\{\boldsymbol{\theta}_0, \boldsymbol{\Omega}\}$, which is standard for the one-step update approach; see Section 3.3 for detailed discussion of $\{\widehat{\boldsymbol{\theta}}_{\text{IN}}, \widehat{\boldsymbol{\Omega}}\}$. In Assumption 3, we do not require $\mu^*(\cdot) = \mu(\cdot)$, allowing for misspecification of the phenotyping model. We consider $\mu^*(\cdot)$ bounded because $\widehat{\mu}(\cdot)$ is usually calculated as an estimated conditional probability, which means its limit $\mu^*(\cdot) \in [0,1]$. Conditions (3.7)–(3.8) specify the rates of terms involving $\widehat{\mu}(\cdot)$ that appear in the remainders of the expansion of $\widehat{\boldsymbol{\theta}}$. We emphasize sequences $\{a_{n,\infty}, a_{n,2}\}$ are allowed to *diverge*. Remark 2 suggests $a_{n,\infty}\log(2+a_{n,\infty}) = o(n^{1/2})$ and $a_{n,2} = o(1)$ are sufficient for $n^{1/2}$-consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}$.

Theorem 1. *Under Assumptions 1–3, our estimators $\widehat{\boldsymbol{\theta}}$ have the following expansion:* $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 =$

$$\boldsymbol{\Omega}[\alpha\, n^{-1}\textstyle\sum_{i=1}^{n}\{D_i - \mu^*(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}_0) + \alpha N_1^{-1}\sum_{i=1}^{N_1}\mu^*(\mathbf{Z}_i)\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}_0) +$$
$$(1-\alpha)N_0^{-1}\textstyle\sum_{i=N_1+1}^{N}\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}_0)] + O_p(r_n) + o_p(n^{-1/2}), \tag{3.9}$$

$$\textit{where } r_n := u_n v_n + u_n^2 + (a_{n,2} + u_n a_{n,\infty})\log(2 + u_n a_{n,\infty}/a_{n,2})/n^{1/2}. \tag{3.10}$$

Further, suppose $r_n = o(n^{-1/2})$, $\mathbb{E}\{\|\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\|^{2(1+c_1)}\} < \infty$ *and* $\lambda_{\min}\{\mathbf{A}_n(\mu^*)\} \geq c_2$ *for some positive constants* $\{c_1, c_2\}$, *where*

$$\mathbf{A}_n(\mu^*) := \alpha^2\text{cov}[\{D - (1-\delta_n)\mu^*(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0) \mid S = 1] +$$
$$\alpha^2\delta_n(1-\delta_n)\text{cov}\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0) \mid S = 1\} +$$
$$(1-\alpha)^2\tau(1-\tau)^{-1}\delta_n\text{cov}\{\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0) \mid S = 0\}. \tag{3.11}$$

$$\textit{Then } n^{1/2}\{\boldsymbol{\Omega}\mathbf{A}_n(\mu^*)\boldsymbol{\Omega}\}^{-1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{N}(\mathbf{0},\mathbf{I}) \textit{ as } n \to \infty. \tag{3.12}$$

Remark 2 (Rate conditions for the asymptotic normality of $\widehat{\boldsymbol{\theta}}$)**.** We can see in Theorem 1 the key to the distributional result (3.12) is the rate condition $r_n = o(n^{-1/2})$ with sequence r_n given in (3.10). The arguments in Section 3.3 indicate typical rates of $\{u_n, v_n\}$ are $u_n = O(n^{-1/2})$ and $v_n = o(1)$. These imply $a_{n,2} + a_{n,\infty}\log\{2 + a_{n,\infty}/(n^{1/2}a_{n,2})\}/n^{1/2} = o(1)$ suffices to guarantee $r_n = o(n^{-1/2})$. Thus, the properties of $\widehat{\mu}(\cdot)$ needed for (3.12) are

$$a_{n,\infty}\log(2 + a_{n,\infty}) = o(n^{1/2}) \quad \text{and} \quad a_{n,2} = o(1) \tag{3.13}$$

with $\{a_{n,\infty}, a_{n,2}\}$ given in (3.7)–(3.8). Here we use the fact that $a_{n,2}$ generally cannot converge faster than $n^{-1/2}$. We can see the L_∞ error $\sup_{\mathbf{z}\in\mathcal{Z}}|\widehat{\mu}_m(\mathbf{z}) - \mu^*(\mathbf{z})|$ is actually allowed to diverge. Rate condition (3.13) is quite reasonable for various regression and machine learning approaches used to calculate $\widehat{\mu}_m(\cdot)$, under suitable conditions of the estimating function $\boldsymbol{\psi}(\cdot,\cdot)$. For example, if the components of $\boldsymbol{\psi}(\cdot,\cdot)$ are bounded, then the second part of (3.13) is equivalent to $\mathbb{E}^{1/2}_{\mathbf{Z}|S=1}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}^2] = o_p(1)$. Compared with assuming $\sup_{\mathbf{z}\in\mathcal{Z}}|\widehat{\mu}_m(\mathbf{z}) - \mu^*(\mathbf{z})| = o_p(1)$ or requiring specific convergence rates of $\widehat{\mu}_m(\cdot)$ (e.g., $a_{n,2} = O(n^{-1/2})$), such an L_2 convergence condition is considerably weaker. Throughout, we place no restriction on the form of $\widehat{\mu}_m(\cdot)$. In the numerical studies of Sections 4 and 6, we will specify some parametric, semiparametric and nonparametric examples of $\widehat{\mu}_m(\cdot)$ for the implementation of our method.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0373

Remark 3 (Semiparametric efficiency of $\widehat{\boldsymbol{\theta}}$)**.** We can show that given the phenotyping model is correctly specified, i.e., $\mu^*(\cdot) = \mu(\cdot)$, our estimators $\widehat{\boldsymbol{\theta}}$ attain under appropriate semiparametric models the *semiparametric efficiency* defined in Tsiatis (2007), when $\delta \equiv \lim_{n\to\infty}(n/N_1)$ is either positive or zero. The detailed derivations are technical and a bit lengthy so we relegate them to Section S3 of the Supplementary Material.

3.3 Choices of $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ and $\widehat{\boldsymbol{\Omega}}$

A natural choice of $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ is the solution to (3.3) with $\widehat{\mu}(\cdot) \equiv 0$, which discards the nonvalidated candidate cases. That is, we obtain $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ from solving

$$\mathbf{0} = \alpha\, n^{-1}\textstyle\sum_{i=1}^{n} D_i\, \boldsymbol{\psi}(\mathbf{W}_i, \widehat{\boldsymbol{\theta}}_{\text{IN}}) + (1-\alpha)N_0^{-1}\textstyle\sum_{i=N_1+1}^{N} \boldsymbol{\psi}(\mathbf{W}_i, \widehat{\boldsymbol{\theta}}_{\text{IN}}), \quad (3.14)$$

which can be written as $\mathbf{0} = \sum_{i=1}^{N} \pi_i \boldsymbol{\psi}(\mathbf{W}_i, \widehat{\boldsymbol{\theta}}_{\text{IN}})$ with $\pi_i := \mathbb{1}(i \le n)\alpha\, D_i/n + \mathbb{1}(i > N_1)(1-\alpha)/N_0 \ge 0$. This can be easily solved as a weighted empirical version of (2.1). The $n^{1/2}$-consistency of $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ in (3.14) is ensured by the classical M-estimation theory (Van der Vaart, 2000, Chapter 5) under standard regularity conditions. Hence, the sequence u_n in Assumption 2 typically satisfies $u_n = O(n^{-1/2})$. To estimate matrix $\boldsymbol{\Omega}$, we set $-\widehat{\boldsymbol{\Omega}}^{-1} \equiv$

$$\alpha\, n^{-1}\textstyle\sum_{i=1}^{n} D_i \boldsymbol{\psi}'(\mathbf{W}_i, \widehat{\boldsymbol{\theta}}_{\text{IN}}) + (1-\alpha)N_0^{-1}\textstyle\sum_{i=N_1+1}^{N} \boldsymbol{\psi}'(\mathbf{W}_i, \widehat{\boldsymbol{\theta}}_{\text{IN}}). \quad (3.15)$$

The convergence of $\widehat{\boldsymbol{\Omega}}$ to $\boldsymbol{\Omega}$ shown below proves $\widehat{\boldsymbol{\Omega}}$ satisfies Assumption 2.

Proposition 1. *Suppose* $\|\widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0\| = o_p(1)$ *and* $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\mathcal{B}_0} \|\boldsymbol{\psi}''_{[j]}(\mathbf{W}, \boldsymbol{\theta})\|\} < \infty$ *for* $j = 1, \ldots, p$. *Then* $\widehat{\boldsymbol{\Omega}}$ *given in* (3.15) *satisfies* $\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\| = o_p(1)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0373

4. Simulations

We now examine the numerical performance of our method on simulated data. Throughout, the dimensions of $\mathbf{Z}$ and $\mathbf{X}$ are $d = 500$ and $p = 12$. The random vector $\mathbf{Z}$ follows the d-dimensional normal distribution with a zero mean and a covariance matrix whose (i, j)th entry equals $0.5^{|i-j|}$. We let $\mathbf{X} \equiv (1, \mathbf{Z}_{[1]}, \ldots, \mathbf{Z}_{[p-1]})^{\mathrm{T}}$ and $Y \equiv \mathbf{Z}_{[p]}$. The first component 1 of $\mathbf{X}$ is included to capture intercept terms in regression models. We use several different combinations of sample sizes (n, N): $n \in \{200, 400\}$ and $N \in \{5 \times 10^3, 10^4, 2.5 \times 10^4\}$. The ratio of candidate cases to controls in the case-control samples is set to $\tau = 0.4$, so the value of $\delta_n \equiv n/N_1 \equiv n/(\tau N)$ ranges from 0.02 to 0.2 in these setups, including cases where the validation set is much smaller or comparable in size to the whole sample. Observations of S and D are generated from the following mechanism:

$$\mathbb{E}(S \mid \mathbf{Z}) \equiv h(2 \textstyle\sum_{j=1}^{d} \mathbf{Z}_{[j]}/d^{1/2}), \;\; \mu(\mathbf{Z}) \equiv \mathbb{E}(D \mid \mathbf{Z}, S = 1) \equiv h\{\rho(\mathbf{Z})\} \quad (4.1)$$

with $h(x) := \{1 + \exp(-x)\}^{-1}$. The prevalence of candidate cases in model (4.1) is $\eta \equiv \mathbb{E}(S) = 0.5$. We consider five different forms of $\mu(\mathbf{Z})$ in (4.1) by setting $\rho(\mathbf{Z})$ to the choices (a)–(e) listed in Section S5 of the Supplementary Material, which simulate a variety of phenotyping effects, including linear, quadratic and interaction ones that are common in biomedical studies. The sparsity level of the high dimensional phenotyping model $\mu(\mathbf{Z})$, which is represented by q, takes two different values, $q = \lceil d^{1/2} \rceil$ and $q = n$, corresponding to sparse and dense models, respectively.

Our target is to estimate the odds ratio parameters (S1) for the primary analysis and the linear regression parameters (S3) (with $f(x) \equiv x$) for the secondary analysis, which are specified in Section S2 of the Supplementary Material. To construct our estimators $\widehat{\boldsymbol{\theta}}$ proposed in (3.5), we plug in the initial estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{IN}}$ that solves (3.14), along with the Jacobian estimate $\widehat{\boldsymbol{\Omega}}$ in (3.15). Regarding the phenotyping model estimator $\widehat{\mu}(\cdot)$, we adopt the cross-fitting strategy (3.6) with $M = 10$ and four choices of $\widehat{\mu}_m : \mathbb{R}^d \mapsto \mathbb{R}$:

(i) a *parametric* estimator

$$\widehat{\mu}_m(\mathbf{z}) \equiv \{1 + \exp(-\widehat{\boldsymbol{\beta}}_m^{\mathrm{T}} \mathbf{z})\}^{-1} \tag{4.2}$$

with $\widehat{\boldsymbol{\beta}}_m \in \mathbb{R}^d$ from penalized logistic regression of D on $\mathbf{Z}$ using $\mathcal{V}_m^-$, where the L_1 penalty is used with a tuning parameter chosen by cross validation;

(ii) a *semiparametric* estimator

$$\widehat{\mu}_m(\mathbf{z}) \equiv \textstyle\sum_{i \in \mathcal{I}_m^-} D_i K_b\{\widehat{\boldsymbol{\beta}}_m^{\mathrm{T}}(\mathbf{z} - \mathbf{Z}_i)\} / \sum_{i \in \mathcal{I}_m^-} K_b\{\widehat{\boldsymbol{\beta}}_m^{\mathrm{T}}(\mathbf{z} - \mathbf{Z}_i)\},$$

where $\widehat{\boldsymbol{\beta}}_m$ is as in (4.2) and $K_b(x) := \exp\{-x^2/(2b^2)\}$ denotes the Gaussian kernel with a bandwidth b chosen by cross validation;

(iii) a *machine learning* estimator $\widehat{\mu}_m(\mathbf{z})$ yielded by applying the random forest algorithm to $\mathcal{V}_m^-$, which treats D as the outcome, grows 500 trees and randomly draws $\lceil d^{1/2} \rceil$ components of $\mathbf{Z}$ as candidates at each split.

The implementations of the above (i), (ii) and (iii) make use of the `R` packages `glmnet`, `np` and `randomForest`. In the following, all the results are summarized over 500 iterations. To save space, the results of the secondary analysis are reported in the Supplementary Material

4.1 Estimation results: relative efficiencies

We investigate the estimation quality. The consistent initial estimator $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ in (3.14) serves as a benchmark. Tables 1 and S1 (in the Supplementary Material) report efficiencies of $\widehat{\boldsymbol{\theta}}$ relative to $\widehat{\boldsymbol{\theta}}_{\text{IN}}$, i.e.,

$$\mathbb{E}(\|\widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0\|^2)/\mathbb{E}(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2). \tag{4.3}$$

Since $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ uses only controls and cases in the validation set, ignoring the nonvalidated candidate case pool $\mathcal{N}$, criterion (4.3) reflects if and to what extent our method improves the estimation of $\boldsymbol{\theta}_0$ through exploiting $\mathcal{N}$. Also, we provide the maximum relative efficiency in each setting:

$$\text{MRE} := \text{tr}\{\boldsymbol{\Omega}\mathbf{A}_n(0)\boldsymbol{\Omega}\}/\text{tr}\{\boldsymbol{\Omega}\mathbf{A}_n(\mu)\boldsymbol{\Omega}\}, \tag{4.4}$$

where $\text{tr}(\cdot)$ denotes the trace of a square matrix; see (3.4) and (3.11) for the forms of matrices $\boldsymbol{\Omega} \equiv -\{\boldsymbol{\Phi}'(\boldsymbol{\theta}_0)\}^{-1}$ and $\mathbf{A}_n(\cdot)$. This represents the optimal asymptotic efficiency that is attainable only if the phenotyping model $\mu(\cdot)$ is correctly specified and the estimation errors of $\{\widehat{\mu}_m(\cdot), \widehat{\boldsymbol{\theta}}_{\text{IN}}, \widehat{\boldsymbol{\Omega}}\}$ are zero. Apparently, the maximum relative efficiency is *unrealistic* in a finite sample.

We present it just as a reference for estimation quality. The true value of $\boldsymbol{\theta}_0$, as well as the unknown quantities in (4.4), is calculated by Monte Carlo based on a sample $\{(D_i, S_i, \mathbf{Z}_i) : i = 1, \ldots, 5 \times 10^4\}$ that is independent of the data involved in estimation.

Across all the simulation configurations listed in Tables 1 and S1, regardless of whether the phenotyping model $\mu(\cdot)$ is misspecified by the nuisance estimator $\widehat{\mu}_m(\cdot)$ or not, we observe uniform superiority of our estimators $\widehat{\boldsymbol{\theta}}$ over the benchmark $\widehat{\boldsymbol{\theta}}_{\text{IN}}$, indicated by the relative efficiencies that significantly exceed one. The advantages become more notable as the whole sample size N increases. These results corroborate our method achieves robust and efficient use of the nonvalidated candidate cases, which are discarded by $\widehat{\boldsymbol{\theta}}_{\text{IN}}$. When the validation set size n rises from 200 to 400, there emerges a clear trend of numbers in the two tables approaching the corresponding maximum relative efficiencies. This validates the asymptotic optimality of our method under correct specification of $\mu(\cdot)$ as clarified in Remark 3. Comparing the three different types of nuisance estimators, we can see for sparse models with $q = \lceil d^{1/2} \rceil$, the parametric one using penalized logistic regression generally produces the best results. When it comes to denser models with $q = n$, the machine learning method based on random forest shows advantages in most of the cases. This observation can be

Table 1: Simulation results of the primary analysis in Section 4: relative efficiencies (4.3) of our estimators $\widehat{\boldsymbol{\theta}}$ to the benchmark $\widehat{\boldsymbol{\theta}}_{\text{IN}}$. The nuisance estimator $\widehat{\mu}(\cdot)$ in $\widehat{\boldsymbol{\theta}}$ is constructed using logistic regression (LR), kernel smoothing (KS) or random forest (RF). Here q is the sparsity level of the phenotyping model $\mathbb{E}(D \mid \mathbf{Z}, S = 1)$, $d \equiv 500$ the dimension of the predictors $\mathbf{Z}$, N the whole sample size, n the validation set size, $\rho(\mathbf{Z})$ the function in data generating model (4.1) and MRE as defined in (4.4). The choices (a)–(e) of $\rho(\mathbf{Z})$ are listed in Section S5 of the Supplementary Material. Results in settings with (a) $\rho(\mathbf{Z}) \equiv 0.7$ are displayed in the upper panel only because they are not affected by the sparsity level q.

$q = \lceil d^{1/2} \rceil$		$N = 5000$				$N = 10000$				$N = 25000$			
n	$\rho(\mathbf{Z})$	LR	KS	RF	MRE	LR	KS	RF	MRE	LR	KS	RF	MRE
200	(a)	2.31	2.20	2.33	2.44	2.74	2.56	2.76	2.76	3.00	2.74	3.01	3.00
	(b)	2.68	2.22	1.75	4.18	2.92	2.40	1.79	5.59	3.37	2.57	1.95	7.09
	(c)	2.17	1.91	2.21	3.29	2.34	2.03	2.42	4.03	2.77	2.36	2.84	4.69
	(d)	1.94	1.72	1.93	3.25	2.10	1.82	2.09	3.94	2.29	1.95	2.27	4.53
	(e)	2.67	2.35	2.02	3.76	2.88	2.37	2.10	4.85	3.37	2.62	2.44	5.92
400	(a)	1.94	1.91	1.94	2.01	2.40	2.35	2.41	2.44	2.77	2.68	2.77	2.83
	(b)	2.45	2.28	1.63	2.84	3.15	2.79	1.80	4.18	3.88	3.20	1.95	6.01
	(c)	1.92	1.85	1.92	2.45	2.19	2.07	2.19	3.29	2.61	2.42	2.60	4.23
	(d)	1.75	1.69	1.69	2.45	1.99	1.87	1.94	3.25	2.20	2.05	2.14	4.11
	(e)	2.37	2.27	1.81	2.65	2.88	2.67	2.02	3.76	3.73	3.33	2.34	5.16
$q = n$		$N = 5000$				$N = 10000$				$N = 25000$			
n	$\rho(\mathbf{Z})$	LR	KS	RF	MRE	LR	KS	RF	MRE	LR	KS	RF	MRE
200	(b)	2.42	2.02	2.40	4.93	2.72	2.17	2.71	7.18	2.97	2.29	2.98	10.03
	(c)	2.63	2.33	2.69	3.91	2.91	2.56	2.99	5.14	3.48	2.94	3.60	6.39
	(d)	2.25	1.96	2.31	3.80	2.45	2.12	2.54	4.89	2.76	2.27	2.92	5.94
	(e)	2.75	2.42	2.81	4.54	3.31	2.74	3.38	6.43	4.00	3.25	4.08	8.68
400	(b)	2.25	2.06	2.23	3.11	2.94	2.58	2.96	5.05	3.69	3.07	3.68	8.49
	(c)	2.27	2.15	2.29	2.85	3.07	2.79	3.11	4.32	4.08	3.53	4.13	6.50
	(d)	2.07	1.95	2.09	2.75	2.52	2.26	2.56	4.01	3.02	2.63	3.08	5.68
	(e)	2.47	2.36	2.48	3.01	3.55	3.25	3.58	4.84	4.87	4.36	4.91	8.01

attributed to the relatively strong ability of machine learning approaches in capturing nonlinear and dense structures of complex phenotyping models.

It highlights the benefits of allowing the form of $\widehat{\mu}_m(\cdot)$ to be arbitrary, which is a remarkable feature of our method. Concerning the semiparametric nuisance estimator that combines logistic regression and kernel smoothing, we notice substantial improvement of its performance when n increases from 200 to 400. We believe this is because the bandwidth selection in the kernel smoothing procedure is more stable with a larger sample size. All these three approaches largely facilitate estimation for the target parameters $\boldsymbol{\theta}_0$ based on the contaminated case-control samples through successfully learning the relation between the true status D and predictors $\mathbf{Z}$ among candidate cases. In addition we consider in some of the simulation settings a plug-in estimator $\widehat{\boldsymbol{\theta}}_{\text{PI}}$, which is the solution to

$$\alpha\, N_1^{-1}\textstyle\sum_{i=1}^{N_1}\widehat{\mu}(\mathbf{Z}_i)\boldsymbol{\psi}(\mathbf{W}_i,\widehat{\boldsymbol{\theta}}_{\text{PI}})+(1-\alpha)N_0^{-1}\sum_{i=N_1+1}^{N}\boldsymbol{\psi}(\mathbf{W}_i,\widehat{\boldsymbol{\theta}}_{\text{PI}})=\mathbf{0}. \quad (4.5)$$

The estimating equation (4.5) naively substitutes the phenotyping model estimator $\widehat{\mu}(\mathbf{Z})$ for the possibly unknown status variable D without debiasing. Its efficiency relative to $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ is recorded in Table S3 of the Supplementary Material to save space. Comparing the numbers in Tables 1, S1 and S3, we can see apparent efficiency inferiority of $\widehat{\boldsymbol{\theta}}_{\text{PI}}$ to our estimators $\widehat{\boldsymbol{\theta}}$. In fact $\widehat{\boldsymbol{\theta}}_{\text{PI}}$ performs even worse than the initial estimator $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ in some cases. These results emphasize the necessity of adopting appropriate debiasing procedures after imputation, especially when $\widehat{\mu}(\cdot)$ is calculated based

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0373

on high dimensional or machine learning models.

4.2 Inference results: confidence intervals

We now establish 95% confidence intervals for the components of $\boldsymbol{\theta}_0$ using the samples of size $N = 5 \times 10^3$, based on limiting distribution (3.12) and a consistent estimate $\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{A}}_n\widehat{\boldsymbol{\Omega}}/n$ for the covariance matrix $\boldsymbol{\Omega}\mathbf{A}_n(\mu^*)\boldsymbol{\Omega}/n$ therein. Here $\widehat{\boldsymbol{\Omega}}$ is as in (3.15) and $\widehat{\mathbf{A}}_n :=$

$$\alpha^2\widehat{\text{cov}}_{\mathcal{V}}[\{D-(1-\delta_n)\widehat{\mu}(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}})] + \alpha^2\delta_n(1-\delta_n)\widehat{\text{cov}}_{\mathcal{V}\cup\mathcal{N}}\{\widehat{\mu}(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}})\}$$

$$+ (1-\alpha)^2\tau(1-\tau)^{-1}\delta_n\widehat{\text{cov}}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}})\}$$

with $\widehat{\text{cov}}_{\mathcal{V}}(\cdot)$, $\widehat{\text{cov}}_{\mathcal{V}\cup\mathcal{N}}(\cdot)$ and $\widehat{\text{cov}}_{\mathcal{C}}(\cdot)$ referring to the empirical covariance matrices calculated from the data sets $\mathcal{V}$, $\mathcal{V}\cup\mathcal{N}$ and $\mathcal{C}$, respectively. Results in the settings where $N \in \{10^4, 2.5\times 10^4\}$ appear similar in pattern. We thus omit them in the interest of space. Since the target $\boldsymbol{\theta}_0$ is a p-dimensional vector, we calculate for each case two numbers that summarize the coverage rates and lengths of the p componentwise confidence intervals $\{\text{CI}_1, \ldots, \text{CI}_p\}$: (a) the deviation of coverage rates from the nominal level,

$$\text{DCR} := p^{-1}\textstyle\sum_{j=1}^{p}|\text{CR}_j \times 100 - 95| \tag{4.6}$$

with CR_j the coverage rate of CI_j, and (b) the average length of $\{\text{CI}_1, \ldots, \text{CI}_p\}$.

In Tables 2 and S2 (in the Supplementary Material), we can see the confidence intervals possess satisfactory average lengths along with negligible deviations of coverage rates across all the settings, which reveal high precision and accuracy of our inference method. These observations verify the distributional result obtained in Theorem 1, confirming the asymptotic normality of $\widehat{\boldsymbol{\theta}}$ relies on neither specific forms of the nuisance estimator nor correct specification of the phenotyping model. Interestingly, when $q = n$, our method still yields coverage rates that are very close to the nominal level 95%. Considering the typical L_2 convergence rate of $\widehat{\boldsymbol{\beta}}_m$ in (4.2), for example, is $O_p((q \log d/n)^{1/2})$ (Wainwright, 2019), limiting behaviors of $\widehat{\mu}_m(\cdot)$ in (4.2) are quite hard to specify in dense models with $q = n$: even the existence of its limit is uncertain. The decent performance in such harsh settings suggests notable insensitivity of our inference method to asymptotics of the nuisance estimation, which will be further investigated in Section 5. Overall, the simulation results demonstrate based on contaminated case-control data, our method is able to provide robust and efficient inference for $\boldsymbol{\theta}_0$ without knowledge concerning (a) structures of the phenotyping model and (b) convergence properties of the nuisance estimator.

5. Relaxation of conditions on the nuisance estimator

Noticing in Section 4 that our method yields satisfactory numerical performance under high dimensional dense phenotyping models, where limiting

Table 2: Simulation results of the primary analysis in Section 4: componentwise confidence intervals for $\boldsymbol{\theta}_0$ established based on our estimators $\widehat{\boldsymbol{\theta}}$ given in (3.5). The sample size is $N = 5000$. The nuisance estimator $\widehat{\mu}(\cdot)$ in $\widehat{\boldsymbol{\theta}}$ is constructed using logistic regression (LR), kernel smoothing (KS) or random forest (RF). Here q is the sparsity level of the phenotyping model $\mathbb{E}(D \mid \mathbf{Z}, S = 1)$, $d \equiv 500$ the dimension of the predictors $\mathbf{Z}$, n the validation set size, $\rho(\mathbf{Z})$ the function in data generating model (4.1), DCR as defined in (4.6) and AL stands for "average length". The choices (a)–(e) of $\rho(\mathbf{Z})$ are listed in Section S5 of the Supplementary Material.

		$q = \lceil d^{1/2} \rceil$						$q = n$					
		LR		KS		RF		LR		KS		RF	
n	$\rho(\mathbf{Z})$	DCR	AL	DCR	AL	DCR	AL	DCR	AL	DCR	AL	DCR	AL
200	(a)	0.90	0.28	0.97	0.28	0.98	0.28	0.90	0.28	0.97	0.28	0.98	0.28
	(b)	0.73	0.31	0.93	0.33	0.87	0.38	0.85	0.28	0.80	0.30	0.93	0.28
	(c)	0.93	0.29	0.65	0.31	0.90	0.29	0.95	0.26	0.88	0.27	1.15	0.25
	(d)	1.12	0.34	0.85	0.36	1.08	0.34	0.78	0.29	1.02	0.31	0.78	0.29
	(e)	0.85	0.28	0.90	0.30	0.98	0.32	0.93	0.24	0.88	0.26	1.03	0.24
400	(a)	0.48	0.21	0.40	0.22	0.52	0.21	0.48	0.21	0.40	0.22	0.52	0.21
	(b)	0.70	0.23	0.93	0.24	0.67	0.28	1.35	0.19	1.15	0.20	1.18	0.19
	(c)	0.62	0.22	0.75	0.23	0.65	0.23	0.80	0.18	0.78	0.19	0.70	0.18
	(d)	0.83	0.25	1.07	0.26	1.05	0.25	0.65	0.20	0.70	0.21	0.62	0.20
	(e)	0.85	0.21	0.67	0.22	0.47	0.24	1.05	0.17	1.10	0.17	1.10	0.17

behaviors of $\widehat{\mu}_m(\cdot)$ can be quite uncertain and Assumption 3 may not hold, we attempt to provide some explanations for this observation by exploring properties of our estimators $\widehat{\boldsymbol{\theta}}$ without Assumption 3. Another motivation of this section is the fact that $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ in (3.14) is a regular M-estimator without any nuisance function, which enjoys $n^{1/2}$-consistency and asymptotic normality for $\boldsymbol{\theta}_0$ under standard regularity conditions. However, the simulation results in Section 4.1 indicate $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ is generally inefficient due to ignoring the nonvalidated candidate cases. Compared with $\widehat{\boldsymbol{\theta}}_{\text{IN}}$, our method involves nuisance estimator $\widehat{\mu}_m(\cdot)$ to achieve semiparametric efficiency. We naturally wish to minimize costs of the efficiency gain, i.e., to strengthen robustness of $\widehat{\boldsymbol{\theta}}$ to assumptions on $\widehat{\mu}_m(\cdot)$ to the greatest extent possible.

Consider this question: what if no convergence condition is imposed on the nuisance estimation? Theorem 2 shows in this scenario, our estimators $\widehat{\boldsymbol{\theta}}$ are still $n^{1/2}$-consistent. To attain asymptotic normality, we introduce a variant of $\widehat{\boldsymbol{\theta}}$ adopting the *sample-splitting* strategy (Cox, 1975): divide the validation set $\mathcal{V}$ into two disjoint subsets, $\widetilde{\mathcal{V}}_1 := \{(D_i, S_i \equiv 1, \mathbf{Z}_i) : 1 \leq i \leq n_1\}$ and $\widetilde{\mathcal{V}}_2 := \{(D_i, S_i \equiv 1, \mathbf{Z}_i) : n_1 < i \leq n\}$, of sizes n_1 and $n_2 := n - n_1$. Let $\widetilde{\mu}_1(\cdot)$ be a random function *involving* $\widetilde{\mathcal{V}}_1$ *only*. The sample-splitting variant of our estimators for $\boldsymbol{\theta}_0$ is $\widetilde{\boldsymbol{\theta}} :=$

$$\widehat{\boldsymbol{\theta}}_{\text{IN}} + \widehat{\boldsymbol{\Omega}}[\alpha(N_1 - n_1)^{-1}\textstyle\sum_{i=n_1+1}^{N_1}\widetilde{\mu}_1(\mathbf{Z}_i)\boldsymbol{\psi}(\mathbf{W}_i, \widehat{\boldsymbol{\theta}}_{\text{IN}}) + (1-\alpha)N_0^{-1}\times$$
$$\textstyle\sum_{i=N_1+1}^{N}\boldsymbol{\psi}(\mathbf{W}_i, \widehat{\boldsymbol{\theta}}_{\text{IN}}) + \alpha\, n_2^{-1}\sum_{i=n_1+1}^{n}\{D_i - \widetilde{\mu}_1(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i, \widehat{\boldsymbol{\theta}}_{\text{IN}})]. \tag{5.1}$$

We can see $\sum_{i=n_1+1}^{N_1}\widetilde{\mu}_1(\mathbf{Z}_i)\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta})$, $\sum_{i=N_1+1}^{N}\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta})$ and $\sum_{i=n_1+1}^{n}\{D_i - \widetilde{\mu}_1(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta})$ in (5.1) are three sums of *conditionally independent* terms given $\widetilde{\mathcal{V}}_1$, since $\widetilde{\mu}_1(\cdot)$ is calculated from $\widetilde{\mathcal{V}}_1 \equiv \{(D_i, S_i \equiv 1, \mathbf{Z}_i) : 1 \leq i \leq n_1\}$ only. This is a key feature that enables us to derive the conditional limiting distribution of $\widetilde{\boldsymbol{\theta}}$ given $\widetilde{\mathcal{V}}_1$ by treating $\widetilde{\mu}_1(\cdot)$ as *nonrandom*. Then, the unconditional asymptotic normality of $\widetilde{\boldsymbol{\theta}}$ follows.

Theorem 2. *Let $\widehat{\mu}(\cdot)$ in* (3.5) *be the cross-fitting estimator* (3.6)*. Suppose* $\{\widetilde{\mu}_1(\cdot), \widehat{\mu}_1(\cdot), \cdots, \widehat{\mu}_M(\cdot)\}$ *are bounded and Assumptions 1–2 hold. Then estimators $\widehat{\boldsymbol{\theta}}$ in* (3.5) *satisfies* $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(n^{-1/2} + u_n v_n + u_n^2)$*. Concerning $\widetilde{\boldsymbol{\theta}}$ in* (5.1)*, we have* $n_2^{1/2}(\boldsymbol{\Omega}\widetilde{\mathbf{A}}_n\boldsymbol{\Omega})^{-1/2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{I})$ *as $n_2 \to \infty$ given* $u_n v_n + u_n^2 = o(n_2^{-1/2})$, $\mathbb{E}\{\|\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\|^{2(1+c)}\} < \infty$ *for some constant $c > 0$, and* $\lambda_{\min}^{-1}(\widetilde{\mathbf{A}}_n) = O_p(1)$*. The form of matrix $\widetilde{\mathbf{A}}_n$ is specified in Section S6 of the Supplementary Material.*

Remark 4 (Interpretation and extension of results in Theorem 2)**.** The only requirement on the nuisance estimators is boundedness. This is realistic because $\{\widetilde{\mu}_1(\cdot), \widehat{\mu}_m(\cdot)\}$ typically aim to predict some conditional probabilities. More generally, assuming certain moments of the nuisance estimators to be of order $O_p(1)$ should suffice for the results in Theorem 2. Compared with the standard definition of "robustness" in the presence of nuisance estimation in the literature such as Bang and Robins (2005), Athey and Imbens (2015) and Dai et al. (2023), the robustness of $\widehat{\boldsymbol{\theta}}$ established in Theorem 2 is superior in the sense that the $n^{1/2}$-consistency of $\widehat{\boldsymbol{\theta}}$ does not rely on any asymptotic property of the nuisance estimator, while the standard robustness just defends against misspecification of nuisance estimators' limits. Regarding $\widetilde{\boldsymbol{\theta}}$, one may not use it to estimate $\boldsymbol{\theta}_0$ in practice due to efficiency loss caused by discarding n_1 validated cases in formula (5.1). However, by proposing $\widetilde{\boldsymbol{\theta}}$ and investigating its asymptotic behaviors, we have illustrated the possibility of valid inference without any asymptotic knowledge of nuisance estimators. We believe the distributional result of $\widetilde{\boldsymbol{\theta}}$ in Theorem 2 can partially explain the numerical results in Section 4.2: the confidence intervals produced by our method always yield satisfactory coverage rates for the components of $\boldsymbol{\theta}_0$, even if properties of $\widehat{\mu}(\cdot)$ are intractable. In Section S4 of the Supplementary Material, we illustrate generalizability of Theorem 2 by deriving analogous properties, which are entirely free of convergence assumptions on nuisance estimation, for approaches to two other important problems involving nuisance functions: semi-supervised inference for mean response (Zhang and Bradic, 2022) and average treatment effect estimation in randomized experiments (Wager et al., 2016).

6. Real data analysis

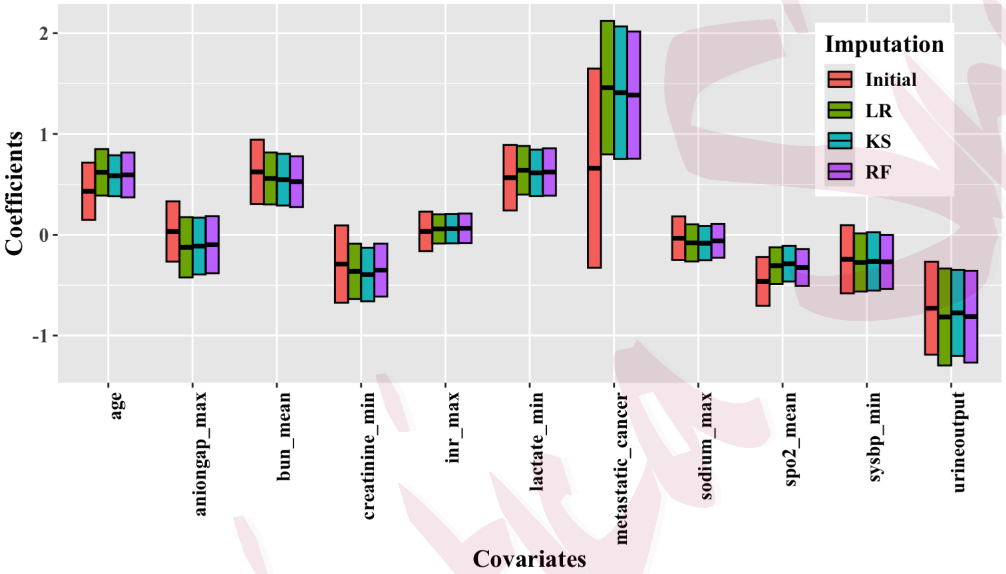
This section illustrates the applicability of our method through an association study of sepsis-related death with a set of critical risk factors, which was based on a data set extracted from MIMIC III (Johnson et al., 2016). For each patient, the data set provided records of the thirty-day vital status S and covariates $\mathbf{X} \in \mathbb{R}^{12}$ (including a constant one as the first component) that were identified as important risk factors for patients' mortality by Hou et al. (2020). Detailed descriptions of these covariates can be found in Table S4 of the Supplementary Material. Noticing some individuals do *not* have any sepsis-related billing code, we realized they were very likely to be sepsis-free and should thereby be excluded from the study that focused on sepsis-related death. To handle this potential cohort contamination, we treated only those with *at least two of the six common sepsis-related billing codes* as eligible observations. Complete records of the six codes were available for 3411 survivals and 116 deaths in the data set. According to our criterion, 1327 survivals and 80 deaths were classified as sepsis patients. The former suffices to serve as the control pool in our study while the latter provides too few cases. We attempted to make use of another 579 deaths who were recorded in the data set but had unknown sepsis statuses. The two-sided t/Z-test between the groups of 116 validated and 579 nonvalidated deaths was conducted for each of the continuous/binary covariates in $\mathbf{X}$, yielding p-values all above 0.05. Hence we treated the validation set as a random subset of all the deaths. Considering the inclusion of ineligibles and the incompleteness of phenotyping information, it was suitable to apply our method to conduct the primary analysis on relationship between mortality and the 12 covariates among sepsis patients. In addition to $\mathbf{X}$, the data set contained observations for another 68 predictors as well, including

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0373

a number of demographic and clinical variables that were informative for sepsis status. In summary, we had $N_0 = 1327$ controls ($S = 0$ and $D = 2$) as well as $N_1 = 116 + 579 = 695$ candidate cases ($S = 1$ and $D \in \{0, 1\}$) of which only $n = 116$ possessed known status D. We aimed at the odds ratio parameters of the vital status S on the covariates $\mathbf{X} \in \mathbb{R}^{12}$ among individuals with $D \neq 0$, as defined in (S1) of the Supplementary Material. The phenotyping model $\mathbb{E}(D \mid \mathbf{Z}, S = 1)$ was established based on the predictors in $\mathbf{Z} \in \mathbb{R}^{80}$, including $\mathbf{X}$ as a subvector.

In Figure 1, we display 95% confidence intervals of the 11 odds ratio parameters constructed based on our estimators $\widehat{\boldsymbol{\theta}}$ in (3.5). The intercept term was excluded because it is usually not of interest in a case-control study. Our method was implemented as described in the second paragraph of Section 4. The benchmark was still the initial estimator $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ given in (3.14). Compared with $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ (red bars), our estimators $\widehat{\boldsymbol{\theta}}$ with any of the three imputation functions (green, blue and purple bars) produced substantially shorter confidence intervals for all the parameters. In particular, the covariate “metastatic_cancer” was detected as significant by our approach. This finding coincided with the medical knowledge that metastatic cancer increases the risk of death. In contrast, the benchmark inference method that discarded the nonvalidated candidate cases missed this significance.

Figure 1: Results of the data analysis in Section 6, which studied relationship between mortality and 11 covariates among sepsis patients: 95% confidence intervals of the odds ratio parameters defined in (S1) of the Supplementary Material, which were constructed based on the initial estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{IN}}$ in (3.14) or our estimators $\widehat{\boldsymbol{\theta}}$ in (3.5). The imputation function $\widehat{\mu}(\cdot)$ in $\widehat{\boldsymbol{\theta}}$ was calculated using logistic regression (LR), kernel smoothing (KS) or random forest (RF). All the continuous covariates were standardized in advance.



Among the three imputation functions, the semiparametric (blue bars) and machine learning (purple bars) ones, which leveraged kernel smoothing and random forest, outperformed the parametric one (green bars) based on logistic regression. This finding again illustrated benefits of compatibility with any reasonable regression or machine learning approaches employed for the nuisance estimation, which is a desirable feature of our method. Also we calculated the plug-in estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{PI}}$ in (4.5), whose components were displayed in Table S5 of the Supplementary Material in the interest of space. Therein $\widehat{\boldsymbol{\theta}}_{\text{PI}}$ showed severe deviation from our estimators presented in Figure 1, which can be attributed to naive imputation without debiasing. Due to not being appropriately debiased in the construction of $\widehat{\boldsymbol{\theta}}_{\text{PI}}$, the nuisance estimator $\widehat{\mu}(\cdot)$ yielded first-order errors that possibly dominated the limiting behavior of $\widehat{\boldsymbol{\theta}}_{\text{PI}}$. Considering $\widehat{\mu}(\cdot)$ was based on high dimensional or machine learning models with intractable properties, the plug-in estimator $\widehat{\boldsymbol{\theta}}_{\text{PI}}$ generally did not have an explicit asymptotic distribution, thus not allowing for confidence interval construction. This fact highlighted another critical advantage of our approach – the ability to provide valid inference.

Supplementary Material

The Supplementary Material collects several important materials that cannot be accommodated in the main article. All the programs and data set used in Sections 4 and 6 are available at `https://github.com/guorongdai/case_contamination_M_estimation`.

Acknowledgments

Dr. Jinbo Chen was supported by grants R01-HL138306 and R01-CA236468.
Dr. Guorong Dai was supported by the NSF-China grant 72271060.

References

Athey, S. and G. Imbens (2015). A measure of robustness to misspecification. *American Economic Review 105*(5), 476–80.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–973.

Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association 91*(433), 14–28.

Casey, J. A., B. S. Schwartz, W. F. Stewart, and N. E. Adler (2016). Using electronic health records for population health research: A review of methods and applications. *Annual Review of Public Health 37*, 61–81.

Chen, Y.-H. (2000). A robust imputation method for surrogate outcome data. *Biometrika 87*(3), 711–716.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika 62*(2), 441–444.

Dai, G., Y. Ma, J. Hasler, J. Chen, and R. J. Carroll (2023). A robust approach for electronic health record–based case-control studies with contaminated case pools. *Biometrics 79*(3), 2023–2035.

Hou, N., M. Li, L. He, B. Xie, L. Wang, R. Zhang, et al. (2020). Predicting 30-days mortality for MIMIC-III patients with sepsis-3: A machine learning approach using XGboost. *Journal* *of Translational Medicine 18*(1), 1–14.

Johnson, A. E., T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data 3*(1), 1–9.

Klarin, D., J. Lynch, K. Aragam, M. Chaffin, T. L. Assimes, J. Huang, et al. (2019). Genome-wide association study of peripheral artery disease in the million veteran program. *Nature Medicine 25*(8), 1274–1279.

Newey, W. K. and J. R. Robins (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Palen, T. E., D. Price, S. Shetterly, and K. B. Wallace (2012). Comparing virtual consults to traditional consults using an electronic health record: An observational case–control study. *BMC Medical Informatics and Decision Making 12*(1), 1–10.

Pathak, J., A. N. Kho, and J. C. Denny (2013). Electronic health records-driven phenotyping: Challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association 20*(e2), e206–e211.

Prentice, R. L. and R. Pyke (1979). Logistic disease incidence models and case-control studies. *Biometrika 66*(3), 403–411.

Tchetgen Tchetgen, E. J. (2014). A general regression framework for a secondary outcome in case–control studies. *Biostatistics 15*(1), 117–128.

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

Media.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, Volume 3. Cambridge University Press.

Wager, S., W. Du, J. Taylor, and R. J. Tibshirani (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences 113*(45), 12673–12678.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Volume 48. Cambridge University Press.

Wang, L., J. Schnall, A. Small, R. A. Hubbard, J. H. Moore, S. M. Damrauer, et al. (2021). Case contamination in electronic health records based case-control studies. *Biometrics 77*(1), 67–77.

Zhang, Y. and J. Bradic (2022). High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika 109*(2), 387–403.

Guorong Dai (corresponding author), Department of Statistics and Data Science,
School of Management, Fudan University, Shanghai 200433, China;
E-mail: guorongdai@fudan.edu.cn

Jinbo Chen, Department of Biostatistics, Epidemiology and Informatics,
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA;
E-mail: jinboche@pennmedicine.upenn.edu