8-2023

A Nested Semiparametric Method for Case-control study with missingness

Ge Zhao
Portland State University, gzhao@pdx.edu

Yanyuan Ma
Penn State University

Jill Schnall Hasler
University of Pennsylvania Perelman School of Medicine, Philadelphia

Scott Damrauer
University of Pennsylvania Perelman School of Medicine, Philadelphia

Michael Levin
University of Pennsylvania Perelman School of Medicine, Philadelphia

See next page for additional authors

Citation Details

Zhao, G., Ma, Y., Hasler, J. S., Damrauer, S., Levin, M., & Chen, J. A Nested Semiparametric Method for Case-Control Study with Missingness. Scandinavian Journal of Statistics.

Authors

Ge Zhao, Yanyuan Ma, Jill Schnall Hasler, Scott Damrauer, Michael Levin, and Jinbo Chen

Received: 15 September 2021 | Revised: 30 January 2023 | Accepted: 19 June 2023

DOI: 10.1111/sjos.12673

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

A nested semiparametric method for case-control study with missingness

Ge Zhao[1] | Yanyuan Ma[2] | Jill Schnall Hasler[3] | Scott Damrauer[3] | Michael Levin[3] | Jinbo Chen[3]

[1]Department of Mathematics and Statistics, Portland State University, Portland, Oregon, USA

[2]Department of Statistics, Penn State University, University Park, Pennsylvania, USA

[3]Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

Correspondence
Ge Zhao, Department of Mathematics and Statistics, Portland State University, Portland, OR 97201, USA.
Email: gzhao@pdx.edu

Abstract

We propose a nested semiparametric model to analyze a case-control study where genuine case status is missing for some individuals. The concept of a noncase is introduced to allow for the imputation of the missing genuine cases. The odds ratio parameter of the genuine cases compared to controls is of interest. The imputation procedure predicts the probability of being a genuine case compared to a noncase semiparametrically in a dimension reduction fashion. This procedure is flexible, and vastly generalizes the existing methods. We establish the root-n asymptotic normality of the odds ratio parameter estimator. Our method yields stable odds ratio parameter estimation owing to the application of an efficient semiparametric sufficient dimension reduction estimator. We conduct finite sample numerical simulations to illustrate the performance of our approach, and apply it to a dilated cardiomyopathy study.

KEYWORDS

case-control study, missingness, semiparametrics

1 INTRODUCTION

Our work is motivated by a case-control study of dilated cardiomyopathy conducted using the University of Pennsylvania hospital electronic health record (EHR). Cases and controls were identified from Penn EHRs using separate rules that were created based on EHR data elements. The rule for identifying controls was rigorous so that controls were identified accurately as in typical EHR-based case-control studies. A more relaxed rule was used for identifying candidate cases. A larger number of genuine cases could be included in the study using this more relaxed rule, which is essential for ensuring study power and generalizability of study results. However, such a relaxed rule led to the inclusion of patients who are not genuine cases and also do not satisfy the control definition. These patients are referred to as "non-cases" (Wang et al., 2020). Noncases differ from genuine cases, and they differ from controls as well, making them ineligible for the study. When estimating odds ratio association parameters, naively treating noncases as genuine cases will lead to biased results (Little & Rubin, 2019). Stemming from the fact that it is often very difficult to create a binary decision rule for discerning patients with or without a condition among the candidate cases, this challenge is common when conducting EHR-based case-control studies. In this work, we propose an innovative method to effectively account for inaccurate case selection.

Our problem can be viewed as belonging to the missing data framework, where the true status of being a noncase or genuine case is unknown for the identified candidate cases. More specifically, the probability model for predicting genuine cases in the combined population of genuine cases and noncases automatically serves as a model for the missingness and naturally brings us to the missing at random (MAR) framework. Our method imputes the true status by modeling the relationship between the genuine case and noncase from a validated subset. This setup represents one key feature of our method. We form a two-layer nested case-control study by treating the genuine cases and noncases as a new case-control data structure along with the primary case-control data. Another key feature of our method is that we impute the missing case status through a semiparametric model which is sufficiently flexible and allows for many covariates.

The imputation step in our approach is nonstandard and plays a different role from what is typically done in the classical imputation literature. Imputation is a widely applied approach for accommodating missing data (Aerts et al., 2002; Little & Rubin, 1987), including missing binary outcomes (Mukaka et al., 2016). But few works focus on case-control studies with missing genuine case status when there is a third group of individuals who are ineligible for the study. Wang et al. (2020) proposed a parametric imputation method for case-control studies in this framework. This method introduces imputation to the estimating equation which corrects the bias caused by the missing genuine case status. To retain the flexibility while bypassing the curse of dimensionality (Wang et al., 2004), we propose a semiparametric sufficient dimension reduction model, and apply an efficient procedure (Ma & Zhu, 2012) to obtain the efficient probability prediction in our imputation procedure. This leads to a stabilized odds ratio parameter estimation in the main model. This modeling and estimation approach allows us to impose minimal assumptions on the missingness scheme while limiting its influence on our odds ratio parameter estimation. In addition, we perform imputation with a probability instead of a randomly generated outcome in an intermediate step of our method. This practice minimizes the potential bias, especially when the prediction probability is extreme (Bernaards et al., 2007), and stabilizes the computation of the overall method.

2 | MODELING THE CASE-CONTROL DATA WITH MISSINGNESS

Let D denote the outcome, with $D = 0$ indicating the controls, $D = 1$ indicating the genuine cases, and $D = 2$ indicating the noncases. The definition of noncases here is simply anyone who is neither a genuine case nor a control. Let N_1 be the sample size of candidate cases ($i = 1,\ldots,N_1$), which includes both genuine cases ($D = 1$) and noncases ($D = 2$). There are also N_0 controls ($D_i = 0, i = N_1 + 1,\ldots,N \equiv N_1 + N_0$). Further, n_1 observations ($i = 1,\ldots,n_1$) from the N_1 candidate-cases are fully observed, and we use the indicator S to denote the validated outcome status. Specifically, $S = 1$ indicates a genuine case and $S = 0$ denotes a noncase. Let $\mathbf{X}$ be a p dimensional covariate vector and $\mathbf{Z}$ be a q-dimensional covariate vector. $\mathbf{X}$ and $\mathbf{Z}$ are allowed to share common components or can even be identical.

Our goal is to fit a logistic regression model using the the genuine cases and controls. When all patients in the model are either genuine cases or controls, the probability that the patient is a genuine case is

$$\mathrm{pr}(D = 1 \mid \mathbf{X}, \mathbf{Z}, D \neq 2) = \mathrm{pr}(D = 1 \mid \mathbf{X}, D \neq 2) = \frac{\exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})}{1 + \exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})}. \tag{1}$$

Note that there is no $\mathbf{Z}$ in model (1). In other words, we use $\mathbf{X}$ to represent all the covariates that are responsible for separating genuine cases from controls in the combined population of genuine cases and controls. Hence, the probability that the patient is a control is

$$\mathrm{pr}(D = 0 \mid \mathbf{X}, \mathbf{Z}, D \neq 2) = \mathrm{pr}(D = 0 \mid \mathbf{X}, D \neq 2) = \frac{1}{1 + \exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})},$$

where $\beta_c \in \mathbf{R}$, $\boldsymbol{\beta}_1 \in \mathbf{R}^p$. We note that the indicators D indexed from $i = n_1 + 1, \ldots, N_1$ in the sample are not observed. We therefore propose to recover the missingness in D by utilizing the underlying structure among the candidate cases. To do this, we assume that given an observation is a candidate case (i.e., is not a control), the probability of being a genuine case is

$$\begin{aligned}\mathrm{pr}(S = 1 \mid \mathbf{X}, \mathbf{Z}, D \neq 0) &= \mathrm{pr}(D = 1 \mid \mathbf{X}, \mathbf{Z}, D \neq 0)\\ &= \mathrm{pr}(S = 1 \mid \mathbf{Z}, D \neq 0)\\ &= \mathrm{pr}(D = 1 \mid \mathbf{Z}, D \neq 0)\\ &= \frac{\exp\{\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})\}}{1 + \exp\{\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})\}}.\end{aligned} \tag{2}$$

Note that there is no $\mathbf{X}$ in (2). The covariates that are responsible for predicting genuine cases from noncases in the combined population of genuine cases and noncases are collected as $\mathbf{Z}$. Hence, the probability of being a noncase is

$$\begin{aligned}\mathrm{pr}(S = 0 \mid \mathbf{X}, \mathbf{Z}, D \neq 0) &= \mathrm{pr}(D = 2 \mid \mathbf{X}, \mathbf{Z}, D \neq 0)\\ &= \mathrm{pr}(S = 0 \mid \mathbf{Z}, D \neq 0)\\ &= \mathrm{pr}(D = 2 \mid \mathbf{Z}, D \neq 0)\\ &= \frac{1}{1 + \exp\{\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})\}},\end{aligned}$$

where $\boldsymbol{\gamma} \in \mathbf{R}^{q\times d}$ and $\eta(\cdot) : \mathbf{R}^d \mapsto \mathbf{R}$ is an arbitrary function. Note that we have used $\mathbf{X}$ to represent all the predictive covariates for (1), and have used $\mathbf{Z}$ to represent all the predictive covariates for (2). Because $\mathbf{X}$ and $\mathbf{Z}$ are allowed to overlap or even be identical, we are not imposing any additional independence assumptions. The function η is unspecified in (2), and the dimension d will be selected via the data-driven method Validated Information Criterion (VIC) (Ma & Zhang, 2015) in practice. When d is selected to be q, the parameter $\boldsymbol{\gamma}$ becomes the identity matrix, and (2) becomes a purely nonparametric model. In this sense, (2) can be viewed as a maximally flexible model and enjoys the same robustness as any nonparametric model against model misspecification.

In this study, we have assumed (2) is correct. Because (2) serves as a missingness mechanism model in our problem formulation, this directly brings us to the MAR framework. In the special case when $\mathbf{Z}$ contains only 1, corresponding to the intercept term, the problem degenerates to missing completely at random (MCAR), and our method will still apply. On the other hand, if some important covariates that are related to D are not included in $\mathbf{Z}$, then (2) will be a misspecified model. In this case, the estimation procedure will break down. Indeed, in this case, the missingness of S will be dependent on that unobserved covariates, which may be further related to whether or not $D = 0$ or $D = 1$ in the population of controls and genuine cases combined. Hence, we are actually in the missing not at random (MNAR) framework. It is well known that any method developed by assuming MAR while the true data structure is MNAR will produce biased results.

3 | NESTED SEMIPARAMETRIC METHODOLOGY

3.1 | Estimating equation

According to our proposed model, the estimating equation for the case-control odds ratio parameter $\boldsymbol{\beta}_c$ and $\boldsymbol{\beta}_1$ in (1) is equivalent to

$$\begin{aligned}
&\sum_{i=1}^{n_1}(1,\mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}} S_i \left\{ 1 - \frac{\exp(\beta_c^* + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)}{1+\exp(\beta_c^* + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)} \right\} \\
&+ \sum_{i=n_1+1}^{N_1}(1,\mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}} \widetilde{S}_i \left\{ 1 - \frac{\exp(\beta_c^* + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)}{1+\exp(\beta_c^* + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)} \right\} \\
&+ \sum_{i=N_1+1}^{N}(1,\mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}} \left\{ 0 - \frac{\exp(\beta_c^* + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)}{1+\exp(\beta_c^* + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)} \right\} = \mathbf{0},
\end{aligned}$$

where $\widetilde{S}_i, i = n_1 + 1, \ldots, N_1$, denotes the hypothetical case indicator within the unobserved $n_2 \equiv N_1 - n_1$ samples whose status D can be either a genuine case or noncase. In other words, $\widetilde{S}_i = 1$ if the ith candidate case is a genuine case and $\widetilde{S}_i = 0$ if it is a noncase. Because $\widetilde{S}_i$'s are not available, our intention is to first impute $\widetilde{S}_i$'s using (2).

3.2 | Semiparametric imputation model

Following the multiple imputation idea, imputing $\widetilde{S}_i$ is equivalent to replacing $\widetilde{S}_i$'s with the probabilities predicted by (2). We first need to estimate the parameters in (2). To this end, we take advantage of an efficient semiparametric method (Ma & Zhu, 2012) to estimate the unknown component $\eta(\cdot)$ and the high-dimensional parameter $\boldsymbol{\gamma}$ simultaneously. In order to avoid the identifiability issue, we assume the upper $d \times d$ block of $\boldsymbol{\gamma}$ is the identity and only the lower $(p-d) \times d$ block of $\boldsymbol{\gamma}$ needs to be estimated. In the semiparametric model (2), the nuisance parameters are $\eta(\cdot)$ and $\mathbf{f_Z}(\mathbf{z})$, the density of covariate $\mathbf{Z}$. The corresponding nuisance tangent space is $\Lambda = \Lambda_1 \oplus \Lambda_2$, where

$$\Lambda_1 = \left[\mathbf{a}(\mathbf{Z}) : E\{\mathbf{a}(\mathbf{Z})\} = 0, \;\; \text{for all } \;\mathbf{a}(\mathbf{Z}) \in \mathbf{R}^{(q-d)d}\right],$$
$$\Lambda_2 = \left[\left\{S - \frac{e^{\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})}}{1+e^{\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})}}\right\}\mathbf{h}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}) : \;\; \text{for all } \;\mathbf{h}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}) \in \mathbf{R}^{(q-d)d}\right].$$

The efficient score is

$$\operatorname{vecl}\left[\{\mathbf{Z} - E(\mathbf{Z} \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})\}\boldsymbol{\eta}'(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})^{\mathrm{T}}\left\{S - \frac{e^{\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})}}{1+e^{\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})}}\right\}\right], \tag{3}$$

where "vecl" is vectorizing the lower $(p-d) \times d$ block of a matrix. Hence we can solve the estimating equation

$$\sum_{i=1}^{n_1}\operatorname{vecl}\left[\{\mathbf{Z}_i - E(\mathbf{Z}_i \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)\}\boldsymbol{\eta}'(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)^{\mathrm{T}}\left\{S_i - \frac{e^{\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)}}{1+e^{\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)}}\right\}\right] = \mathbf{0}, \tag{4}$$

to obtain an efficient estimator for $\boldsymbol{\gamma}$.

We need to point out that when efficiency of the estimation of $\boldsymbol{\gamma}$ is not sought after, (4) can be generalized to the following form

$$\sum_{i=1}^{n_1}\operatorname{vecl}\left(\{\mathbf{Z}_i - \widehat{E}(\mathbf{Z}_i \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)\}\left[\mathbf{g}(S_i, \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i) - \widehat{E}\{\mathbf{g}(S_i, \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i) \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i\}\right]\right) = \mathbf{0},$$

where $\mathbf{g}(\cdot,\cdot)$ is an arbitrary nontrivial function on $\mathbf{R}^d$. This estimator retains the consistency of the $\boldsymbol{\gamma}$ estimation as well (Ma & Zhu, 2012).

Since both $E(\mathbf{Z}_i \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)$ and $\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)$ are unknown in (4), we use the following approach to estimate these two quantities. First, we posit a working model η^*, and its corresponding derivative is $\boldsymbol{\eta}^{*\prime}$. Let $\widehat{E}(\mathbf{Z}_i \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)$ be a nonparametric estimator of $E(\mathbf{Z}_i \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)$, for example, a kernel estimation. This yields the estimating equation

$$\sum_{i=1}^{n_1}\operatorname{vecl}\left[\left\{\mathbf{Z}_i - \widehat{E}(\mathbf{Z}_i \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)\right\}\boldsymbol{\eta}^{*\prime}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)^{\mathrm{T}}\left\{S_i - \frac{e^{\eta^*(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)}}{1+e^{\eta^*(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)}}\right\}\right] = \mathbf{0}. \tag{5}$$

Write the estimator as $\widehat{\boldsymbol{\gamma}}_1$ which is a consistent estimator of $\boldsymbol{\gamma}$. Second, we estimate η and $\boldsymbol{\eta}'$ by solving the equations of b_0 and $\mathbf{b}_1$ from

$$\sum_{i=1}^{n_1}\left[S_i - \frac{\exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\mathbf{Z}_i - \mathbf{t})\}}{1+\exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\mathbf{Z}_i - \mathbf{t})\}}\right]\begin{pmatrix} 1 \\ \widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\mathbf{Z}_i - \mathbf{t}\end{pmatrix}K_h(\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\mathbf{Z}_i - \mathbf{t}) = 0, \tag{6}$$

to obtain the estimation of $\widehat{\eta}(\mathbf{t})$ and $\widehat{\boldsymbol{\eta}}'(\mathbf{t})$ at any $\mathbf{t} \in \mathbf{R}^d$. Here, $K(\cdot)$ is a kernel function and $K_h(\cdot) = K(\cdot/h)/h$. Finally, we plug $\widehat{\eta}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})$, $\widehat{\boldsymbol{\eta}}'(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})$ and $\widehat{E}(\mathbf{Z}_i \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)$ into (4) and solve for the efficient estimator $\hat{\boldsymbol{\gamma}}$ from the estimating equation

$$\sum_{i=1}^{n_1} \text{vecl}\left[\left\{\mathbf{Z}_i - \hat{E}(\mathbf{Z}_i \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)\right\}\hat{\boldsymbol{\eta}}'(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)^{\mathrm{T}}\left\{S_i - \frac{e^{\hat{\eta}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)}}{1+e^{\hat{\eta}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)}}\right\}\right] = \mathbf{0}. \tag{7}$$

3.3 | Nested estimating equation for odds ratio parameters

By incorporating the estimation $\hat{\eta}(\cdot)$ and $\hat{\boldsymbol{\gamma}}$ from (6) and (7), we will be able to impute the status of the n_2 candidate cases using the generated model

$$(\widetilde{S}_i \mid D_i \neq 0) \sim \text{Bernoulli}\left[\frac{\exp\{\hat{\eta}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)\}}{1+\exp\{\hat{\eta}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)\}}\right], i = n_1+1,\ldots,N_1. \tag{8}$$

Using multiple imputation, say B imputations and taking the average, then when $B \to \infty$, we obtain

$$B^{-1}\sum_{b=1}^{B}\widetilde{S}_{ib} \to \frac{\exp\{\hat{\eta}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)\}}{1+\exp\{\hat{\eta}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)\}},$$

in probability, and hence, we get the estimating equation

$$\begin{aligned}\mathbf{0} = &\sum_{i=1}^{n_1}(1,\mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}S_i\left\{1 - \frac{\exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)}{1+\exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)}\right\}\\ &+ \sum_{i=n_1+1}^{N_1}(1,\mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}\frac{\exp\{\hat{\eta}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)\}}{1+\exp\{\hat{\eta}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)\}}\left\{1 - \frac{\exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)}{1+\exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)}\right\}\\ &+ \sum_{i=N_1+1}^{N}(1,\mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}\left\{0 - \frac{\exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)}{1+\exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)}\right\}.\end{aligned} \tag{9}$$

We solve this equation to obtain $\hat{\beta}_c$ and $\hat{\boldsymbol{\beta}}_1$. Following Chen and Ibrahim (2014), considering infinite B will eliminate the additional between-imputation variation. Note that the estimation of β_c and $\boldsymbol{\beta}_1$ is completely separated from the estimation of η and $\boldsymbol{\gamma}$. Both estimation procedures are standard, hence the computation is not challenging. Below, we provide the detailed algorithm.

Step 1 Obtain an initial estimation $\hat{\boldsymbol{\gamma}}_1$:
Obtain an initial estimator $\hat{\boldsymbol{\gamma}}_1$ from solving (5) based on data $\{S_i, \mathbf{Z}_i\}_{i=1}^{n_1}$. In (5), η^* and $\boldsymbol{\eta}^{*\prime}$ are from a working model and

$$\hat{E}(\mathbf{Z}_i \mid \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{Z}_i) = \frac{\sum_{k=n_1+1}^{N_1}\mathbf{Z}_k K_h(\boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{Z}_k - \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{Z}_i)}{\sum_{k=n_1+1}^{N_1}K_h(\boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{Z}_k - \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{Z}_i)}.$$

Step 2 Estimate $\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)$, $\boldsymbol{\eta}'(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)$ and $\boldsymbol{\gamma}$:

21 For any $\boldsymbol{\gamma}$ and $\mathbf{t} = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i, i = 1,\ldots, n_1$, compute $\hat{\eta}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)$ and $\hat{\boldsymbol{\eta}}'(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)$ as the solution of b_0 and $\mathbf{b}_1$ to (6), respectively.

22 Insert $\hat{\eta}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)$ and $\hat{\boldsymbol{\eta}}'(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)$ from Step 221 into (7) and solve it to obtain an updated $\hat{\boldsymbol{\gamma}}_1$ based on the data $\{S_i, \mathbf{Z}_i\}_{i=1}^{n_1}$.

23 Repeat Step 21 and Step 22 until convergence. The resulting $\hat{\boldsymbol{\gamma}}_1$ is the efficient estimator of $\boldsymbol{\gamma}$. Let $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}_1$.

Step 3 Apply the imputation model:
Solve for $\hat{\eta}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)$ and $\hat{\boldsymbol{\eta}}'(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)$ from (6) at each $\mathbf{t} = \hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i, i = n_1 + 1,\ldots, N_1$.

Step 4 Obtain $\hat{\beta}_c$ and $\hat{\boldsymbol{\beta}}_1$:
Compute $\hat{\beta}_c$ and $\hat{\boldsymbol{\beta}}_1$ by solving (9), based on $\hat{\boldsymbol{\gamma}}$ from Step 23, $\left\{\hat{\eta}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i), \hat{\boldsymbol{\eta}}'(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)\right\}_{i=n_1+1}^{N_1}$ from Step 3, and data $\{\mathbf{X}_i\}_{i=1}^{N}$, $\{S_i\}_{i=1}^{n_1}$ and $\{\mathbf{Z}_i\}_{i=n_1+1}^{N_1}$.

All the equations in the algorithm are solved by Powell's algorithm (Powell, 1965). Powell's algorithm is designed for solving multivariate nonlinear problems.

4 | ASYMPTOTIC PROPERTIES

We intend to derive the asymptotic properties of the estimator from (9) for $\boldsymbol{\beta}_1$ by taking into account the variability of $\hat{\eta}(\cdot)$ and $\hat{\boldsymbol{\gamma}}$. For simplicity, we denote the expit function H, that is, $H(t) = \exp(t)/\{1 + \exp(t)\}$ for any t. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_c, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{W}_i = (1, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}$. We prove the results for the case where $\boldsymbol{\gamma}$ is a vector. The case where $\boldsymbol{\gamma}$ is a matrix, the results are similar but with more complex notation in handling matrix operations.

First, we list the regularity conditions for deriving the asymptotic properties.

C1 There exists two constants $0 < c_1 < c_2 < \infty$ so that the sample sizes satisfy $c_1 < n_1/n_2 < c_2$ and $c_1 < N_1/N_0 < c_2$.

C2 The univariate kernel function $K(\cdot)$ is Lipschitz, symmetric and has compact support. It satisfies

$$\int K(u)du = 1, \int u^i K(u)du = 0, 1 \leq i \leq m-1,\ 0 \neq \int u^m K(u)du < \infty,$$

for an integer $m > 2$. The d-dimensional kernel function is a product of d univariate kernel functions, that is, $K_h(\mathbf{u}) = K(\mathbf{u}/h)/h^d = \Pi_{j=1}^{d} K_h(u_j) = \Pi_{j=1}^{d} K(u_j/h)/h^d$ for $\mathbf{u} = (u_1,\ldots, u_d)^{\mathrm{T}}$. Here we use the same K regardless of the dimension of its argument.

C3 The bandwidth $h = O(n_1^{-\kappa})$ for $1/(4m) < \kappa < 1/(2d)$.

C4 The density functions of $\mathbf{Z}$ and $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}$, denoted, respectively, by $f_{\mathbf{Z}}(\mathbf{z})$ and $f_{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z})$, are bounded from below and above. Each entry in the matrices $E\left(\mathbf{Z}\mathbf{Z}^{\mathrm{T}} \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}\right)$ is locally Lipschitz-continuous and bounded from above as a function of $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}$.

C5 $E(\mathbf{Z} \mid \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z})f_{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z})$ and $\mathbf{g}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z})$ are mth-order differentiable and their mth derivatives, as well as $f_{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z})$ are locally Lipschitz-continuous.

C6 (*The boundedness.*) The parameter space $\mathcal{B}$ is bounded.

These are very mild conditions. Condition C1 requires that the proportion of cases and controls do not degenerate to zero both in the population and in the sample. Conditions C2 and C3 are common requirements on the kernel function and the bandwidth. Conditions C4 and C5 assume sufficient smoothness and boundedness of the density of covariates corresponding to the efficient semiparametric method. In order to ensure a unique solution of the parameter estimation, we assume the boundedness of parameter space in Condition C6.

Lemma 1. *Under Conditions C2–C5, especially* $n_1^{1/2}h^4 \to 0$ *and* $n_1h^2 \to \infty$*, using results from* Ma and Zhu (2012)*, we obtain that* $\widehat{\boldsymbol\gamma}$ *is a consistent estimator of* $\boldsymbol\gamma$ *satisfying*

$$n_1^{1/2}(\widehat{\boldsymbol\gamma}-\boldsymbol\gamma) = -n_1^{-1/2}\sum_{i=1}^{n_1}\mathbf{A}_\gamma^{-1}\mathbf{f}(S_i,\mathbf{Z}_i,\boldsymbol\gamma,\mathbf{g}) + o_p(1),$$

where

$$\mathbf{f}(S_i,\mathbf{Z}_i,\boldsymbol\gamma,\mathbf{g}) = vecl\big(\{\mathbf{Z}_i - E(\mathbf{Z}_i \mid \boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i)\}\left[\mathbf{g}(S_i,\boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i) - E\{\mathbf{g}(S_i,\boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i) \mid \boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i\}\right]\big),$$

and

$$\mathbf{A}_\gamma = E\left\{\frac{\partial vecl\big(\{\mathbf{Z}_i - E(\mathbf{Z}_i \mid \boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i)\}\left[\mathbf{g}(S_i,\boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i) - E\{\mathbf{g}(S_i,\boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i) \mid \boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i\}\right]\big)}{\partial vecl(\boldsymbol\gamma)^{\mathrm T}} \mid D_i \neq 0\right\}.$$

We do not include the details of Lemma 1 since it was carefully proved and discussed in Ma and Zhu (2013). Following Ma and Zhu (2013), the above expansion still holds if we replace the pre-decided function $\mathbf{g}(\cdot)$ with the estimated version of the function $\mathbf{g}_0(\cdot)$, where

$$\mathbf{g}_0(S_i,\boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i) \equiv [S_i - H\{\eta(\boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i)\}]\frac{\partial\eta(\boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i)}{\partial(\boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i)}.$$

To further estimate $\eta(\cdot)$ regardless which choice of $\mathbf{g}(\cdot)$ function is used in obtaining $\widehat{\boldsymbol\gamma}$, we propose to simply perform a kernel mean regression followed with a logit transformation, that is, at any $\mathbf{u}_0$, we set

$$\widehat\eta(\mathbf{u}_0,\widehat{\boldsymbol\gamma}) \equiv H^{-1}\left\{\frac{\sum_{i=1}^{n_1}S_iK_h(\widehat{\boldsymbol\gamma}^{\mathrm T}\mathbf{Z}_i - \mathbf{u}_0)}{\sum_{i=1}^{n_1}K_h(\widehat{\boldsymbol\gamma}^{\mathrm T}\mathbf{Z}_i - \mathbf{u}_0)}\right\}.$$

Next, we provide the asymptotic properties of $\widehat{\boldsymbol\beta}$ based on the discussion of $\widehat{\boldsymbol\gamma}$ and $\widehat\eta$.

Theorem 1. *Under Conditions C1–C6,* $\widehat{\boldsymbol\beta}$ *is a consistent estimation of* $\boldsymbol\beta$ *and*

$$N^{1/2}(\widehat{\boldsymbol\beta}-\boldsymbol\beta) \to N(\mathbf{0},\mathbf{A}_\beta^{-1}\mathbf{V}\mathbf{A}_\beta^{-1^{\mathrm T}}),$$

where $\mathbf{V} = n_1N^{-1}\mathbf{V}_1 + n_2N^{-1}\mathbf{V}_2 + N_1N^{-1}\mathbf{V}_3$,

$$\begin{aligned}\mathbf{V}_1 = E\Bigg\{\Bigg(&\mathbf{W}_iS_i\{1-H(\boldsymbol\beta^{\mathrm T}\mathbf{W}_i)\} + \frac{n_2}{n_1}E\left[\mathbf{W}_i\{1-H(\boldsymbol\beta^{\mathrm T}\mathbf{W}_i)\} \mid \boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i\right]\\ &\times\{S_i - E(S_i \mid \boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i)\} - \frac{n_2}{n_1}(\mathbf{B}_2+\mathbf{B}_\gamma)\mathbf{A}_\gamma^{-1}\mathbf{f}(S_i,\mathbf{Z}_i,\boldsymbol\gamma,\mathbf{g})\Bigg)^{\otimes 2}\Bigg\},\\ \mathbf{V}_2 = E\big[&\mathbf{W}_i\mathbf{W}_i^{\mathrm T}\{1-H(\boldsymbol\beta^{\mathrm T}\mathbf{W}_i)\}^2\{E(S_i \mid \boldsymbol\gamma^{\mathrm T}\mathbf{Z}_i)\}^2\big],\\ \mathbf{V}_3 = E\{&\mathbf{W}_i\mathbf{W}_i^{\mathrm T}H(\boldsymbol\beta^{\mathrm T}\mathbf{W}_i)^2\},\end{aligned}$$

and

$$\mathbf{A}_\beta = N_1 N^{-1} E[\mathbf{W}_i \mathbf{W}_i^{\mathrm{T}} H\{\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}_i)\} H(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i)\{1 - H(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i)\} \mid D \neq 0] \\ + N_0 N^{-1} E[\mathbf{W}_i \mathbf{W}_i^{\mathrm{T}} H(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i)\{1 - H(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i)\} \mid D_i = 0].$$

The asymptotic variance in Theorem 1 has the typical sandwich form. The matrix $\mathbf{A}_\beta$ results from the derivative of (9) with respect to $\boldsymbol{\beta}$. The matrix $\mathbf{V}$ contains three components. The first component $\mathbf{V}_1$ captures the variability contributed by the randomness of the fully observed genuine and noncases. The second component $\mathbf{V}_2$ corresponds to the variability due to the randomness of the candidate-cases. The third component $\mathbf{V}_3$ corresponds to the variability due to the randomness of the controls.

Theorem 1 shows that the proposed estimator $\widehat{\boldsymbol{\beta}}$ is consistent with a root-n convergence rate. It also provides an approach to estimate the asymptotic variance of $\widehat{\boldsymbol{\beta}}$. The proof of Theorem 1 is in the Appendix S1.

5 | NUMERICAL SIMULATION

5.1 | Data generation procedure

The population can be divided into three parts, $D = 0, 1$, and 2, according to the model. The ratio between $D = 0$ and $D = 1$ is $1 : \exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})$ and the ratio between $D = 1$ and $D = 2$ is $\exp\{\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})\} : 1$. Thus, the ratio between $D = 0 : D = 1 : D = 2$ is $1 : \exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}) : \exp\{\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} - \eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})\}$. Therefore, we use the following data generating process to conduct finite sample studies.

1. Generate a population following the model

$$\mathrm{pr}(D = 0 \mid \mathbf{X}, \mathbf{Z}) = \frac{1}{1 + \exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}) + \exp\{\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} - \eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})\}},$$
$$\mathrm{pr}(D = 1 \mid \mathbf{X}, \mathbf{Z}) = \frac{\exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})}{1 + \exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}) + \exp\{\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} - \eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})\}},$$
$$\mathrm{pr}(D = 2 \mid \mathbf{X}, \mathbf{Z}) = \frac{\exp\{\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} - \eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})\}}{1 + \exp(\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}) + \exp\{\beta_c + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} - \eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})\}}.$$

2. Sample N_0 observations from the $D = 0$ subpopulation.
3. Sample N_1 observations from the $D = 1$ and $D = 2$ subpopulations combined.
4. Sample n_1 observations from the sub-sample with size N_1 above, set $S = 1$ if $D = 1$ and $S = 0$ if $D = 2$. Mask out the D information on all the N_1 observations.

5.2 | Finite sample study

We study the finite sample performance of our method through simulation studies. In each of our studies, we generate 1000 datasets. In the first study, we generate a $p = 6$ dimensional covariate vector $\mathbf{X}$ from the multivariate normal distribution with mean zero and variance-covariance

TABLE 1 Results of Study 1, based on 1000 simulations with 1000 control-cases and 1000 candidate-cases.

		β_1	β_2	β_3	β_4	β_5	β_6
	True	1.3	−1.3	1	−2.5	2.5	−0.5
Semi	Mean	1.304	−1.321	0.985	−2.525	2.490	−0.487
	Bias	0.004	0.021	0.014	0.025	0.009	0.012
	SD	0.188	0.190	0.197	0.188	0.209	0.194
	$\widehat{\text{SD}}$	0.191	0.191	0.190	0.201	0.198	0.188
	CI	(0.93, 1.68)	(−1.70, −0.96)	(0.61, 1.36)	(−2.92, −2.13)	(2.10, 2.87)	(−0.86, −0.13)
	Coverage	97.3%	95.5%	93.2%	94.3%	93.2%	94.8%
OEE	Mean	1.318	−1.317	1.014	−2.539	2.532	−0.510
	Bias	0.018	0.017	0.014	0.039	0.032	0.010
	SD	0.113	0.119	0.112	0.178	0.168	0.107
	$\widehat{\text{SD}}$	0.128	0.128	0.119	0.182	0.179	0.110
	CI	(0.87, 1.74)	(−1.75, −0.91)	(0.55, 1.42)	(−2.95, −2.08)	(2.06, 2.97)	(−0.94, −0.08)
	Coverage	97.4%	96.4%	96.0%	95.6%	97.2%	96.2%
Naive	Mean	1.770	−1.620	0.836	−1.288	0.956	0.010
	Bias	0.470	0.320	0.163	1.211	1.543	0.510
	SD	0.215	0.210	0.203	0.222	0.215	0.195
	$\widehat{\text{SD}}$	0.203	0.199	0.192	0.207	0.201	0.191
	CI	(0.94, 1.63)	(−1.66, −0.95)	(0.63, 1.32)	(−3.15, −2.41)	(1.76, 2.48)	(−0.47, 0.22)
	Coverage	32.9%	59.2%	84.5%	0.0%	0.0%	26.7

Abbreviations: Bias, average of absolute bias; CI, average 95% confidence interval; Coverage, 95% coverage of corresponding estimation; Mean, average of $\widehat{\beta}$; SD, sample standard deviation; $\widehat{\text{SD}}$, average of the estimated standard deviations of the corresponding estimation.

matrix equal to the identity. We set $(Z_1, Z_2, Z_3)^{\mathrm{T}} = (X_1, X_2, X_3)^{\mathrm{T}}$, and generate $(Z_4, Z_5, Z_6)^{\mathrm{T}}$ from the multivariate normal distribution with mean zero and variance–covariance matrix identity $\mathbf{I}_3$. Thus, the dimension of $\mathbf{Z}$ is $q = 6$. We set the true parameter values $\boldsymbol{\gamma} = (1, 1.3, -1.3, 1, -1.5, 1.5)^{\mathrm{T}}$, $\boldsymbol{\beta} = (1.3, -1.3, 1, -2.5, 2.5, -0.5)^{\mathrm{T}}$ and consider the true η function to be $\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}) = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}$. We experiment with sample size $N_0 + N_1 = 2000$. Of the 2000 samples, $n_1 = 500$ are observed candidate cases whose true status, D is known ($D = 1$ or $D = 2$), $N_0 = 1000$ are controls ($D = 0$), and the remaining $N_1 - n_1 = 500$ are candidate cases whose true status is unobserved (unknown D where $D = 1$ or $D = 2$).

In the second study, we repeat the same analysis as in the first study, except here, the true η function is $\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}) = 1 - (\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})^2$. In the first two studies, we set the bandwidth in the nonparametric estimator to be $c\mathrm{SD}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})(N_1 - n_1)^{-1/3}$ and c is a constant in the range of 0.1 to 10. The results are insensitive in this range of c.

The performances of $\widehat{\boldsymbol{\beta}}$ in the first simulation study are in Table 1 and Figure 1. We can see clearly that the estimators of $\boldsymbol{\beta}$ have very small biases and SDs. We also report the estimated SD of the main regression parameter estimator $\widehat{\boldsymbol{\beta}}$ using the asymptotic results provided in Section 4. Clearly, the average estimated SD is close to the sample SD and the resulting 95% confidence interval has coverage close to the nominal level. The estimation of $\boldsymbol{\gamma}$ is also consistent

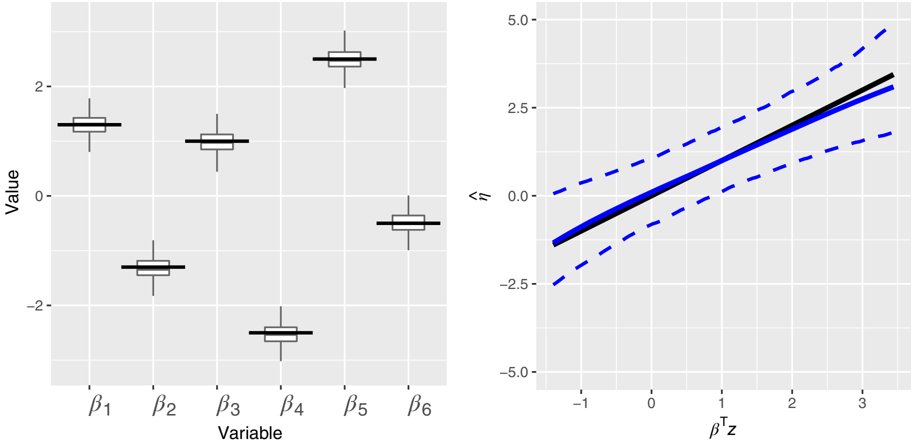

FIGURE 1 Simulation performance in Study 1. Left panel: boxplot of $\hat{\beta}$. Right panel: performance of $\hat{\eta}$. Black solid line is the truth for both panels. Blue solid line: mean of $\hat{\eta}$, Lower blue dashed line: 0.05 quantile of $\hat{\eta}$, Upper blue dashed line: 0.95 quantile of $\hat{\eta}$.

with very small SEs. The absolute biases of $\hat{\boldsymbol{\gamma}}_{-1}$'s are $(0.025, 0.004, 0.052, 0.022, 0.001)^{\mathrm{T}}$ and the corresponding SEs are $(0.609, 0.552, 0.649, 0.599, 0.610)^{\mathrm{T}}$, where subscript $_{-1}$ indicates the indices corresponding to all values of $\hat{\boldsymbol{\gamma}}$ except for the first one. More details of estimating $\boldsymbol{\gamma}$ have been rigorously discussed in Ma and Zhu (2012) and Ma and Zhu (2013). From Figure 1 we can see that the mean of $\hat{\eta}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z})$ is close to the true function $\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z})$ overall, with the performance at the boundary worse than the interior as is typical for all nonparametric estimators. The results of estimating $\boldsymbol{\beta}$ in the second simulation study are in Table 2 and Figure 2. The estimation of $\boldsymbol{\gamma}_{-1}$ has small absolute biases with value $(0.072, 0.051, 0.076, 0.092, 0.086)^{\mathrm{T}}$ and the SEs are $(0.470, 0.511, 0.482, 0.482, 0.518)^{\mathrm{T}}$. The same conclusion can be drawn as in the first simulation. For comparison, we also report the results from a naive method and original EE (OEE) method (Wang et al., 2020). The naive method treats the noncases as genuine cases. The OEE method proposes a weighted estimating equation to overcome the bias from the naive method. The weight is calculated by estimating the probability of being a genuine case given covariates parametrically. By doing so, its odds ratio parameter estimation is unbiased. OEE performs well in the first study because the model is correctly specified. On the contrary, OEE performs poorly in the second study when the model is misspecified. Without surprise, the naive method performs poorly in both studies.

We also conduct a third simulation to evaluate the performance of the proposed model in a high-dimensional covariate case which imitates the dilated cardiomyopathy dataset. In this scenario, we generate $\mathbf{X}$ and $\mathbf{Z}$ from independent standard uniform distribution with dimension $p = q = 20, d = 2$, and they share 10 common covariates. The η function is $\eta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}) = \cos\{(\boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{Z})^2 + (\boldsymbol{\gamma}_2^{\mathrm{T}}\mathbf{Z})^2\}$, where $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ stand for the first and second column vectors in $\boldsymbol{\gamma}$, respectively. The sample sizes are $N_1 = 2000$ and $N_0 = 5000$ for cases and controls. Among the candidate cases, we randomly mask out D for 1000 observations. The bandwidth in the nonparametric estimator is set to be $c\{\mathrm{SD}(\boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{Z}) + \mathrm{SD}(\boldsymbol{\gamma}_2^{\mathrm{T}}\mathbf{Z})\}(N_1 - n_1)^{-1/5}$, and c is a constant in the range of 0.1 to 10. The results are insensitive in this range of c.

The $\boldsymbol{\beta}_{-1}$ estimation of the third simulation is reported in Table 3 along with the corresponding SD estimation and 95% coverage probability. We also illustrate the estimated $\eta(\cdot)$, that is, $\hat{\eta}(\cdot)$, in Figure 3 where the mean and 95% confidence band are reported. We can see the estimation

TABLE 2 Results of Study 2, based on 1000 simulations with 1000 control-cases and 1000 candidate-cases.

		β_1	β_2	β_3	β_4	β_5	β_6
	True	1.3	−1.3	1	−2.5	2.5	−0.5
Semi	Mean	1.281	−1.283	0.981	-2.427	2.462	−0.469
	Bias	0.018	0.017	0.018	0.073	0.037	0.030
	SD	0.164	0.165	0.149	0.240	0.221	0.141
	$\widehat{\text{SD}}$	0.153	0.152	0.142	0.225	0.209	0.138
	CI	(0.97, 1.49)	(−1.50, −0.96)	(0.70, 1.20)	(−2.73, −1.99)	(2.01, 2.75)	(−0.74, −0.23)
	Coverage	93.3%	91.9%	94.2%	91.3%	92.6%	93.4%
OEE	Mean	0.804	−0.804	0.620	−1.400	1.554	−0.262
	Bias	0.495	0.495	0.379	1.099	0.945	0.237
	SD	0.081	0.081	0.076	0.122	0.117	0.091
	$\widehat{\text{SD}}$	0.110	0.110	0.107	0.118	0.122	0.102
	CI	(0.59, 0.99)	(−1.00, −0.59)	(0.42, 0.81)	(−1.44, −1.01)	(1.28, 1.73)	(−0.49, −0.11)
	Coverage	0.1%	0.2%	2.4%	0.0%	0.0%	35.6%
Naive	Mean	1.559	−1.683	0.975	−0.885	1.126	−0.251
	Bias	0.259	0.383	0.024	1.614	1.373	0.248
	SD	0.201	0.207	0.196	0.205	0.224	0.203
	$\widehat{\text{SD}}$	0.198	0.200	0.192	0.200	0.204	0.192
	CI	(0.55, 0.80)	(−0.81, −0.55)	(0.40, 0.64)	(−1.50, −1.19)	(1.11, 1.40)	(−0.33, −0.09)
	Coverage	74.1%	49.9%	93.9%	0.0%	0.0%	75.3%

Abbreviations: Bias, average of absolute bias; CI, average 95% confidence interval; Coverage, 95% coverage of corresponding estimation; Mean, average of $\widehat{\beta}$; SD, sample standard deviation; $\widehat{\text{SD}}$, average of the estimated standard deviations of the corresponding estimation.

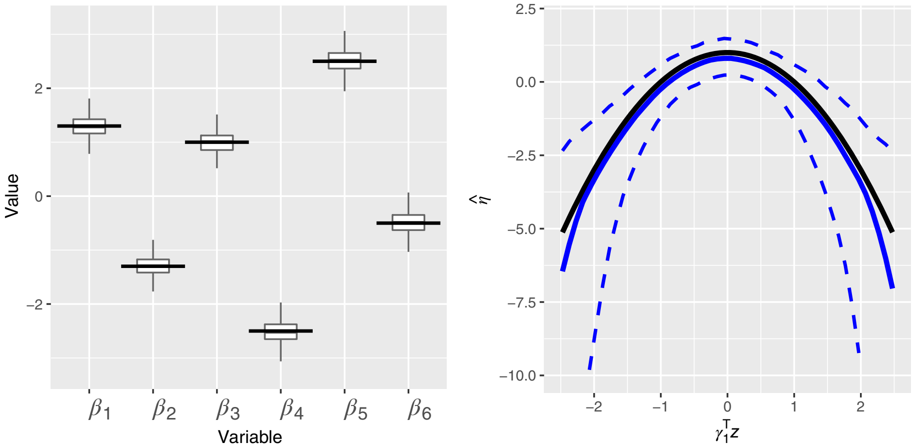

FIGURE 2 Simulation performance in Study 2. Left panel: boxplot of $\widehat{\beta}$. Right panel: performance of $\widehat{\eta}$. Black solid line is the truth for both panels. Blue solid line: mean of $\widehat{\eta}$, Lower blue dashed line: 0.05 quantile of $\widehat{\eta}$, Upper blue dashed line: 0.95 quantile of $\widehat{\eta}$.

TABLE 3 Results of Study 3, based on 1000 simulations with 2000 control-cases and 2000 candidate-cases.

parameter	True	Bias	SD	$\widehat{\text{SD}}$	Coverage	CI
β_1	−0.562	0.021	0.108	0.109	94.3%	(−0.31, 0.19)
β_2	−0.025	4.6×10^{-4}	0.103	0.109	97.9%	(−0.42, 0.07)
β_3	−0.346	3.7×10^{-3}	0.105	0.109	95.2%	(−0.31, 0.19)
β_4	−0.214	4.3×10^{-3}	0.105	0.109	96.5%	(−0.46, 0.03)
β_5	0.059	0.021	0.116	0.109	93.5%	(−0.35, 0.12)
β_6	−0.002	3.5×10^{-3}	0.111	0.109	92.9%	(−0.40, 0.11)
β_7	−0.343	7.7×10^{-3}	0.103	0.109	96.1%	(−0.38, 0.02)
β_8	−0.108	4.8×10^{-3}	0.119	0.109	93.0%	(−0.40, 0.11)
β_9	−0.528	7.4×10^{-5}	0.108	0.110	96.5%	(−0.42, 0.01)
β_{10}	0.283	1.5×10^{-3}	0.106	0.109	94.8%	(−0.37, 0.11)
β_{11}	−0.589	1.2×10^{-2}	0.123	0.110	90.9%	(−0.38, 0.11)
β_{12}	0.025	1.3×10^{-2}	0.112	0.109	97.0%	(−0.38, 0.10)
β_{13}	0.057	2.8×10^{-3}	0.105	0.109	95.4%	(−0.35, 0.12)
β_{14}	−0.555	4.6×10^{-3}	0.110	0.110	95.6%	(−0.40, 0.09)
β_{15}	−0.302	8.9×10^{-3}	0.124	0.109	92.9%	(−0.39, 0.05)
β_{16}	0.375	3.0×10^{-3}	0.116	0.109	93.0%	(−0.39, 0.11)
β_{17}	−0.034	8.2×10^{-3}	0.131	0.109	88.9%	(−0.39, 0.04)
β_{18}	0.314	3.4×10^{-3}	0.117	0.109	92.1%	(−0.33, 0.18)
β_{19}	−0.557	4.5×10^{-3}	0.122	0.109	93.0%	(−0.40, 0.04)
β_{20}	−0.156	4.0×10^{-3}	0.113	0.109	94.0%	(−0.35, 0.11)

Abbreviations: Bias, average of absolute bias; CI, average 95% confidence interval; Coverage, 95% coverage of corresponding estimation; Mean, average of $\widehat{\boldsymbol{\beta}}$; SD, sample standard deviation; $\widehat{\text{SD}}$, average of the estimated standard deviations of the corresponding estimation.

captures the trend of η even in such a high-dimensional situation. A referee points out that the estimated SDs are almost identical to each other. This is because the covariate components in $\mathbf{X}, \mathbf{Z}$ happen to be generated from the same distribution in this simulation.

Following a referee's request, we further conduct two additional simulation studies, where the purpose is to investigate the performance of our method in small sample size situation and in the MNAR situation, respectively. Specifically, in Study 4, the data is generated from the same model and parameter setting as in Study 2 but with sample size $n_1 = 50$, $N_1 = 100$, and $N_0 = 100$, hence the total size is $N_1 + N_0 = 200$. The results are provided in Table 4 and Figure 4. These results show that when the sample total size is 200, our method deteriorates, although it still performs better than the OEE and naive methods. Our method captures the missingness mechanism well in terms of estimating η, although it has a wider confidence band than in Study 2 due to the very small sample size. In the fifth simulation, we set $(Z_1, Z_2, Z_3)^{\mathrm{T}} = (X_1, X_2, X_3)^{\mathrm{T}}$ as before, and generate $(Z_4, Z_5, Z_6, Z_7, Z_8)^{\mathrm{T}}$ from the multivariate normal distribution with mean zero and variance-covariance matrix equal to the identity $\mathbf{I}_5$. Thus, the dimension of $\mathbf{Z}$ is $q = 8$. Otherwise, all the settings are the same as in Study 2. Thus, the data generation mechanism for the true outcome status D depends on all the covariates in $\mathbf{Z}$, while we use only $(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6)^{\mathrm{T}}$

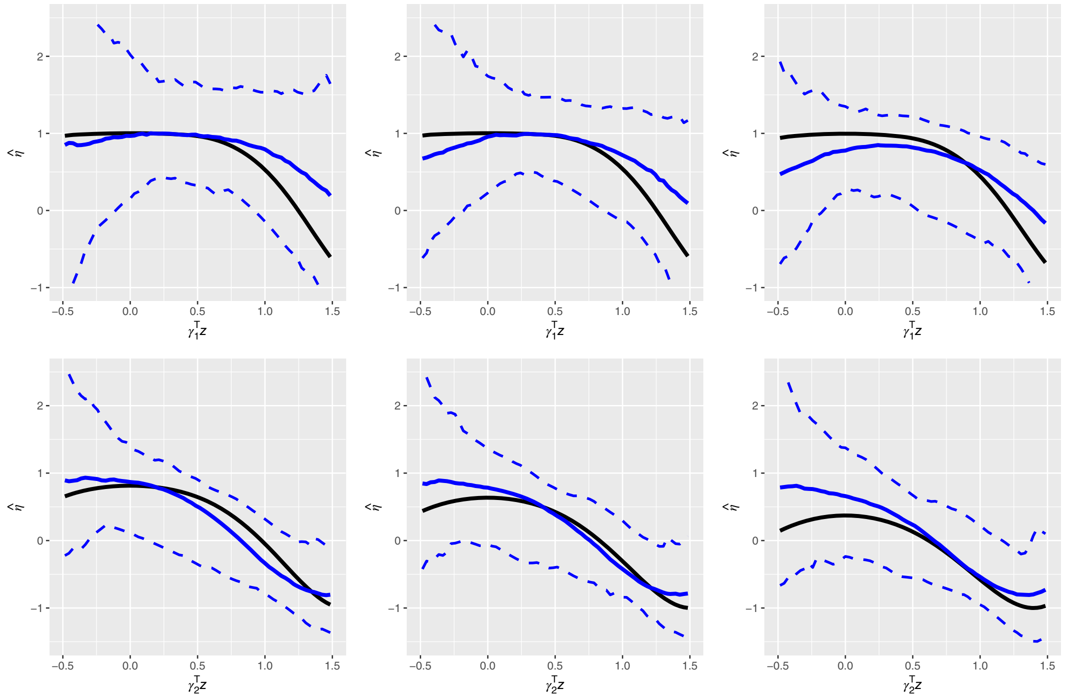

FIGURE 3 Performance of $\hat{\eta}$ in Study 3. First line: $\hat{\eta}$ versus $\boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{Z}$ at $\boldsymbol{\gamma}_2^{\mathrm{T}}\mathbf{Z} = -0.1, 0, 0.3$, respectively, from left to right. Second line: $\hat{\eta}$ versus $\boldsymbol{\gamma}_2^{\mathrm{T}}\mathbf{Z}$ at $\boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{Z} = 0.8, 0.9, 1.1$, respectively, from left to right. Black solid line: true η, Blue solid line: mean of $\hat{\eta}$, blue dashed line: 0.025 and 0.975 quantile of $\hat{\eta}$.

to estimate the imputation model (2). This is an MNAR setting and it mimics the situation that two covariates $(Z_7, Z_8)^{\mathrm{T}}$ are not observed. The estimation of $\boldsymbol{\beta}_{-1}$ and η are reported in Table 5 and Figure 5. Compared to the correctly specified model in the second simulation, our method retains the major trend in the estimation with slight biases. It indicates some degree of robustness of our method when the missingness model is misspecified.

6 | DILATED CARDIOMYOPATHY DATASET ANALYSIS

We apply the proposed model to the analysis of a dilated cardiomyopathy case-control study using data form the University of Pennsylvania EHR. The subjects in this study are patients of European descent who are enrolled in the Penn Biobank. The main goal of the study is to assess the association of the hiPSI TTNtv with the phenotype dilated cardiomyopathy. The adjusting covariates include a patient's gender, age, a collection of ICD-9 and ICD-10 codes related to dilated cardiomyopathy, summarized measures derived from echocardiograms (EKGs), and genetic principal components for helping control for population stratification. Additionally, a number of individuals in the data set are missing summary measures for EKGs, so we include an indicator for each patient to indicate whether or not each of the summary measures is available. Patients' ICD-9 and ICD-10 codes were mapped to PheWAS codes (Haggerty et al., 2019). In this analysis, a candidate case is defined as one who had at least one visit for dilated cardiomyopathy or has had at least one of the following diagnosis codes: I42.0, 425.4, 425.8, 425.9, I42.8, and I42.9.

TABLE 4 Results of Study 4, based on 1000 simulations with 100 control-cases and 100 candidate-cases.

		β_1	β_2	β_3	β_4	β_5	β_6
	True	1.3	−1.3	1	−2.5	2.5	−0.5
Semi	Mean	1.406	−1.424	1.066	−2.534	2.715	−0.433
	Bias	0.384	0.391	0.362	0.669	0.585	0.362
	SD	0.669	0.706	0.597	1.092	1.008	0.588
	$\widehat{\text{SD}}$	0.511	0.517	0.476	0.770	0.733	0.456
	CI	(0.45, 2.27)	(−2.70, −0.40)	(0.17, 1.86)	(−4.58, −1.05)	(1.35, 3.89)	(−1.44, 0.44)
	Coverage	88.5%	89.3%	89.7%	84.6%	86.8%	87.2%
OEE	Mean	0.861	−0.881	0.664	−1.522	1.695	−0.280
	Bias	0.460	0.444	0.361	1.031	0.860	0.280
	SD	0.305	0.316	0.291	0.496	0.471	0.332
	$\widehat{\text{SD}}$	2.427	2.580	2.465	2.697	2.133	2.678
	CI	(−3.87, 5.61)	(−5.92, 4.15)	(−4.15, 5.49)	(−6.79, 3.73)	(−2.47, 5.89)	(−5.52, 4.97)
	Coverage	100.0%	100.0%	100.0%	100.0%	99.8%	100.0%
Naive	Mean	0.675	−0.686	0.522	−1.160	1.319	−0.158
	Bias	0.643	0.622	0.485	1.361	1.205	0.344
	SD	0.199	0.211	0.198	0.260	0.261	0.204
	$\widehat{\text{SD}}$	0.203	0.203	0.198	0.235	0.240	0.194
	CI	(0.28, 1.07)	(−1.09, −0.28)	(0.13, 0.90)	(−1.63, −0.68)	(0.84, 1.77)	(−0.54, 0.22)
	Coverage	17.8%	19.2%	32.5%	1.2%	2.7%	55.1%

Abbreviations: Bias, average of absolute bias; CI, average 95% confidence interval; Coverage, 95% coverage of corresponding estimation; Mean, average of $\hat{\beta}$; SD, sample standard deviation; $\widehat{\text{SD}}$, average of the estimated standard deviations of the corresponding estimation.

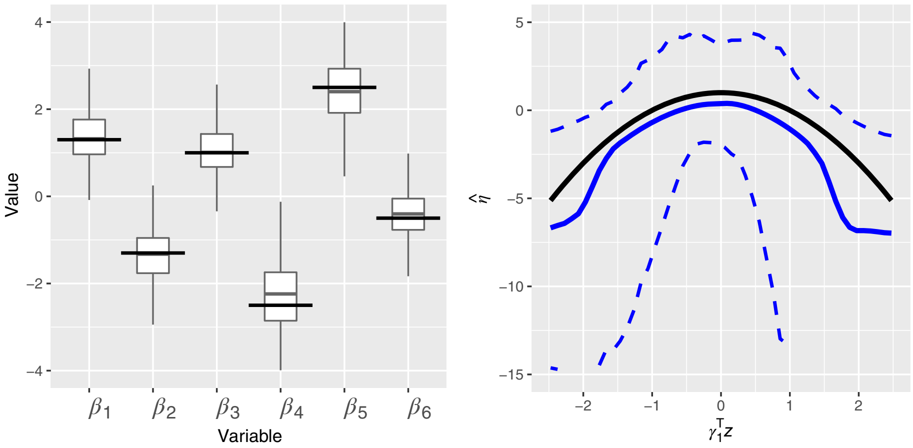

FIGURE 4 Simulation performance in Study 4. Left panel: boxplot of $\hat{\beta}$. Right panel: performance of $\hat{\eta}$. Black solid line is the truth for both panels. Blue solid line: mean of $\hat{\eta}$, Lower blue dashed line: 0.05 quantile of $\hat{\eta}$, Upper blue dashed line: 0.95 quantile of $\hat{\eta}$.

TABLE 5 Results of misspecified Study 2, based on 1000 simulations with 1000 control-cases and 1000 candidate-cases

		β_1	β_2	β_3	β_4	β_5	β_6
	True	1.3	−1.3	1	−2.5	2.5	−0.5
Semi	Mean	1.229	−1.225	0.948	−2.356	2.380	−0.481
	Bias	0.109	0.110	0.095	0.175	0.165	0.087
	SD	0.131	0.139	0.120	0.201	0.208	0.127
	$\widehat{\text{SD}}$	0.135	0.136	0.127	0.187	0.189	0.130
	CI	(0.97, 1.49)	(−1.50, −0.96)	(0.70, 1.20)	(−2.73, −1.99)	(2.01, 2.75)	(−0.74, −0.23)
	Coverage	91.4%	89.9%	92.1%	84.9%	86.3%	92.4%
OEE	Mean	0.791	−0.796	0.611	−1.224	1.506	−0.297
	Bias	0.509	0.504	0.389	1.276	0.994	0.203
	SD	0.081	0.082	0.076	0.113	0.126	0.099
	$\widehat{\text{SD}}$	0.103	0.103	0.100	0.108	0.114	0.097
	CI	(0.59, 0.99)	(−1.00, −0.59)	(0.42, 0.81)	(−1.44, −1.01)	(1.28, 1.73)	(−0.49, −0.11)
	Coverage	0.0%	0.0%	1.5%	0.0%	0.0%	46.3%
Naive	Mean	0.675	−0.680	0.519	−1.347	1.253	−0.207
	Bias	0.625	0.620	0.481	1.153	1.247	0.293
	SD	0.064	0.065	0.058	0.081	0.081	0.065
	$\widehat{\text{SD}}$	0.064	0.064	0.062	0.077	0.076	0.062
	CI	(0.55, 0.80)	(−0.81, −0.55)	(0.40, 0.64)	(−1.50, −1.19)	(1.11, 1.40)	(−0.33, −0.09)
	Coverage	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%

Abbreviations: Bias, average of absolute bias; CI, average 95% confidence interval; Coverage, 95% coverage of corresponding estimation; Mean, average of $\hat{\beta}$; SD, sample standard deviation; $\widehat{\text{SD}}$, average of the estimated standard deviations of the corresponding estimation.

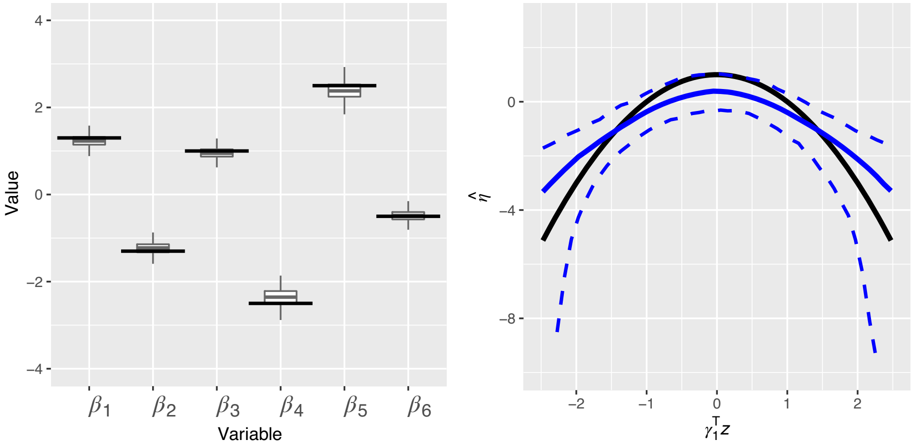

FIGURE 5 Simulation performance in Study 5. Left panel: boxplot of $\hat{\beta}$. Right panel: performance of $\hat{\eta}$. Black solid line is the truth for both panels. Blue solid line: mean of $\hat{\eta}$, Lower blue dashed line: 0.05 quantile of $\hat{\eta}$, Upper blue dashed line: 0.95 quantile of $\hat{\eta}$.

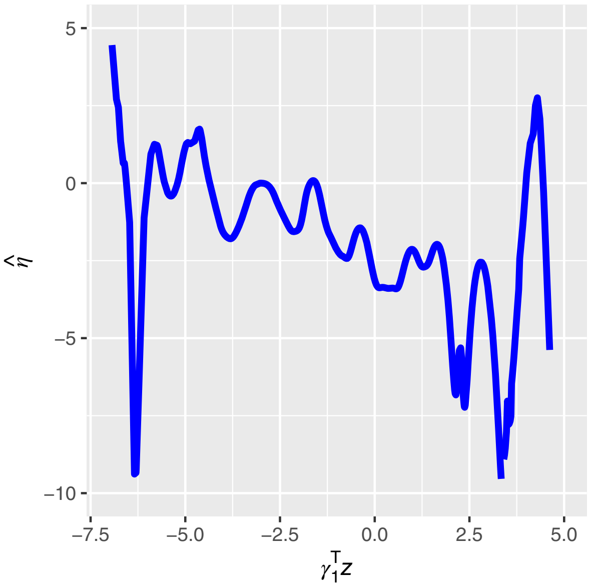

FIGURE 6 Estimated logit probability of genuine cases among candidate-cases in the dilated cardiomyopathy dataset analysis, that is, $\hat{\eta}$.

The dilated cardiomyopathy visits are any encounters with the words "Dilated Cardiomyopathy" in the clinical notes. These encounters are identified using natural language processing, a technique for text mining. The genuine cases were defined using an algorithm validated by the clinician team, and the remaining patients in the case pool who did not meet the genuine case definition were treated as noncases. Everyone who does not match the definition of a candidate case is considered to be a control. The sample size of candidate cases is 1723, where 400 individuals were fully observed. We obtained the validated sample by randomly drawing a subset of 400 individuals from the subset of candidate cases. The genuine case status, D, was retained for these 400 individuals, and D was masked for the remainder of the candidate cases. We also have 6120 controls. The bandwidth in the nonparametric estimator is set to be $SD(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z})1323^{-1/3}$. Before applying the proposed semiparametric method in estimating $\boldsymbol{\gamma}_i, i = 1, 2, \ldots, d$, we first determine the number of indices d by minimizing the VIC (Ma & Zhang, 2015), an information criterion for sufficient dimension reduction models that takes goodness of fit and dimensionality into account simultaneously under mild assumptions. The best selection of d corresponds to the smallest VIC.

The most preferable choice of d is 1, corresponding to VIC = 123.462. The result of η estimation is reported in Figure 6. According to the plot, the probability of being a genuine case is decreasing when $\boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{Z}$ increases with small perturbation. Large variability occurs at both ends due to fewer data points observed. The results of estimating $\boldsymbol{\beta}$ from three methods are reported in Table 6. It is shown that the coefficient for the hiPSI TTNtv is significant with the same sign in all methods. Meanwhile, the estimation efficiency of the odds ratio parameter in the proposed method is higher than the OEE method. OEE estimates Age to be nonsignificant with the opposite sign compared to other methods. All methods conclude that only the first genetic principle component is significant. Compared to the naive analysis, which treats all missing values as genuine cases, the proposed method does not lose too much efficiency.

TABLE 6 Results of the dilated cardiomyopathy dataset analysis.

	Proposed			**OEE**			**naive**		
Parameter	**est**	$\widehat{\text{SD}}$	**CI**	**est**	$\widehat{\text{SD}}$	**CI**	**est**	$\widehat{\text{SD}}$	**CI**
β_0	−3.34	0.196	(−3.72, −2.95)	−2.742	0.292	(−3.31, −2.16)	−1.975	0.162	(−2.29, −1.65)
Age	0.098	0.038	(0.02, 0.17)	−0.033	0.058	(−0.14, 0.08)	0.123	0.031	(0.06, 0.18)
pc1	−9.015	3.491	(−15.85, −2.17)	−9.304	4.704	(−18.52, −0.08)	−9.605	2.831	(−15.15, −4.05)
hiPSI	2.381	0.237	(1.91, 2.84)	2.741	0.343	(2.06, 3.41)	1.891	0.214	(1.47, 2.31)
gender	−0.754	0.085	(−0.92, −0.58)	−0.570	0.224	(−1.00, −0.13)	−0.459	0.064	(−0.58, −0.33)
pc2	6.145	3.520	(−0.75, 13.04)	11.315	4.428	(2.63, 19.99)	4.461	2.800	(−1.02, 9.94)
pc3	−0.933	3.618	(−8.02, 6.15)	1.636	4.751	(−7.67, 10.94)	−0.080	2.704	(−5.37, 5.21)
pc4	1.851	3.559	(−5.12, 8.82)	4.060	4.480	(−4.72, 12.84)	−3.074	2.708	(−8.38, 2.23)
pc5	0.64	3.067	(−5.37, 6.65)	0.816	3.928	(−6.88, 8.51)	−0.894	2.719	(−6.22, 4.43)
pc6	−3.478	3.055	(−9.46, 2.50)	−1.807	4.224	(−10.08, 6.47)	−2.737	2.682	(−7.99, 2.51)
pc7	0.726	3.219	(−5.58, 7.03)	−0.674	4.098	(−8.70, 7.35)	−1.695	2.688	(−6.96, 3.57)
pc8	−5.62	3.177	(−11.84, 0.60)	−2.695	4.274	(−11.07, 5.68)	−4.527	2.678	(−9.77, 0.72)
pc9	−2.103	3.513	(−8.98, 4.78)	−5.773	4.757	(−15.09, 3.55)	0.599	2.760	(−4.81, 6.00)
pc10	0.461	3.300	(−6.00, 6.92)	−1.128	4.077	(−9.11, 6.86)	−1.213	2.708	(−6.52, 4.09)

Abbreviations: CI, 95% confidence interval; est, parameter estimation; $\widehat{\text{SD}}$, estimated standard deviations of the corresponding estimation.

7 | CONCLUSION

We propose a nested semiparametric method for analyzing EHR-based case-control studies where the true outcome status of some of the candidate cases are missing. Our method imputes the missing values by introducing an additional index, denoted as noncases, and by modeling the genuine case/noncase pair semiparametrically. The imputation process is very flexible because of the semiparametric structure and the dimension reduction association. Meanwhile, applying the efficient sufficient semiparametric dimension reduction estimator helps to retain stability in odds ratio parameter estimation in the main model even though the missingness scheme is unknown. Many applicable alternative approaches have been developed in the missing data literature if the imputation model had been known or parametric, such as maximum likelihood estimator (MLE) and the fully Bayesian method (Ibrahim et al., 2005; Mitra & Reiter, 2011). However, a prespecified functional form increases the chance of model misspecification (Si & Reiter, 2013), and misspecifications will lead to biased results (Chen & Ibrahim, 2014). In order to improve the robustness, modifications have been made in both MLE and Bayesian methods, such as incorporating a spline into the algorithm to estimate the nonparametric components (Rizopoulos & Ghosh, 2011; Su & Hogan, 2008). The modified MLE and modified Bayesian methods are alternative approaches to our semiparametric imputation approach. They reflect different general approaches in handling missing data in the literature. Although we only considered binary outcomes that are subject to missingness, the flexibility of the semiparametric modeling allows for a straightforward extension to more complex data formats.

ORCID

Ge Zhao https://orcid.org/0000-0002-2875-8652

REFERENCES

Aerts, M., Claeskens, G., Hens, N., & Molenberghs, G. (2002). Local multiple imputation. *Biometrika*, *89*, 375–388.

Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, *26*, 1368–1382.

Chen, Q., & Ibrahim, J. G. (2014). A note on the relationships between multiple imputation, maximum likelihood and fully bayesian methods for missing responses in linear regression models. *Statistics and Its Interface*, *6*, 315.

Haggerty, C. M., Damrauer, S. M., Levin, M. G., Birtwell, D., Carey, D. J., Golden, A. M., Hartzel, D. N., Hu, Y., Judy, R., Kelly, M. A., Kember, R. L., Lester Kirchner, H., Leader, J. B., Liang, L., McDermott-Roe, C., Babu, A., Morley, M., Nealy, Z., Person, T. N., … DiscovEHR and Penn Medicine Biobank Studies. (2019). Genomics-first evaluation of heart disease associated with titin-truncating variants. *Circulation*, *140*, 42–54.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*, 332–346.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data* (Technical Report. ISBN 0471802549, 9780471802549). New York, John Wiley.

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

Ma, Y., & Zhang, X. (2015). A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika*, *102*, 409–420.

Ma, Y., & Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, *107*, 168–179.

Ma, Y., & Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. *Annals of Statistics*, *41*, 250–268.

Mitra, R., & Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine*, *30*, 627–641.

Mukaka, M., White, S. A., Terlouw, D. J., Mwapasa, V., Kalilani-Phiri, L., & Faragher, E. B. (2016). Is using multiple imputation better than complete case analysis for estimating a prevalence (risk) difference in randomized controlled trials when binary outcome observations are missing? *Trials*, *17*, 1–12.

Powell, M. J. D. (1965). A method for minimizing a sum of squares of non-linear functions without calculating derivatives. *The Computer Journal*, *7*, 303–307.

Rizopoulos, D., & Ghosh, P. (2011). A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, *30*, 1366–1380.

Si, Y., & Reiter, J. P. (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, *38*, 499–521.

Su, L., & Hogan, J. W. (2008). Bayesian semiparametric regression for longitudinal binary processes with missing data. *Statistics in Medicine*, *27*, 3247–3268.

Wang, L., Schnall, J., Small, A., Hubbard, R. A., Moore, J. H., Damrauer, S. M., & Chen, J. (2020). Case contamination in electronic health records based case-control studies. *Biometrics*, *77*, 67–77.

Wang, Q., Linton, O., & Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, *99*, 334–345.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhao, G., Ma, Y., Hasler, J. S., Damrauer, S., Levin, M., & Chen, J. (2023). A nested semiparametric method for case-control study with missingness. *Scandinavian Journal of Statistics*, 1–19. https://doi.org/10.1111/sjos.12673