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*

Supplementary Material

We collect in the following (i) discussion on the difference of our problem from two relevant traditional ones, (ii) concrete examples of our general framework, (iii) detailed derivations of semiparametric efficiency of our estimators, (iv) examples to illustrate generalizability of the results in Section 5, (v) supplements to Sections 4–6 and (vi) all technical proofs.

Notations For any random function $\widehat{\mathbf{g}}(\cdot)$ and random vector $\mathbf{U}$ with copies $\{\mathbf{U}_i : i = 1, \ldots, N\}$, let $\mathbb{E}_{\mathbf{U}}\{\widehat{\mathbf{g}}(\mathbf{U})\} \equiv \int \widehat{\mathbf{g}}(\mathbf{u}) dF_{\mathbf{U}}(\mathbf{u})$ represents the integral of $\widehat{\mathbf{g}}(\cdot)$ with respect to the distribution function $F_{\mathbf{U}}(\cdot)$ of $\mathbf{U}$. Also, recall $\mathbb{E}_{\mathbf{U}|S=s}\{\widehat{\mathbf{g}}(\mathbf{U})\} \equiv \int \widehat{\mathbf{g}}(\mathbf{u}) dF_{\mathbf{U}|S=s}(\mathbf{u})$ represents the integral of $\widehat{\mathbf{g}}(\cdot)$ with respect to the conditional distribution function $F_{\mathbf{U}|S=s}(\cdot)$ of $\mathbf{U}$ given $S = s \in \{0, 1\}$. In the following, let

$$\mathbb{E}_{\mathcal{V}}\{\widehat{\mathbf{g}}(\mathbf{U})\} := n^{-1}\textstyle\sum_{i=1}^{n}\widehat{\mathbf{g}}(\mathbf{U}_i),\ \mathbb{G}_{\mathcal{V}}\{\widehat{\mathbf{g}}(\mathbf{U})\} \equiv n^{1/2}(\mathbb{E}_{\mathcal{V}} - \mathbb{E}_{\mathbf{U}|S=1})\{\widehat{\mathbf{g}}(\mathbf{U})\},$$

$$\mathbb{E}_{\mathcal{N}}\{\widehat{\mathbf{g}}(\mathbf{U})\} := (N_1 - n)^{-1}\textstyle\sum_{i=n+1}^{N_1}\widehat{\mathbf{g}}(\mathbf{U}_i),$$

$$\mathbb{G}_{\mathcal{N}}\{\widehat{\mathbf{g}}(\mathbf{U})\} := (N_1 - n)^{1/2}(\mathbb{E}_{\mathcal{N}} - \mathbb{E}_{\mathbf{U}|S=1})\{\widehat{\mathbf{g}}(\mathbf{U})\},$$

$$\mathbb{E}_{\mathcal{C}}\{\widehat{\mathbf{g}}(\mathbf{U})\} := N_0^{-1}\textstyle\sum_{i=N_1+1}^{N}\widehat{\mathbf{g}}(\mathbf{U}_i),\ \mathbb{G}_{\mathcal{C}}\{\widehat{\mathbf{g}}(\mathbf{U})\} := N_0^{1/2}(\mathbb{E}_{\mathcal{C}} - \mathbb{E}_{\mathbf{U}|S=0})\{\widehat{\mathbf{g}}(\mathbf{U})\},$$

$$\mathbb{E}_{\mathcal{V}\cup\mathcal{N}}\{\widehat{\mathbf{g}}(\mathbf{U})\} := N_1^{-1}\textstyle\sum_{i=1}^{N_1}\widehat{\mathbf{g}}(\mathbf{U}_i),\ \mathbb{G}_{\mathcal{V}\cup\mathcal{N}}\{\widehat{\mathbf{g}}(\mathbf{U})\} := N_1^{1/2}(\mathbb{E}_{\mathcal{V}\cup\mathcal{N}} - \mathbb{E}_{\mathbf{U}|S=1})\{\widehat{\mathbf{g}}(\mathbf{U})\},$$

$$\mathbb{E}_{\widetilde{\mathcal{V}}_2}\{\widehat{\mathbf{g}}(\mathbf{U})\} := n_2^{-1}\textstyle\sum_{i=n_1+1}^{n}\widehat{\mathbf{g}}(\mathbf{U}_i),\ \mathbb{G}_{\widetilde{\mathcal{V}}_2}\{\widehat{\mathbf{g}}(\mathbf{U})\} := n_2^{1/2}(\mathbb{E}_{\widetilde{\mathcal{V}}_2} - \mathbb{E}_{\mathbf{U}|S=1})\{\widehat{\mathbf{g}}(\mathbf{U})\},$$

$$\mathbb{E}_{\widetilde{\mathcal{V}}_2\cup\mathcal{N}}\{\widehat{\mathbf{g}}(\mathbf{U})\} := (N_1 - n_1)^{-1}\textstyle\sum_{i=n_1+1}^{N_1}\widehat{\mathbf{g}}(\mathbf{U}_i),$$

$$\mathbb{G}_{\widetilde{\mathcal{V}}_2\cup\mathcal{N}}\{\widehat{\mathbf{g}}(\mathbf{U})\} := (N_1 - n_1)^{1/2}(\mathbb{E}_{\widetilde{\mathcal{V}}_2\cup\mathcal{N}} - \mathbb{E}_{\mathbf{U}|S=1})\{\widehat{\mathbf{g}}(\mathbf{U})\},$$

$$\mathbb{E}_{\mathcal{L}}\{\widehat{\mathbf{g}}(\mathbf{U})\} := n^{-1}\textstyle\sum_{i=1}^{n}\widehat{\mathbf{g}}(\mathbf{U}_i),\ \mathbb{G}_{\mathcal{L}}\{\widehat{\mathbf{g}}(\mathbf{U})\} \equiv n^{1/2}(\mathbb{E}_{\mathcal{L}} - \mathbb{E}_{\mathbf{U}})\{\widehat{\mathbf{g}}(\mathbf{U})\},$$

$$\mathbb{E}_{\mathcal{U}}\{\widehat{\mathbf{g}}(\mathbf{U})\} := (N-n)^{-1}\textstyle\sum_{i=n+1}^{N}\widehat{\mathbf{g}}(\mathbf{U}_i) \text{ and } \mathbb{G}_{\mathcal{U}}\{\widehat{\mathbf{g}}(\mathbf{U})\} \equiv (N-n)^{1/2}(\mathbb{E}_{\mathcal{U}} - \mathbb{E}_{\mathbf{U}})\{\widehat{\mathbf{g}}(\mathbf{U})\}.$$

Also, for $m = 1, \ldots, M$, denote

$$\mathbb{E}_{\mathcal{V}_m}\{\widehat{\mathbf{g}}(\mathbf{U})\} := (n/M)^{-1}\textstyle\sum_{i\in\mathcal{I}_m}\widehat{\mathbf{g}}(\mathbf{U}_i),\ \mathbb{G}_{\mathcal{V}_m}\{\widehat{\mathbf{g}}(\mathbf{U})\} := (n/M)^{1/2}(\mathbb{E}_{\mathcal{V}_m} - \mathbb{E}_{\mathbf{U}|S=1})\{\widehat{\mathbf{g}}(\mathbf{U})\},$$

$$\mathbb{E}_{\mathcal{L}_m}\{\widehat{\mathbf{g}}(\mathbf{U})\} := (n/M)^{-1}\textstyle\sum_{i\in\mathcal{I}_m}\widehat{\mathbf{g}}(\mathbf{U}_i) \text{ and } \mathbb{G}_{\mathcal{L}_m}\{\widehat{\mathbf{g}}(\mathbf{U})\} := (n/M)^{1/2}(\mathbb{E}_{\mathcal{V}_m} - \mathbb{E}_{\mathbf{U}})\{\widehat{\mathbf{g}}(\mathbf{U})\}.$$

S1. Difference of our problem from two relevant traditional ones

We emphasize the ineligibles in the case pool *cannot* be treated as controls but should be *excluded* from the study if known, so the case contamination problem considered in this work is essentially different from (a) the outcome misclassification framework with some cases labeled as controls or the other way around (Lancaster and Imbens, 1996; Beesley and Mukherjee, 2022), and (b) the exponential tilt mixture model where in addition to the (uncontaminated) case and control pools, an unlabeled data set (with unknown status) is also available (Qin, 1999; Zhang and Tan, 2020). In contrast to our setup with ineligibles, individuals in the study samples of both (a) and (b), although with possibly confused or unknown labels, are either cases or controls that are eligible for the analysis–*none* of them should be excluded.

S2. Concrete examples and remarks of the general M-estimation problem (1)

We enumerate below some important special examples of the general M-estimation problem (1) to illustrate its practical relevance.

Example S1 (Primary analysis)**.** Let $\alpha \equiv \tau$ and $\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}) \equiv \mathbf{X}\{S - h(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X})\}$ with $h(x) \equiv \{1 + \exp(-x)\}^{-1}$. Here $\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta})$ is the estimating equation for the logistic regression model of S on $\mathbf{X}$ with a parameter vector $\boldsymbol{\theta}$. Then equation (1) becomes

$$\tau\mathbb{E}[D\mathbf{X}\{1 - h(\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{X})\} \mid S = 1] - (1 - \tau)\mathbb{E}\{\mathbf{X}h(\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{X}) \mid S = 0\} = \mathbf{0}. \quad \text{(S1)}$$

Consider a logistic regression model of D on $\mathbf{X}$ among cases and controls:

$$T(\mathbf{X}) := \mathbb{P}(D = 1 \mid \mathbf{X})/\mathbb{P}(D = 2 \mid \mathbf{X}) = \exp(\overline{\boldsymbol{\theta}}_0^{\mathrm{T}}\mathbf{X}) \quad \text{(S2)}$$

for some $\overline{\boldsymbol{\theta}}_0 \in \mathbb{R}^p$, which is the most fundamental and frequently used model for analyzing the association of the case-control status with a set of covariates based on a case-control sample (Prentice and Pyke, 1979; Breslow, 1996). Recalling the first component of $\mathbf{X}$ equals one, Prentice and Pyke (1979) demonstrated the intercept term $\overline{\boldsymbol{\theta}}_{0[1]}$ is unidentifiable from case-control data unless $\eta \equiv \mathbb{E}(S)$ is known, while $\overline{\boldsymbol{\theta}}_{0[j]} = \boldsymbol{\theta}_{0[j]}$ for $j = 2, \ldots, p$. Hence, the vector $\boldsymbol{\theta}_0$ given in (S1) can be used to calculate the odds ratio between two individuals with covariates $\mathbf{X}_1$ and $\mathbf{X}_2$, i.e., $T(\mathbf{X}_1)/T(\mathbf{X}_2) = \exp\{\sum_{j=2}^{p} \overline{\boldsymbol{\theta}}_{0[j]}(\mathbf{X}_{1[j]} - \mathbf{X}_{2[j]})\} = \exp\{\sum_{j=2}^{p} \boldsymbol{\theta}_{0[j]}(\mathbf{X}_{1[j]} - \mathbf{X}_{2[j]})\}$. Investigations of associations between the primary outcome D and covariates $\mathbf{X}$ through estimating the *odds ratio parameters* $\boldsymbol{\theta}_0$ are referred to as the "primary analysis" in the case-control literature.

Example S2 (Secondary analysis)**.** When $\alpha \equiv \eta$ and $\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}) \equiv \mathbf{X}\{Y - f(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X})\}$ for some known "inverse link" function $f(\cdot)$, the vector $\boldsymbol{\theta}_0$ solves the equation

$$\begin{aligned} \mathbf{0} &= \eta\, \mathbb{E}[D\mathbf{X}\{Y - f(\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{X})\} \mid S = 1] + (1 - \eta)\mathbb{E}[\mathbf{X}\{Y - f(\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{X})\} \mid S = 0] \\ &= \mathbb{E}[\mathbb{1}(D \neq 0)\mathbf{X}\{Y - f(\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{X})\}]. \qquad \text{(S3)} \end{aligned}$$

From (S3), we can see $\boldsymbol{\theta}_0$ is in fact the *generalized linear model parameters* of Y on $\mathbf{X}$ among individuals with $D \neq 0$, corresponding to the following model of Y and $\mathbf{X}$:

$$\mathbb{E}(Y \mid \mathbf{X}, D \neq 0) = f(\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{X}). \qquad \text{(S4)}$$

Special cases include the linear and logistic regression models with $f(x) \equiv x$ and $f(x) \equiv \{1 + \exp(-x)\}^{-1}$, respectively. If we set $f(x) \equiv x$, $p = 1$ and $\mathbf{X} \equiv 1$, then equation (S3) implies $\boldsymbol{\theta}_0 \equiv \mathbb{E}(Y \mid D \neq 0)$, which is the mean outcome in the population without ineligibles. Thus, the vector $\boldsymbol{\theta}_0$ in (S3) is quite a fundamental and important quantity in the so-called "secondary analysis" that studies how the covariates $\mathbf{X}$ affect the secondary outcome Y.

Remark S1 (Existence and uniqueness of $\boldsymbol{\theta}_0$)**.** By rewriting (1) as

$$\mathbb{E}\{g(S, D)\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} = \mathbf{0}$$

with $g(S, D) := \{\alpha\, S/\eta + (1 - \alpha)(1 - S)/(1 - \eta)\}\mathbb{1}(D \neq 0) \geq 0$, we notice the equations in Examples S1–S2 with $\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}) \equiv \mathbf{X}\{S - h(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X})\}$ or $\mathbf{X}\{Y - f(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X})\}$ have similar forms to those considered in Tian et al. (2007). The theory in Section 2 and Appendix 1 of Tian et al. (2007) guarantees the existence and uniqueness of $\boldsymbol{\theta}_0$ in Examples S1–S2 under mild conditions such as positive definiteness of $\mathbb{E}\{g(S, D)\mathbf{X}\mathbf{X}^{\mathrm{T}}\}$ and finiteness of $\mathbb{E}(Y)$.

Remark S2 (Model-free nature of our setup)**.** Although we introduce (working) models (S2) and (S4) to illustrate the practical relevance of $\boldsymbol{\theta}_0$ in the primary and secondary analyses, the validity of relationship (S2) and (S4) is actually *not* required for any results in this article. In fact, even without any model assumption, our target parameters $\boldsymbol{\theta}_0$ are still *well-defined* by the general M-estimation problem (1), as well as the two special cases (S1) and (S3). This model-free setup allows for the unified theoretical analysis in Sections 3 and 5 without adaptation for the model assumptions in each special example of our general framework, as well as for direct applications of our method to other parameters defined by estimating equations in the form of (1).

Remark S3 (Identifiability of $\boldsymbol{\theta}_0$ in the secondary analysis)**.** To estimate the generalized linear model parameters $\boldsymbol{\theta}_0$ defined in (S3), we need the population prevalence $\eta \equiv \mathbb{E}(S)$ to be known. This is satisfied, for example, when the case–control sample is nested within a well-defined cohort study (Tchetgen Tchetgen, 2014), and is also assumed in Wei et al. (2013), Tchetgen Tchetgen (2014) and Wei et al. (2016), among others. Various alternative assumptions are common in the secondary analysis literature as well, e.g., rare disease approximations (Jiang et al., 2006; Wei et al., 2013) and parametric forms of the model between Y and $\mathbf{X}$ (Ma and Carroll, 2016; Liang et al., 2018). Due to retrospective sampling, consistent estimation of parameters defined with respect to the true population distribution is generally impossible in the secondary analysis of case-control data, if none of the above-mentioned assumptions is made (Ma and Carroll, 2016). Considering the model-free nature of our setup as clarified in Remark S2, we thus view the knowledge of the population prevalence η as the bottom line for the identifiability of $\boldsymbol{\theta}_0$ defined in (S3), and avoid other requirements on the underlying data generating mechanism that may have negative impact on the generality and uniformity of our framework.

S3. Semiparametric efficiency of $\widehat{\boldsymbol{\theta}}$

Theorem 1 considers behaviors of $\widehat{\boldsymbol{\theta}}$ in the general setting with $\delta \equiv \lim_{n\to\infty}(n/N_1) \geq 0$. Since the two cases $\delta > 0$ and $\delta = 0$ correspond to different semiparametric models in semiparametric efficiency analysis, we first state in Corollary S1 the asymptotic properties of $\widehat{\boldsymbol{\theta}}$ when $\delta = 0$.

Corollary S1. *Suppose $\delta = 0$. Under the conditions in Theorem 1, we have*

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \boldsymbol{\Omega}[\alpha\, n^{-1}\textstyle\sum_{i=1}^{n}\{D_i - \mu^*(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0) + \alpha\,\mathbb{E}\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1\} + (1-\alpha)\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 0\}] + o_p(n^{-1/2}) \quad \textit{and} \tag{S5}$$

$$n^{1/2}(\boldsymbol{\Omega}\mathbf{A}_0\boldsymbol{\Omega})^{-1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{I}) \quad \textit{as} \quad n \to \infty, \tag{S6}$$

where $\mathbf{A}_0 := \alpha^2 \mathrm{cov}[\{D - \mu^*(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1]$.

We now corroborate if the phenotyping model is correctly specified, i.e., $\mu^*(\cdot) = \mu(\cdot)$, our estimators $\widehat{\boldsymbol{\theta}}$ in (6) achieve under appropriate semipara-

---

metric models the *semiparametric efficiency* defined in Tsiatis (2007), when $\delta \equiv \lim_{n\to\infty}(n/N_1)$ is either positive or zero. We will analyze two cases separately: (a) $\delta > 0$ and (b) $\delta = 0$. The proofs of the following claims (S8), (S11), (S14) and (S15) can be found in Section S8.

Case (a) with $\delta > 0$ We introduce a *nonrandom* indicator R_i that represents whether an individual has been validated ($R_i = 1$) or not ($R_i = 0$), i.e., $R_i := \mathbb{1}(i \in \{1, \ldots, n\} \cup \{N_1 + 1, \ldots, N\})$ for $i = 1, \ldots, N$. Then our study sample can be written as $\{\mathbf{V}_i := (R_i D_i, R_i, S_i, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}} : i = 1, \ldots, N\}$. Given

$$\delta > 0,\ \mu^*(\cdot) = \mu(\cdot) \text{ and the conditions in Theorem 1 hold,} \tag{S7}$$

equation (10) implies the following expansion:

$$\begin{aligned}
&\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = N^{-1}{\textstyle\sum_{i=1}^N}\boldsymbol{\varphi}(\mathbf{V}_i) + o_p(n^{-1/2}),\ \text{where } \boldsymbol{\Omega}^{-1}\boldsymbol{\varphi}(\mathbf{V}_i) := \qquad \text{(S8)}\\
&[\alpha(\delta\tau)^{-1}R_iS_i\{D_i - \mu(\mathbf{Z}_i)\} + \alpha\,\tau^{-1}S_i\mu(\mathbf{Z}_i) + (1-\alpha)(1-\tau)^{-1}(1-S_i)]\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0) - \\
&\alpha\,\tau^{-1}S_i\,\mathbb{E}\{\mu(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1\} - (1-\alpha)(1-\tau)^{-1}(1-S_i)\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 0\}
\end{aligned}$$

with $\mathbf{W}_i \equiv (S_i, Y_i, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}$. Since the values of S_i and R_i ($i = 1, \ldots, N$) are deterministic, the observations $\{\mathbf{V}_i : i = 1, \ldots, N\}$ are independent but *not* identically distributed. However, the theory from Section 2 of Ma (2010) established asymptotic equivalence between the case-control sample and an independent and identically distributed one from a hypothetical population with the prevalence of candidate cases equal to N_1/N, allowing us to view $\{\mathbf{V}_i : i = 1, \ldots, N\}$ as N independent copies of a base observation $\mathbf{V} := (RD, R, \widetilde{S}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$, where $R, \widetilde{S} \in \{0, 1\}$ satisfy

$$\mathbb{E}(\widetilde{S}) = \tau,\ P_{D,\mathbf{Z}|\widetilde{S}} = P_{D,\mathbf{Z}|S},\ R \perp\!\!\!\perp (D, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}} \mid \widetilde{S} \ \text{ and } \ \mathbb{E}(R \mid \widetilde{S}) = \widetilde{S}\delta + (1 - \widetilde{S}) \tag{S9}$$

with $\tau, \delta \in (0, 1)$. It is possible that $\mathbb{E}(\widetilde{S}) \neq \mathbb{E}(S) \equiv \eta$ due to the case-control sampling. That is, $\widetilde{S}$ and the original S may follow different marginal distributions, but the conditional distributions $P_{D,\mathbf{Z}|S}$ and $P_{D,\mathbf{Z}|\widetilde{S}}$ are the same. Under the semiparametric model

$$\mathcal{M} := \{P_{D,R,\widetilde{S},\mathbf{Z}} : \text{(S9) is satisfied while } P_{D,\mathbf{Z}|\widetilde{S}} \text{ is unrestricted}\}, \tag{S10}$$

we show in Section S8 that the efficient influence function $\boldsymbol{\varphi}_{\text{EFF}}(\mathbf{V})$ for estimating $\boldsymbol{\theta}_0$ is

$$\boldsymbol{\varphi}_{\text{EFF}}(\mathbf{V}) = \boldsymbol{\varphi}(\mathbf{V}). \tag{S11}$$

Then, according to the arguments in Section 2 of Ma (2010), all first-order asymptotic results established for N independent and identically distributed observations of $\mathbf{V}$ hold for the original case-control sample as well. By treating $\{\mathbf{V}_i : i = 1, \ldots, N\}$ as N independent copies of $\mathbf{V}$, we notice from expansion (S8) if condition (S7) holds, our estimators $\widehat{\boldsymbol{\theta}}$ attain the efficient influence function $\boldsymbol{\varphi}_{\text{EFF}}(\mathbf{V}) \equiv \boldsymbol{\varphi}(\mathbf{V})$ and are (locally) *semiparametric efficient* (Tsiatis, 2007, Chapter 4) for estimating $\boldsymbol{\theta}_0$ under semiparametric model $\mathcal{M}$ given in (S10).

Case (b) with $\delta = 0$ Consider the semiparametric model

$$\mathcal{M}_0 := \{P_{D,S,\mathbf{Z}} : P_{S,\mathbf{Z}} \text{ is known while } P_{D|\mathbf{Z},S=1} \text{ is unrestricted}\}. \quad \text{(S12)}$$

Since $S \equiv \mathbb{1}(D \neq 2)$, the distribution $P_{D|S=0,\mathbf{Z}}$ is degenerate, given by $\mathbb{P}(D = d \mid S = 0, \mathbf{Z}) = \mathbb{1}(d = 2)$. The only unknown component in $\mathcal{M}_0$ is $P_{D|\mathbf{Z},S=1}$. Given

$$\mu^*(\cdot) = \mu(\cdot) \text{ and the conditions (including } \delta = 0\text{) in Corollary S1 hold,} \quad \text{(S13)}$$

we know from (S5) that

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = n^{-1}\textstyle\sum_{i=1}^{n}\boldsymbol{\varphi}_0(D_i, \mathbf{Z}_i) + o_p(n^{-1/2}), \text{ where} \quad \text{(S14)}$$
$$\boldsymbol{\varphi}_0(D_i, \mathbf{Z}_i) := \alpha\{D_i - \mu(\mathbf{Z}_i)\}\boldsymbol{\Omega}\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0) \equiv \alpha\{D_i - \mu(\mathbf{Z}_i)\}\boldsymbol{\Omega}\boldsymbol{\psi}\{(1, Y_i, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}, \boldsymbol{\theta}_0\}$$

with $(Y_i, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}$ a subvector of $\mathbf{Z}_i$. Here we write $\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0) \equiv \boldsymbol{\psi}\{(1, Y_i, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}, \boldsymbol{\theta}_0\}$ since $\mathbf{W}_i \equiv (S_i, Y_i, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}$ and $S_i \equiv 1$ for $i = 1, \ldots, n$. In Section S8, we prove that under $\mathcal{M}_0$ in (S12), the efficient influence function $\boldsymbol{\varphi}_{\text{EFF}}^{(0)}(D, \mathbf{Z})$ for estimating $\boldsymbol{\theta}_0$ satisfies

$$\boldsymbol{\varphi}_{\text{EFF}}^{(0)}(D, \mathbf{Z}) = \boldsymbol{\varphi}_0(D, \mathbf{Z}) \equiv \alpha\{D - \mu(\mathbf{Z})\}\boldsymbol{\Omega}\boldsymbol{\psi}\{(1, Y, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}, \boldsymbol{\theta}_0\}. \quad \text{(S15)}$$

Noticing $\{(D_i, \mathbf{Z}_i) : i = 1, \ldots, n\}$ are independent and identically distributed observations from distribution $P_{D,\mathbf{Z}|S=1}$, we know from (S14) that under condition (S13), our estimators $\widehat{\boldsymbol{\theta}}$ enjoy (local) *semiparametric efficiency* (Tsiatis, 2007, Chapter 4) for estimating $\boldsymbol{\theta}_0$ under semiparametric model $\mathcal{M}_0$ given in (S12).

S4. Illustrations of generalizability of the results in Section 5

We consider in this section (i) semi-supervised mean response estimation and (ii) average treatment effect estimation in randomized experiments, establishing for estimators in these problems results similar to Theorem 2, which are entirely free of convergence assumptions on nuisance estimation. These problems, as well as the case contamination one in the main article, share a common feature: there exist some simple $n^{1/2}$-consistent estimators without nuisance functions (e.g., supervised/inverse-probability-weighted estimators and $\widehat{\boldsymbol{\theta}}_{\text{IN}}$ in (15)), while advanced methods involve nuisance estimation to improve efficiency. It is hence desirable to reduce the reliance on assumptions of the nuisance estimators so that possible performance degradation arising from violation of these assumptions can be avoided, especially when the nuisance estimators are based on some high dimensional or black-box models whose limiting behaviors can be hard to specify. As elaborated in Section 5.1 of Zhang et al. (2019), problem (ii) is actually a two-arm version of (i) and approaches to (i) can be directly used for (ii), so the following discussion will focus on (i) while all the methods and results in this section apply to (ii) as well.

Suppose our study sample consists of two independent data sets: (i) a labeled data set $\mathcal{L} := \{(Y_i, \mathbf{X}_i) : i = 1, \dots, n\}$ and (ii) an unlabeled data set $\mathcal{U} := \{\mathbf{X}_i : i = n+1, \dots, N\}$, which contain n and $(N-n)$ independent copies of base observations $(Y, \mathbf{X})$ and $\mathbf{X}$, respectively. Here $Y \in \mathbb{R}$ is a response while $\mathbf{X} \in \mathbb{R}^d$ represents a set of possibly high dimensional covariates. This is the so-called "semi-supervised" setting. Our target parameter is $\theta_0 := \mathbb{E}(Y)$. Estimation of θ_0 in such a setting has been investigated by Bloniarz et al. (2016), Wager et al. (2016), Zhang et al. (2019), Zhang and Bradic (2022) in the context of semi-supervised inference or causal inference in randomized experiments. We relegate comparison of our results with those in the existing literature to Remark S4.

It is easy to see the supervised estimator $\overline{Y} := n^{-1}\sum_{i=1}^{n} Y_i$ enjoys $n^{1/2}$-consistency and asymptotic normality as long as $\mathbb{E}(Y^2) = O(1)$, but $\overline{Y}$ is generally suboptimal when unlabeled data are available due to efficiency loss caused by ignoring $\mathcal{U}$. We hence attempt to devise semi-supervised estimators, which take account of both $\mathcal{L}$ and $\mathcal{U}$, based on the following identity:

$$\mathbb{E}(Y) = \mathbb{E}\{\phi(\mathbf{X})\} = \mathbb{E}\{\phi(\mathbf{X})\} + \mathbb{E}\{Y - \phi(\mathbf{X})\} = \mathbb{E}\{\phi^*(\mathbf{X})\} + \mathbb{E}\{Y - \phi^*(\mathbf{X})\} \quad \text{(S16)}$$

with $\phi(\mathbf{x}) := \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$ and $\phi^* : \mathbb{R}^d \mapsto \mathbb{R}$ an arbitrary function. Then our semi-supervised estimator is the empirical version (S16), that is,

$$\widehat{\theta} := N^{-1}\textstyle\sum_{i=1}^{N}\widehat{\phi}(\mathbf{X}_i) + n^{-1}\sum_{i=1}^{n}\{Y_i - \widehat{\phi}(\mathbf{X}_i)\}, \quad \text{(S17)}$$

where $\widehat{\phi}(\cdot)$ is a random function involving $\mathcal{L}$ only. Term $n^{-1}\sum_{i=1}^{n}\{Y_i - \widehat{\phi}(\mathbf{X}_i)\}$ in (S17) plays the same debiasing role as $\alpha\, n^{-1}\sum_{i=1}^{n}\{D_i - \widehat{\mu}(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i, \widehat{\boldsymbol{\theta}}_{\text{IN}})$ in (6); see the discussion concerning usefulness of such debiasing terms at the end of Section 3.1. As in (7), we apply cross-fitting to calculate $\widehat{\phi}(\mathbf{X}_i)$: 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 $\widehat{\phi}_m(\cdot)$ be a random function based on $\mathcal{L}_m^- := \{(Y_i, \mathbf{X}_i) : i \in \mathcal{I}\backslash\mathcal{I}_m\}$ $(m = 1, \ldots, M)$. Then we set

$$\widehat{\phi}(\mathbf{X}_i) \;\equiv\; \textstyle\sum_{m=1}^{M}\{\mathbb{1}(i \in \mathcal{I}_m)\widehat{\phi}_m(\mathbf{X}_i) + \mathbb{1}(i > n)\widehat{\phi}_m(\mathbf{X}_i)/M\} \quad (i = 1, \ldots, N). \quad \text{(S18)}$$

Also, we propose a sample-splitting variant of $\widehat{\theta}$: divide $\mathcal{L}$ into two disjoint subsets: $\widetilde{\mathcal{L}}_1 := \{(Y_i, \mathbf{X}_i) : 1 \leq i \leq n_1\}$ and $\widetilde{\mathcal{L}}_2 := \{(Y_i, \mathbf{X}_i) : n_1 < i \leq n\}$ of sizes n_1 and $n_2 := n - n_1$. Let $\widetilde{\phi}_1(\cdot)$ be a random function involving $\widetilde{\mathcal{L}}_1$

only. The sample-splitting variant $\widehat{\theta}$ is

$$\widetilde{\theta} := (N - n_1)^{-1}\textstyle\sum_{i=n_1+1}^{N}\widetilde{\phi}_1(\mathbf{X}_i) + n_2^{-1}\textstyle\sum_{i=n_1+1}^{n}\{Y_i - \widetilde{\phi}_1(\mathbf{X}_i)\}. \tag{S19}$$

We now explain why $\widehat{\theta}$ is $n^{1/2}$-consistent and $\widetilde{\theta}$ is asymptotically normal even if no assumption is imposed on the convergence behaviors of $\{\widehat{\phi}_m(\cdot), \widetilde{\phi}_1(\cdot)\}$. With $\widehat{\phi}(\mathbf{X}_i)$ as in (S18), we can write $\widehat{\theta} - \theta_0 \equiv T_1 + (1 - n/N)\widehat{T}_2$, where

$$\begin{aligned}
T_1 &:= n^{-1}\textstyle\sum_{i=1}^{n}(Y_i - \theta_0) \quad \text{and} \\
\widehat{T}_2 &:= (N-n)^{-1}\textstyle\sum_{i=n+1}^{N}\widehat{\phi}(\mathbf{X}_i) - n^{-1}\textstyle\sum_{i=1}^{n}\widehat{\phi}(\mathbf{X}_i) \\
&= M^{-1}\textstyle\sum_{m=1}^{M}\{(N-n)^{-1}\textstyle\sum_{i=n+1}^{N}\widehat{\phi}_m(\mathbf{X}_i) - (n/M)^{-1}\textstyle\sum_{i\in\mathcal{I}_m}\widehat{\phi}(\mathbf{X}_i)\} \\
&= M^{-1}\textstyle\sum_{m=1}^{M}([(N-n)^{-1}\textstyle\sum_{i=n+1}^{N}\widehat{\phi}_m(\mathbf{X}_i) - \mathbb{E}_{\mathbf{X}}\{\widehat{\phi}_m(\mathbf{X})\}] - \\
&\qquad [(n/M)^{-1}\textstyle\sum_{i\in\mathcal{I}_m}\widehat{\phi}_m(\mathbf{X}_i) - \mathbb{E}_{\mathbf{X}}\{\widehat{\phi}_m(\mathbf{X})\}])
\end{aligned}$$

It is straightforward to show $T_1 = O_p(n^{-1/2})$. Moreover, notice that

$$\{\widehat{\phi}_m(\mathbf{X}_i) : i \in \mathcal{I}_m \cup \{n+1, \ldots, N\}\}$$

are conditionally independent and identically distributed given $\mathcal{L}_m^-$, since only $\mathcal{L}_m^-$ is used to calculate $\widehat{\phi}_m(\cdot)$. Therefore, conditionally on $\mathcal{L}_m^-$, we can treat $\widehat{\phi}_m(\cdot)$ as nonrandom and view

$$[(N-n)^{-1}\textstyle\sum_{i=n+1}^{N}\widehat{\phi}_m(\mathbf{X}_i) - \mathbb{E}_{\mathbf{X}}\{\widehat{\phi}_m(\mathbf{X})\}] \text{ and } [(n/M)^{-1}\textstyle\sum_{i\in\mathcal{I}_m}\widehat{\phi}_m(\mathbf{X}_i) - \mathbb{E}_{\mathbf{X}}\{\widehat{\phi}_m(\mathbf{X})\}]$$

as two terms of the form "sample mean minus population mean", which are of order $O_p(n^{-1/2})$ given $\mathbb{E}_{\mathbf{X}}[\{\widehat{\phi}_m(\mathbf{X})\}^2] = O_p(1)$ according to Chebyshev's inequality. The unconditional $n^{1/2}$-convergences follow. Considering M is fixed, we know that $\widehat{T}_2 = O_p(n^{-1/2})$ and thereby that

$$\widehat{\theta} - \theta_0 \equiv T_1 + \widehat{T}_2 = O_p(n^{-1/2}).$$

Concerning the data-splitting variant $\widetilde{\theta}$, we can see from (S19) that

$$\mathbb{E}(\widetilde{\theta} - \theta_0 \mid \widetilde{\mathcal{L}}_1) = \mathbb{E}_{\mathbf{X}}\{\widetilde{\phi}_1(\mathbf{X})\} + \mathbb{E}(Y - \theta_0) - \mathbb{E}_{\mathbf{X}}\{\widetilde{\phi}_1(\mathbf{X})\} = 0, \tag{S20}$$

because $\widetilde{\mathcal{L}}_1 \perp\!\!\!\perp (\widetilde{\mathcal{L}}_2 \cup \mathcal{U})$. In addition, notice that $\widetilde{\theta} - \theta_0 \equiv$

$$\textstyle\sum_{i=n_1+1}^{n}[n_2^{-1}\{Y_i - \widetilde{\phi}_1(\mathbf{X}_i)\} + (N-n_1)^{-1}\widetilde{\phi}_1(\mathbf{X}_i)] + \sum_{i=n+1}^{N}(N-n_1)^{-1}\widetilde{\phi}_1(\mathbf{X}_i), \tag{S21}$$

where the summands

$$\{n_2^{-1}\{Y_i - \widetilde{\phi}_1(\mathbf{X}_i)\} + (N-n_1)^{-1}\widetilde{\phi}_1(\mathbf{X}_i) : i = n_1+1,\ldots,n\} \cup$$
$$\{(N-n_1)^{-1}\widetilde{\phi}_1(\mathbf{X}_i) : i = n+1,\ldots,N\}$$

are conditionally independent given $\widetilde{\mathcal{L}}_1$ since $\widetilde{\phi}_1(\cdot)$ involves $\widetilde{\mathcal{L}}_1$ only. Combining (S20) and (S21), we can show the conditional asymptotic normality of $(\widetilde{\theta} - \theta_0)$ given $\widetilde{\mathcal{L}}_1$ under some suitable moment conditions, using the Lyapunov central limit theorem. The unconditional asymptotic normality follows. In the theorem below, we formally state the above-mentioned properties of $\{\widehat{\theta}, \widetilde{\theta}\}$ without imposing assumptions on convergence behaviors of $\{\widehat{\phi}_m(\cdot), \widetilde{\phi}_1(\cdot)\}$. The proof can be found in Section S8.

Theorem S1. *Suppose* $\mathbb{E}(Y^2) = O(1)$. *If* $\mathbb{E}_{\mathbf{X}}[\{\widehat{\phi}_m(\mathbf{X})\}^2] = O_p(1)$ *for* $m = 1, \ldots, M$, *then* $\widehat{\theta}$ *in* (S17) *satisfies* $\widehat{\theta} - \theta_0 = O_p(n^{-1/2})$. *Further, with* $\widetilde{\nu}_n := n_2/(N - n_1)$, *denote*

$$\widetilde{\sigma}_n := [\mathrm{var}\{Y - (1 - \widetilde{\nu}_n)\widetilde{\phi}_1(\mathbf{X}) \mid \widetilde{\mathcal{L}}_1\} + \widetilde{\nu}_n(1 - \widetilde{\nu}_n)\mathrm{var}\{\widetilde{\phi}_1(\mathbf{X}) \mid \widetilde{\mathcal{L}}_1\}]^{1/2}.$$

Then, given $\widetilde{\sigma}_n^{-1} = O_p(1)$ *and* $\mathbb{E}\{|Y|^{2(1+c)}\} + \mathbb{E}_{\mathbf{X}}\{|\widetilde{\phi}_1(\mathbf{X})|^{2(1+c)}\} = O_p(1)$ *for some constant* $c > 0$, *we have* $n_2^{1/2}\, \widetilde{\sigma}_n^{-1/2}(\widetilde{\theta} - \theta_0) \xrightarrow{d} \mathbf{N}(0, 1)$ *as* $n_2 \to \infty$.

The moment conditions $\mathbb{E}_{\mathbf{X}}[\{\widehat{\phi}_m(\mathbf{X})\}^2] = O_p(1)$ and $\mathbb{E}_{\mathbf{X}}\{|\widetilde{\phi}_1(\mathbf{X})|^{2(1+c)}\} = O_p(1)$ are fairly mild, typically holding when, for example, the response Y is bounded and $\{\widehat{\phi}_m(\cdot), \widetilde{\phi}_1(\cdot)\}$ are calculated based on some algorithms targeting $\mathbb{E}(Y \mid \mathbf{X})$. More generally, we can always truncate $\{\widehat{\phi}_m(\cdot), \widetilde{\phi}_1(\cdot)\}$ to make them bounded. Analogously to Remark 4, we can also summarize properties of $\widehat{\theta}$ into three strata according to different assumptions on $\widehat{\phi}_m(\cdot)$:

(i) $n^{1/2}$-consistency whenever $\mathbb{E}_{\mathbf{X}}[\{\widehat{\phi}_m(\mathbf{X})\}^2] = O_p(1)$;

(ii) asymptotic normality if $\widehat{\phi}_m(\cdot)$ converge (in the L_2 sense) to some function $\phi^*(\cdot)$ satisfying $\mathbb{E}\{|\phi^*(\mathbf{X})|^{2(1+c_1)}\} < c_2$ for some positive constants $\{c_1, c_2\}$;

(iii) semiparametric efficiency given the function $\phi^*(\cdot)$ in (ii) equals $\mathbb{E}(Y \mid \mathbf{X} = \cdot)$.

The above (i) is from Theorem S1, while (ii) and (iii) were established in Theorem 5 of Zhang and Bradic (2022).

Remark S4. [Comparison with results in the existing literature] As pointed out at the end of the second paragraph in this section, estimation of $\mathbb{E}(Y)$ when unlabeled data are available in addition to labeled data has been investigated by some existing works in different contexts. Among them, the estimators proposed in Zhang et al. (2019) and Bloniarz et al. (2016) can be viewed as a special example of $\widehat{\theta}$ in (S17) with $\widehat{\phi}(\cdot)$ based on a working linear model of Y on $\mathbf{X}$. Due to not using cross-fitting, their theoretical analysis was conducted under fairly stringent conditions, e.g., that the dimension d of $\mathbf{X}$ (Zhang et al., 2019) or the sparsity of the working linear model (Bloniarz et al., 2016) is of order $o(n^{1/2})$, which were needed to ensure

$$n^{-1}\textstyle\sum_{i=1}^{n}\{\widehat{\phi}(\mathbf{X}_i) - \phi^*(\mathbf{X}_i)\} - \mathbb{E}_{\mathbf{X}}\{\widehat{\phi}(\mathbf{X}) - \phi^*(\mathbf{X})\} = o_p(n^{-1/2})$$

with $\phi^*(\cdot)$ the probability limit of $\widehat{\phi}(\cdot)$. The estimator $\widehat{\theta}$ in (S1) was considered by Wager et al. (2016) and Zhang and Bradic (2022) as well. These two articles adopted the cross-fitting strategy as in (S18), allowing the form of $\widehat{\phi}_m(\cdot)$ to be arbitrary and only requiring its L_2 convergence to some function. All these existing works assumed the nuisance estimator $\widehat{\phi}(\cdot)$ (or

$\widehat{\phi}_m(\cdot)$ if cross-fitting is applied) has a probability limit, which may not be the case when, for example, $d \gg n$ and we are agnostic to any structure information (e.g., sparsity level) of the working model used to construct $\widehat{\phi}(\cdot)$. In contrast, Theorem S1 imposes no condition on the convergence of $\widehat{\phi}_m(\cdot)$, providing theoretical guarantees for the performance of $\widehat{\theta}$ under violation of the above-mentioned assumptions. Even if the probability limit of $\widehat{\phi}_m(\cdot)$ does not exist, the estimator $\widehat{\theta}$ still enjoys $n^{1/2}$-consistency as long as $\mathbb{E}_{\mathbf{X}}[\{\widehat{\phi}_m(\mathbf{X})\}^2] = O_p(1)$.

S5. Supplement to simulations in Section 4

We set function $\rho(\mathbf{Z})$ in the data generating mechanism (17) to the following five different forms with sparsity level $q \in \{\lceil d^{1/2} \rceil, n\}$:

(a) $\rho(\mathbf{Z}) \equiv 0.7$, which yields a *constant* model;

(b) $\rho(\mathbf{Z}) \equiv 5 \sum_{j=1}^{q} \mathbf{Z}_{[j]}/q^{1/2}$, which yields a *linear* model;

(c) $\rho(\mathbf{Z}) \equiv 3 \sum_{j=1}^{q} \mathbf{Z}_{[j]}/q^{1/2} + 3(\sum_{j=1}^{q} \mathbf{Z}_{[j]})^2/(2q)$, which yields a *single-index* model;

(d) $\rho(\mathbf{Z}) \equiv (3 \sum_{j=1}^{q} \mathbf{Z}_{[j]}/q^{1/2})(1 + \sum_{j=1}^{\lceil q/2 \rceil} \mathbf{Z}_{[j]}/q^{1/2}) - 3(\boldsymbol{\kappa}^{\mathrm{T}} \mathbf{Z}_q)^2/q$ with $\boldsymbol{\kappa} := (1, 0, 1, 0, \ldots)^{\mathrm{T}} \in \mathbb{R}^q$ and $\mathbf{Z}_q := (\mathbf{Z}_{[1]}, \ldots, \mathbf{Z}_{[q]})^{\mathrm{T}}$, which yields a *multiple-index* model;

(e) $\rho(\mathbf{Z}) \equiv 3\sum_{j=1}^{q}\{\mathbf{Z}_{[j]}/q^{1/2} + \mathbf{Z}_{[j]}^2/(2q)\}$, which yields an *additive* model.

In Tables S1–S2, we present the simulation results of the secondary analysis in Section 4. Table S3 presents the relative efficiencies of the plug-in estimator $\widehat{\boldsymbol{\theta}}_{\text{PI}}$ in (21) to the benchmark $\widehat{\boldsymbol{\theta}}_{\text{IN}}$.

S6. Supplement to Theorem 2 in Section 5

With $\widetilde{\delta}_n := n_2/(N_1 - n_1)$, matrix $\widetilde{\mathbf{A}}_n$ in Theorem 2 is defined as

$$\begin{aligned}\widetilde{\mathbf{A}}_n := \alpha^2 \text{cov}[\{D - (1-\widetilde{\delta}_n)\widetilde{\mu}_1(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1, \widetilde{\mathcal{V}}_1] + \\ \alpha^2\widetilde{\delta}_n(1-\widetilde{\delta}_n)\text{cov}\{\widetilde{\mu}_1(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1, \widetilde{\mathcal{V}}_1\} + \\ (1-\alpha)^2\tau(1-\tau)^{-1}(n_2/N_1)\text{cov}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 0\}.\end{aligned}$$

S7. Supplement to the data analysis in Section 6

In Table S4 below we list the names, descriptions and summary statistics of the covariates considered in the data analysis of Section 6. Table S5 records the components of the plug-in estimator $\widehat{\boldsymbol{\theta}}_{\text{PI}}$ in (21).

Table S1: Simulation results of the secondary analysis in Section 4: relative efficiencies (19) 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 (17) and MRE as defined in (20). The choices (a)–(e) of $\rho(\mathbf{Z})$ are listed in Section S5. 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)	2.14	2.05	2.15	2.34	2.55	2.41	2.56	2.71	2.84	2.65	2.87	3.03
	(b)	2.58	2.28	1.84	3.81	3.16	2.64	2.13	5.64	3.54	2.97	2.19	8.25
	(c)	2.13	1.96	2.14	3.31	2.61	2.35	2.60	4.33	2.82	2.49	2.83	5.41
	(d)	1.92	1.77	1.87	3.23	2.22	2.01	2.14	4.28	2.40	2.12	2.31	5.42
	(e)	2.69	2.45	2.16	3.76	3.39	2.92	2.58	5.31	3.84	3.24	2.75	7.25
400	(a)	1.75	1.73	1.75	1.88	2.25	2.20	2.25	2.34	2.67	2.58	2.66	2.81
	(b)	2.09	1.99	1.60	2.45	2.96	2.78	1.98	3.81	3.68	3.26	2.18	6.29
	(c)	1.78	1.75	1.75	2.33	2.27	2.20	2.23	3.31	2.60	2.50	2.54	4.63
	(d)	1.68	1.65	1.61	2.27	2.12	2.08	1.99	3.23	2.29	2.21	2.14	4.59
	(e)	2.14	2.10	1.79	2.49	3.13	2.94	2.38	3.76	3.83	3.58	2.62	5.82
$q=n$		$N=5000$				$N=10000$				$N=25000$			
n	$\rho(\mathbf{Z})$	LR	KS	RF	MRE	LR	KS	RF	MRE	LR	KS	RF	MRE
200	(b)	2.24	1.93	2.20	4.33	2.60	2.14	2.61	6.62	3.00	2.37	2.96	10.04
	(c)	2.47	2.28	2.50	3.70	2.87	2.56	2.93	5.07	3.30	2.89	3.38	6.63
	(d)	2.05	1.88	2.09	3.50	2.42	2.12	2.49	4.73	2.54	2.20	2.63	6.12
	(e)	2.66	2.35	2.68	4.28	3.33	2.82	3.36	6.30	3.94	3.33	3.97	9.03
400	(b)	2.06	1.91	2.05	2.85	3.01	2.70	2.93	4.69	3.53	3.03	3.53	8.30
	(c)	2.11	2.01	2.13	2.67	3.03	2.77	3.04	4.10	3.68	3.21	3.72	6.41
	(d)	1.86	1.78	1.85	2.52	2.55	2.36	2.55	3.76	3.01	2.71	3.02	5.66
	(e)	2.38	2.29	2.38	2.88	3.58	3.36	3.54	4.67	4.64	4.18	4.67	8.01

Table S2: Simulation results of the secondary analysis in Section 4: componentwise confidence intervals for $\boldsymbol{\theta}_0$ established based on our estimators $\widehat{\boldsymbol{\theta}}$ given in (6). 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 (17), DCR as defined in (22) and AL stands for "average length". The choices (a)–(e) of $\rho(\mathbf{Z})$ are listed in Section S5.

		$q = \lceil d^{1/2} \rceil$						$q = n$					
		LR		KS		RF		LR		KS		RF	
n	$\rho(\mathbf{Z})$	DCR	AL	DCR	AL	DCR	AL	DCR	AL	DCR	AL	DCR	AL
200	(a)	0.60	0.10	0.55	0.10	0.63	0.10	0.60	0.10	0.55	0.10	0.63	0.10
	(b)	0.93	0.09	1.08	0.10	0.72	0.11	0.70	0.10	0.65	0.11	0.67	0.10
	(c)	0.92	0.10	0.93	0.11	0.85	0.10	0.73	0.10	0.58	0.10	0.68	0.10
	(d)	0.65	0.11	1.00	0.11	0.68	0.11	0.93	0.10	1.02	0.11	0.93	0.10
	(e)	0.83	0.09	0.70	0.10	0.68	0.10	0.78	0.09	0.83	0.10	0.87	0.09
400	(a)	0.70	0.08	0.70	0.08	0.77	0.08	0.70	0.08	0.70	0.08	0.77	0.08
	(b)	1.20	0.08	1.38	0.08	1.23	0.09	0.60	0.08	0.52	0.08	0.75	0.08
	(c)	1.45	0.08	1.17	0.08	1.43	0.08	0.98	0.08	0.78	0.08	0.92	0.08
	(d)	0.87	0.09	0.85	0.09	0.73	0.09	0.70	0.08	0.62	0.08	0.83	0.08
	(e)	1.08	0.08	0.77	0.08	0.98	0.08	0.83	0.07	0.78	0.07	0.88	0.07

Table S3: Simulation results: relative efficiencies of the plug-in estimator $\widehat{\boldsymbol{\theta}}_{\text{PI}}$ in (21) to the benchmark $\widehat{\boldsymbol{\theta}}_{\text{IN}}$. The sample size is $N = 5000$. The validation set size is $n = 200$. The nuisance estimator $\widehat{\mu}(\cdot)$ 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 $\mathbf{Z}$, $\rho(\mathbf{Z})$ the function in data generating model (17). The choices (a)–(e) of $\rho(\mathbf{Z})$ are listed in Section S5. 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.

		Primary Analysis			Secondary Analysis		
q	$\rho(\mathbf{Z})$	LR	KS	RF	LR	KS	RF
$\lceil d^{1/2} \rceil$	(a)	1.41	1.35	1.21	1.36	1.32	1.20
	(b)	1.09	0.97	1.25	1.14	1.00	1.23
	(c)	1.27	1.02	1.23	1.32	1.19	1.19
	(d)	1.36	0.99	1.32	1.38	1.19	1.24
	(e)	1.10	0.93	1.23	1.17	1.02	1.19
n	(b)	1.27	1.06	1.23	1.24	1.07	1.19
	(c)	1.27	1.13	1.17	1.24	1.15	1.15
	(d)	1.41	1.20	1.24	1.33	1.20	1.21
	(e)	1.21	1.06	1.15	1.17	1.06	1.13

Table S4: Names, descriptions and summary statistics of the covariates considered in the data analysis of Section 6. Here "SD" stands for "standard deviation".

Name	**Description**	**Mean**	**SD**
age	Age (year)	67.1	17.0
aniongap_max	Maximum anion gap (mmol/L)	18.0	5.5
bun_mean	Mean level of blood urea nitrogen (mmol/L)	36.2	26.3
cretinine_min	Minimum cretinine concentration in blood (mmol/L)	1.6	1.4
inr_max	Maximum international normalized ratio	1.9	1.6
lactate_min	Minimum lactate concentration in blood (mmol/L)	1.9	1.5
metastatic_cancer	Having metastatic cancer (1) or not (0)	0.1	0.3
sodium_max	Maximum sodium concentration in blood (mmol/L)	140.7	6.3
spo2_mean	Mean fraction of oxygen-saturated hemoglobin relative to total hemoglobin (unsaturated+saturated) in the blood	96.7	2.9
sysbp_min	Minimum systolic blood pressure (mmHg)	83.2	16.8
urineoutput	Urine output (ml/kg/hr)	1666.7	1419.8

Table S5: Results of the real data analysis in Section 6: components of the plug-in estimator $\widehat{\boldsymbol{\theta}}_{\text{PI}}$ in (21). The nuisance estimator $\widehat{\mu}(\cdot)$ is constructed using logistic regression (LR), kernel smoothing (KS) or random forest (RF).

	LR	KS	RF
age	-0.65	-0.78	-0.65
aniongap_max	0.53	0.52	0.55
bun_mean	0.39	0.47	0.34
cretinine_min	-0.28	-0.28	-0.30
inr_max	0.42	0.49	0.40
lactate_min	0.01	-0.03	0.01
metastatic_cancer	0.07	0.04	0.10
sodium_max	-0.40	-0.44	-0.36
spo2_mean	-0.26	-0.29	-0.24
sysbp_min	0.07	0.08	0.06
urineoutput	1.24	1.29	1.21

S8. An auxiliary lemma and technical proofs

Lemma S1. *Let $\widehat{\mathbf{M}} \in \mathbb{R}^{p\times p}$ be an estimator for a matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$. Suppose that $\mathbf{M}$ and $\widehat{\mathbf{M}}$ are both invertible, and that $\|\widehat{\mathbf{M}} - \mathbf{M}\| = o_p(1)$. Then $\|\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1}\| = O_p(\|\widehat{\mathbf{M}} - \mathbf{M}\|)$.*

Proof of Lemma S1: Since $\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1} = \widehat{\mathbf{M}}^{-1}(\mathbf{M} - \widehat{\mathbf{M}})\mathbf{M}^{-1}$, we know

$$
\begin{aligned}
\|\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1}\| &\leq \|\widehat{\mathbf{M}}^{-1}\|\|\widehat{\mathbf{M}} - \mathbf{M}\|\|\mathbf{M}^{-1}\| \\
&\leq \|\mathbf{M}^{-1}\|(\|\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1}\| + \|\mathbf{M}^{-1}\|)\|\widehat{\mathbf{M}} - \mathbf{M}\|
\end{aligned}
$$

$$\leq \ c_1(\|\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1}\| + c_2)\|\widehat{\mathbf{M}} - \mathbf{M}\|,$$

which implies

$$(1 - c_1\|\widehat{\mathbf{M}} - \mathbf{M}\|)\|\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1}\| \leq c_2\|\widehat{\mathbf{M}} - \mathbf{M}\|. \tag{S22}$$

Considering the assumption that $\|\widehat{\mathbf{M}} - \mathbf{M}\| = o_p(1)$, we have

$$\mathbb{P}(c_1 \|\widehat{\mathbf{M}} - \mathbf{M}\| \leq 1/2) \to 1. \tag{S23}$$

Further, we can obtain from (S22) that on the event $\{c_1 \|\widehat{\mathbf{M}} - \mathbf{M}\| \leq 1/2\}$,

$$\|\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1}\| \leq (1 - c_1\|\widehat{\mathbf{M}} - \mathbf{M}\|)^{-1}c_2\|\widehat{\mathbf{M}} - \mathbf{M}\| \leq c\,\|\widehat{\mathbf{M}} - \mathbf{M}\|,$$

which combined with (S23) gives $\|\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1}\| = O_p(\|\widehat{\mathbf{M}} - \mathbf{M}\|)$.

<u>Proof of Theorem 1</u>: In the following proof, the fact that $\mathbb{P}(\widehat{\boldsymbol{\theta}}_{\text{IN}} \in \mathcal{B}_0) \to 1$ from Assumption 2 will be used implicitly here and there. Write

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \equiv \widehat{\mathbf{T}}_1 + \widehat{\boldsymbol{\Omega}}(\widehat{\mathbf{T}}_2 + \widehat{\mathbf{T}}_3), \text{ where} \tag{S24}$$

$$\widehat{\mathbf{T}}_1 := \widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0 + \widehat{\boldsymbol{\Omega}}[\alpha\,\mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\}],$$

$$\widehat{\mathbf{T}}_2 := \alpha(1 - \delta_n)(\mathbb{E}_{\mathcal{N}} - \mathbb{E}_{\mathcal{V}})[\{\widehat{\mu}(\mathbf{Z}) - \mu^*(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})] \text{ and}$$

$$\widehat{\mathbf{T}}_3 := \alpha(\mathbb{E}_{\mathcal{V}\cup\mathcal{N}} - \mathbb{E}_{\mathcal{V}})\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\}.$$

We first control $\widehat{\mathbf{T}}_1$. According to Assumption 1, we have from Taylor's expansion that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{B}_0$,

$$|D\{\boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta}_1) - \boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta}_2)\}| \leq \sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \boldsymbol{\theta})\|\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \tag{S25}$$

with $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\boldsymbol{\psi}'_{[j]}(\mathbf{W},\boldsymbol{\theta})\|^2 \mid S=1\} < \infty$ $(j=1,\ldots,p)$. Hence, it follows from Example 19.7 of Van der Vaart (2000) that $\{D\boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}) : \boldsymbol{\theta}\in\mathcal{B}_0\}$ is $P_{\mathbf{W}|S=1}$-Donsker. Similarly, we can show $\{\boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}) : \boldsymbol{\theta}\in\mathcal{B}_0\}$ is $P_{\mathbf{W}|S=0}$-Donsker. Moreover, applying Taylor's expansion yields for $j = 1,\ldots,p$ that

$$\begin{aligned}
&\mathbb{E}_{D,\mathbf{W}|S=1}[D^2\{\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\}^2] \\
&\le \|\widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0\|^2\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\boldsymbol{\psi}'_{[j]}(\mathbf{W},\boldsymbol{\theta})\|^2 \mid S=1\} = O_p(u_n^2) = o_p(1) \text{ and}
\end{aligned} \tag{S26}$$

$$\mathbb{E}_{\mathbf{W}|S=0}[\{\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\}^2] = o_p(1).$$

Here we use Assumptions 1 and 2. The above derivations have verified the conditions of Lemma 19.24 in Van der Vaart (2000), which ensures

$$\mathbb{G}_{\mathcal{V}}[D\{\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\}] = o_p(1) \quad \text{and} \quad \mathbb{G}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\} = o_p(1).$$

Therefore, we have

$$\begin{aligned}
&\alpha\,\mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\} \\
&= \alpha\,\mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\} + \\
&\alpha\,\mathbb{E}_{D,\mathbf{W}|S=1}\{D\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\} + (1-\alpha)\mathbb{E}_{\mathbf{W}|S=0}\{\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\} + o_p(n^{-1/2}).
\end{aligned} \tag{S27}$$

Using Taylor's expansion again, we obtain

$$\alpha\,\mathbb{E}_{D,\mathbf{W}|S=1}\{D\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\} + (1-\alpha)\mathbb{E}_{\mathbf{W}|S=0}\{\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\}$$

$$= \boldsymbol{\Phi}'(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0) + O_p(\|\widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0\|^2) = \boldsymbol{\Phi}'(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0) + O_p(u_n^2) \tag{S28}$$

$$= O_p(u_n), \tag{S29}$$

where the first step holds by Assumption 1 and the last two steps are due to Assumption 2. Let $\widehat{\mathbf{D}} := \widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}$. Combining (S27) and (S29) yields

$$\widehat{\mathbf{D}}[\alpha\, \mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\}]$$

$$= \widehat{\mathbf{D}}[\alpha\, \mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} + O_p(u_n) + o_p(n^{-1/2})]$$

$$= \widehat{\mathbf{D}}\{O_p(n^{-1/2}) + O_p(u_n) + o_p(n^{-1/2})\} = O_p(u_n v_n) + o_p(n^{-1/2}), \tag{S30}$$

where the second step holds by the central limit theorem and the last step is due to Assumption 2. Then, putting (S27), (S28) and (S30) together gives

$$\widehat{\boldsymbol{\Omega}}[\alpha\, \mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\}]$$

$$= \boldsymbol{\Omega}[\alpha\, \mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\}] + O_p(u_n v_n) + o_p(n^{-1/2})$$

$$= \boldsymbol{\Omega}[\alpha\, \mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\}] - (\widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0) +$$

$$O_p(u_n v_n + u_n^2) + o_p(n^{-1/2}),$$

which implies

$$\widehat{\mathbf{T}}_1 = \boldsymbol{\Omega}[\alpha\, \mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\}] +$$

$$O_p(u_n v_n + u_n^2) + o_p(n^{-1/2}). \tag{S31}$$

Next, we handle $\widehat{\mathbf{\Omega}}\widehat{\mathbf{T}}_2$. For $j = 1, \ldots, p$, let

$$\widehat{U}_\infty(\mathcal{V}_m^-) := \sup_{\mathbf{z}\in\mathcal{Z}}|\widehat{\mu}_m(\mathbf{z}) - \mu^*(\mathbf{z})|, \ \mathcal{F}_{j,m} := \{\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}\boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{B}_n\} \ \text{ and}$$

$$\widehat{U}_{2,j}(\mathcal{V}_m^-) := \mathbb{E}^{1/2}_{\mathbf{Z},\mathbf{W}|S=1}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}^2 \sup_{\boldsymbol{\theta}\in\mathcal{B}_n}\boldsymbol{\psi}^2_{[j]}(\mathbf{W}, \boldsymbol{\theta})] \quad (j = 1, \ldots, p),$$

where $\mathcal{B}_n := \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le t_n u_n\}$ for some positive sequence $t_n \to \infty$ satisfying $t_n u_n = o(1)$ and $\mathcal{B}_n \subset \mathcal{B}_0$ for every n. Using Taylor's expansion, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z},\mathbf{W}|S=1}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}^2 \sup_{\boldsymbol{\theta}\in\mathcal{B}_n}\{\boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta}) - \boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta}_0)\}^2]) \\
&\le \widehat{U}^2_\infty(\mathcal{V}_m^-)\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \boldsymbol{\theta})\|^2\}\sup_{\boldsymbol{\theta}\in\mathcal{B}_n}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 = O_p(t_n^2 u_n^2 a^2_{n,\infty}),
\end{aligned}$$

where the last step holds by Assumptions 1 and 3. Therefore, Assumption 3 ensures

$$\begin{aligned}
\widehat{U}^2_{2,j}(\mathcal{V}_m^-) \le c\,(&\mathbb{E}_{\mathbf{Z},\mathbf{W}|S=1}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}^2\boldsymbol{\psi}^2_{[j]}(\mathbf{W}, \boldsymbol{\theta}_0)] + \\
&\mathbb{E}_{\mathbf{Z},\mathbf{W}|S=1}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}^2 \sup_{\boldsymbol{\theta}\in\mathcal{B}_n}\{\boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta}) - \boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta}_0)\}^2]) \\
= O_p(&a^2_{n,2} + t_n^2 u_n^2 a^2_{n,\infty}).
\end{aligned} \tag{S32}$$

In addition, we also have

$$\widehat{U}_{2,j}(\mathcal{V}_m^-) \ge \widetilde{U}_{2,j}(\mathcal{V}_m^-) := \mathbb{E}^{1/2}_{\mathbf{Z},\mathbf{W}|S=1}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}^2\boldsymbol{\psi}^2_{[j]}(\mathbf{W}, \boldsymbol{\theta}_0)]. \tag{S33}$$

In the following, we will use the symbols $N_{[]}(\cdot,\cdot,\cdot)$ and $J_{[]}(\cdot,\cdot,\cdot)$ to represent the bracketing number and the bracketing integral defined in Van der Vaart and Wellner (1996). Since for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{B}_n$,

$$|\boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_1) - \boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_2)| \leq \sup_{\boldsymbol{\theta}\in\mathcal{B}_n}\|\boldsymbol{\psi}'_{[j]}(\mathbf{W},\boldsymbol{\theta})\|\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \quad \text{and}$$

$$\mathbb{E}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}^2 \sup_{\boldsymbol{\theta}\in\mathcal{B}_n}\|\boldsymbol{\psi}'_{[j]}(\mathbf{W},\boldsymbol{\theta})\|^2] \leq c\,\widehat{U}^2_\infty(\mathcal{V}_m^-)$$

due to Assumption 1, we know from Example 19.7 of Van der Vaart (2000) that

$$\begin{aligned} N_{[]}\{\epsilon, \mathcal{F}_{j,m} \mid \mathcal{V}_m^-, L_2(P_{\mathbf{Z}|S=1})\} &\leq \max\{1, c\,\{t_n u_n \widehat{U}_\infty(\mathcal{V}_m^-)/\epsilon\}^p\} \leq 1 + c\,\{t_n u_n \widehat{U}_\infty(\mathcal{V}_m^-)/\epsilon\}^p \\ &\leq \{1 + c\,t_n u_n \widehat{U}_\infty(\mathcal{V}_m^-)/\epsilon\}^p \end{aligned}$$

for any ϵ. It follows that

$$\begin{aligned} J_{[]}\{1, \mathcal{F}_{j,m} \mid \mathcal{V}_m^-, L_2(P_{\mathbf{Z}|S=1})\} &\equiv \int_0^1 [1 + \log N_{[]}\{\epsilon\,\widehat{U}_{2,j}(\mathcal{V}_m^-), \mathcal{F}_{j,m} \mid \mathcal{V}_m^-, L_2(P_{\mathbf{Z}|S=1})\}]^{1/2} d\epsilon \\ &\leq \int_0^1 1 + \log N_{[]}\{\epsilon\,\widehat{U}_{2,j}(\mathcal{V}_m^-), \mathcal{F}_{j,m} \mid \mathcal{V}_m^-, L_2(P_{\mathbf{Z}|S=1})\} d\epsilon \\ &\leq c\,[1 + \int_0^1 \log\{\epsilon + t_n u_n \widehat{U}_\infty(\mathcal{V}_m^-)/\widehat{U}_{2,j}(\mathcal{V}_m^-)\} d\epsilon - \int_0^1 \log\epsilon\, d\epsilon] \\ &\leq c\,[1 + \log\{1 + t_n u_n \widehat{U}_\infty(\mathcal{V}_m^-)/\widehat{U}_{2,j}(\mathcal{V}_m^-)\}] \\ &\leq c\,[1 + \log\{1 + t_n u_n \widehat{U}_\infty(\mathcal{V}_m^-)/\widetilde{U}_{2,j}(\mathcal{V}_m^-)\}] \\ &= O_p(\log(2 + t_n u_n a_{n,\infty}/a_{n,2})), \end{aligned} \tag{S34}$$

where the fifth step uses (S33) and the last step holds by Assumption 3. The above derivations are conditional on $\mathcal{V}_m^-$ and treats the function $\widehat{\mu}_m(\cdot)$

as nonrandom. Because $\{\{\widehat{\mu}_m(\mathbf{Z}_i) - \mu^*(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}) : i = n+1, \ldots, N_1\}$ are conditionally independent given $\mathcal{V}_m^-$, Theorem 2.14.2 of Van der Vaart and Wellner (1996) ensures

$$\mathbb{E}_{\mathbf{Z},\mathbf{W}|S=1}(\sup_{\boldsymbol{\theta}\in\mathcal{B}_n}|\mathbb{G}_{\mathcal{N}}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}\boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta})]|)$$
$$\leq c\, J_{[]}\{1, \mathcal{F}_{j,m} \mid \mathcal{V}_m^-, L_2(P_{\mathbf{Z}|S=1})\}\widehat{U}_{2,j}(\mathcal{V}_m^-) = O_p((a_{n,2} + t_n u_n a_{n,\infty}) \log(2 + t_n u_n a_{n,\infty}/a_{n,2})),$$

where the last step holds by (S32) and (S34). Then, applying Markov's inequality, we have for any positive sequence $s_n \to \infty$,

$$\mathbb{P}(\sup_{\boldsymbol{\theta}\in\mathcal{B}_n}|\mathbb{G}_{\mathcal{N}}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}\boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta})]| \geq$$
$$s_n(a_{n,2} + t_n u_n a_{n,\infty}) \log(2 + t_n u_n a_{n,\infty}/a_{n,2}) \mid \mathcal{V}_m^-) = o_p(1).$$

which, combined with Lemma 6.1 of Chernozhukov et al. (2018), implies

$$\sup_{\boldsymbol{\theta}\in\mathcal{B}_n}|\mathbb{G}_{\mathcal{N}}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}\boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta})]| = O_p((a_{n,2} + t_n u_n a_{n,\infty}) \log(2 + t_n u_n a_{n,\infty}/a_{n,2})).$$

Considering the fact that $\mathbb{P}(\widehat{\boldsymbol{\theta}}_{\text{IN}} \in \mathcal{B}_n) \to 1$ from Assumption 2, it follows that

$$\mathbb{G}_{\mathcal{N}}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}\boldsymbol{\psi}_{[j]}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})] = O_p((a_{n,2} + t_n u_n a_{n,\infty}) \log(2 + t_n u_n a_{n,\infty}/a_{n,2})).$$

Because t_n can diverge arbitrarily slowly, we know

$$\mathbb{G}_{\mathcal{N}}[\{\widehat{\mu}_m(\mathbf{Z}) - \mu^*(\mathbf{Z})\}\boldsymbol{\psi}_{[j]}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})] = O_p((a_{n,2} + u_n a_{n,\infty}) \log(2 + u_n a_{n,\infty}/a_{n,2})). \quad \text{(S35)}$$

Noticing that $\{\{\widehat{\mu}_m(\mathbf{Z}_i)-\mu^*(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}): i\in\mathcal{I}_m\}$ are also conditionally independent given $\mathcal{V}_m^-$ according to (7), we can show for $j=1,\dots,p$,

$$\mathbb{G}_{\mathcal{V}_m}[\{\widehat{\mu}_m(\mathbf{Z})-\mu^*(\mathbf{Z})\}\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})]=O_p((a_{n,2}+u_na_{n,\infty})\log(2+u_na_{n,\infty}/a_{n,2})), \quad \text{(S36)}$$

similarly to (S35). Combining (S35)–(S36) yields

$$\begin{aligned}
|\widehat{\mathbf{T}}_{2[j]}| &\le \textstyle\sum_{m=1}^M|\{(N_1-n)^{-1/2}\mathbb{G}_{\mathcal{N}}-(n/M)^{-1/2}\mathbb{G}_{\mathcal{V}_m}\}[\{\widehat{\mu}_m(\mathbf{Z})-\mu^*(\mathbf{Z})\}\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})]| \\
&\le \textstyle\sum_{m=1}^M(|(N_1-n)^{-1/2}\mathbb{G}_{\mathcal{N}}[\{\widehat{\mu}_m(\mathbf{Z})-\mu^*(\mathbf{Z})\}\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})]|+ \\
&\qquad |n^{-1/2}\mathbb{G}_{\mathcal{V}_m}[\{\widehat{\mu}_m(\mathbf{Z})-\mu^*(\mathbf{Z})\}\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})]|) \\
&= O_p((a_{n,2}+u_na_{n,\infty})\log(2+u_na_{n,\infty}/a_{n,2})/n^{1/2})+o_p(n^{-1/2}).
\end{aligned}$$

Hence, due to the fact that $\widehat{\boldsymbol{\Omega}}=\boldsymbol{\Omega}+\widehat{\mathbf{D}}=O_p(1+v_n)=O_p(1)$ from Assumption 2, we have

$$\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{T}}_2=O_p((a_{n,2}+u_na_{n,\infty})\log(2+u_na_{n,\infty}/a_{n,2})/n^{1/2})+o_p(n^{-1/2}). \quad \text{(S37)}$$

We now deal with $\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{T}}_3$. Since $\mu^*(\cdot)$ is bounded according to Assumption 3, we can show $\{\mu^*(\mathbf{Z})\boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}):\boldsymbol{\theta}\in\mathcal{B}_0\}$ is $P_{\mathbf{Z}|S=1}$-Donsker by the arguments around (S25). Also, similarly to (S26), we can obtain by using Taylor's expansion that for $j=1,\dots,p$,

$$\mathbb{E}_{\mathbf{Z},\mathbf{W}|S=1}([\mu^*(\mathbf{Z})\{\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})-\boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\}]^2)=o_p(1).$$

Then, Lemma 19.24 in Van der Vaart (2000) gives

$$\mathbb{G}_{\mathcal{V}\cup\mathcal{N}}[\mu^*(\mathbf{Z})\{\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})-\boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\}]=o_p(1) \ \text{ and}$$

$$\mathbb{G}_{\mathcal{V}}[\mu^*(\mathbf{Z})\{\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\}] = o_p(1).$$

It follows that

$$\begin{aligned}
\widehat{\mathbf{T}}_3 &= \alpha(\mathbb{E}_{\mathcal{V}\cup\mathcal{N}} - \mathbb{E}_{\mathcal{V}})\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\} + \\
&\quad \alpha(N_1^{-1/2}\mathbb{G}_{\mathcal{V}\cup\mathcal{N}} - n^{-1/2}\mathbb{G}_{\mathcal{V}})[\mu^*(\mathbf{Z})\{\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\}] \\
&= \alpha(\mathbb{E}_{\mathcal{V}\cup\mathcal{N}} - \mathbb{E}_{\mathcal{V}})\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\} + o_p(n^{-1/2}) \qquad \text{(S38)} \\
&= \alpha(N_1^{-1/2}\mathbb{G}_{\mathcal{V}\cup\mathcal{N}} - n^{-1/2}\mathbb{G}_{\mathcal{V}})\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\} + o_p(n^{-1/2}) = O_p(n^{-1/2}),
\end{aligned}$$

where the last step uses Chebyshev's inequality under the conditions that $\mu^*(\cdot)$ is bounded from Assumption 3. It follows that $\widehat{\mathbf{D}}\widehat{\mathbf{T}}_3 = O_p(v_n/n^{1/2}) = o_p(n^{-1/2})$ due to Assumption 2. This coupled with (S38) implies

$$\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{T}}_3 = \boldsymbol{\Omega}\widehat{\mathbf{T}}_3 + \widehat{\mathbf{D}}\widehat{\mathbf{T}}_3 = \alpha\,\boldsymbol{\Omega}(\mathbb{E}_{\mathcal{V}\cup\mathcal{N}} - \mathbb{E}_{\mathcal{V}})\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\} + o_p(n^{-1/2}). \qquad \text{(S39)}$$

Finally, putting (S24), (S31), (S37) and (S39) together yields the expansion (10), which can be rewritten as

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \boldsymbol{\Omega}\mathbf{G}_n + o_p(1) \qquad \text{(S40)}$$

under the condition $r_n = o_p(n^{-1/2})$, where

$$\begin{aligned}
\mathbf{G}_n := n^{1/2}(&\alpha\,\mathbb{E}_{\mathcal{V}}[\{D - (1-\delta_n)\mu^*(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)] + \alpha\,\mathbb{E}_{\mathcal{N}}\{(1-\delta_n)\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\} + \\
&(1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\}).
\end{aligned}$$

Then, we have $\mathrm{cov}(\mathbf{G}_n) \equiv \mathbf{A}_n(\mu^*)$ and

$$\mathbb{E}(\mathbf{G}_n/n^{1/2}) = \alpha\,\mathbb{E}[\{D - (1-\delta_n)\mu^*(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1] +$$
$$\alpha\,\mathbb{E}\{(1-\delta_n)\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1\} + (1-\alpha)\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 0\}$$
$$= \alpha\,\mathbb{E}\{D\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1\} + (1-\alpha)\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 0\} = \mathbf{0}.$$

Further, Assumption 3, the condition that $\mathbb{E}\{\|\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\|^{2(1+c)}\} < \infty$ for some $c > 0$, and the fact that $\|\boldsymbol{\Omega}\| < \infty$ together ensure

$$n\,\mathbb{E}[\|\boldsymbol{\Omega}\{D - (1-\delta_n)\mu^*(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\|^{2(1+c)} \mid S = 1]/n^{1+c} = o(1),$$
$$(N_1 - n)\mathbb{E}[\|\boldsymbol{\Omega}\{(1-\delta_n)\mu^*(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\|^{2(1+c)} \mid S = 1]/\{(N_1 - n)/n^{1/2}\}^{2(1+c)} = o(1),$$
$$N_0\,\mathbb{E}\{\|\boldsymbol{\Omega}\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\|^{2(1+c)} \mid S = 0\}/(N_0/n^{1/2})^{2(1+c)} = o(1),$$

which, combined with the fact that

$$\lambda_{\min}\{\mathrm{cov}(\boldsymbol{\Omega}\mathbf{G}_n)\} = \lambda_{\min}\{\boldsymbol{\Omega}\mathbf{A}_n(\mu^*)\boldsymbol{\Omega}\} \geq \lambda_{\min}\{\mathbf{A}_n(\mu^*)\}\lambda^2_{\min}(\boldsymbol{\Omega}) \geq c,$$

verify the Lyapunov condition. Hence, Lyapunov's central limit theorem gives

$$\{\boldsymbol{\Omega}\mathbf{A}_n(\mu^*)\boldsymbol{\Omega}\}^{-1/2}\mathbf{G}_n \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{I}).$$

This, coupled with (S40), concludes (13).

Proof of Proposition 1: Denote $\mathcal{D}_j := \{D\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{B}_0\}$. Using

Taylor's expansion, we have for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{B}_0$,

$$\|D\{\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \boldsymbol{\theta}_1) - \boldsymbol{\psi}'_{[j]}(\mathbf{W}, \boldsymbol{\theta}_2)\}\| \leq \sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\boldsymbol{\psi}''_{[j]}(\mathbf{W}, \boldsymbol{\theta})\| \, \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$

with $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\boldsymbol{\psi}''_{[j]}(\mathbf{W}, \boldsymbol{\theta})\| \mid S = 1\} < \infty$ as assumed. Therefore, Example 19.7 of Van der Vaart (2000) indicates $N_{[]}\{\epsilon, \mathcal{D}_j, L_1(P_{D,\mathbf{W}|S=1})\} < \infty$ for any $\epsilon > 0$. Then, Theorem 19.4 of Van der Vaart (2000) implies $\mathcal{D}_j$ is $P_{D,\mathbf{W}|S=1}$-Glivenko-Cantelli, i.e.,

$$\sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \boldsymbol{\theta})\} - \mathbb{E}\{D\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \boldsymbol{\theta}) \mid S = 1\}\| = o_p(1),$$

which, combined with the fact that $\mathbb{P}(\widehat{\boldsymbol{\theta}}_{\text{IN}} \in \mathcal{B}_0) \to 1$, ensures

$$\|\mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\} - \mathbb{E}_{D,\mathbf{W}|S=1}\{D\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\}\| = o_p(1). \tag{S41}$$

Similarly, we can show

$$\|\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\} - \mathbb{E}_{\mathbf{W}|S=0}\{\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\}\| = o_p(1). \tag{S42}$$

In addition, applying Taylor's expansion again, we can obtain from the assumption that

$$\begin{aligned}
&\|\mathbb{E}_{D,\mathbf{W}|S=1}[D\{\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}'_{[j]}(\mathbf{W}, \boldsymbol{\theta}_0)\}]\| \\
&\leq \sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\mathbb{E}\{\boldsymbol{\psi}''_{[j]}(\mathbf{W}, \boldsymbol{\theta})\}\| \, \|\widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0\| = o_p(1) \text{ and} \qquad \text{(S43)} \\
&\|\mathbb{E}_{\mathbf{W}|S=0}\{\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}'_{[j]}(\mathbf{W}, \boldsymbol{\theta}_0)\}\| = o_p(1). \qquad \text{(S44)}
\end{aligned}$$

Putting (S41)–(S44) together yields $\widehat{\boldsymbol{\Omega}}^{-1} - \boldsymbol{\Omega}^{-1} = o_p(1)$. Then, Lemma S1 implies

$$\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega} = o_p(1).$$

Proof of Theorem 2: Write

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \equiv \widehat{\mathbf{T}}_1 + \widehat{\boldsymbol{\Omega}}\widetilde{\mathbf{T}}_2, \text{ where} \tag{S45}$$

$$\widehat{\mathbf{T}}_1 \equiv \widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0 + \widehat{\boldsymbol{\Omega}}[\alpha\, \mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\}],$$

$$\widetilde{\mathbf{T}}_2 := \alpha(1-\delta_n)(\mathbb{E}_{\mathcal{N}} - \mathbb{E}_{\mathcal{V}})\{\widehat{\mu}(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\}$$

$$\equiv \alpha(1-\delta_n)M^{-1}\textstyle\sum_{m=1}^{M}(\mathbb{E}_{\mathcal{N}} - \mathbb{E}_{\mathcal{V}_m})\{\widehat{\mu}_m(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\}.$$

Recalling (S31), we have

$$\begin{aligned}
\widehat{\mathbf{T}}_1 &= \boldsymbol{\Omega}[\alpha\, \mathbb{E}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\}] + O_p(u_n v_n + u_n^2) + o_p(n^{-1/2}) \\
&= \boldsymbol{\Omega}[\alpha\, \mathbb{G}_{\mathcal{V}}\{D\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} + (1-\alpha)\mathbb{G}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\}] + O_p(u_n v_n + u_n^2) + o_p(n^{-1/2}) \\
&= O_p(n^{-1/2} + u_n v_n + u_n^2),
\end{aligned} \tag{S46}$$

where the last step holds by the central limit theorem. Regarding $\widetilde{\mathbf{T}}_2$, Taylor's expansion guarantees for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{B}_0$,

$$|\widehat{\mu}_m(\mathbf{Z})\{\boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta}_1) - \boldsymbol{\psi}_{[j]}(\mathbf{W}, \boldsymbol{\theta}_2)\}| \leq \sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \boldsymbol{\theta})\|\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$

with $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\boldsymbol{\psi}'_{[j]}(\mathbf{W}, \boldsymbol{\theta})\|^2 \mid S = 1\} < \infty$ $(j = 1, \ldots, p)$, because of Assumption 1 and the boundedness of $\widehat{\mu}_m(\cdot)$. Hence, it follows from Example

19.7 of Van der Vaart (2000) that $\{\widehat{\mu}_m(\mathbf{Z})\boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{B}_0\}$ is (conditionally) $P_{\mathbf{W}|S=1}$-Donsker given $\mathcal{V}_m^-$. Further, applying Taylor's expansion yields for $j = 1, \ldots, p$ that

$$\begin{aligned}&\mathbb{E}_{\mathbf{W}|S=1}[\widehat{\mu}_m^2(\mathbf{Z})\{\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\rm IN}) - \boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\}^2]\\ &\le \|\widehat{\boldsymbol{\theta}}_{\rm IN} - \boldsymbol{\theta}_0\|^2\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\mathcal{B}_0}\|\boldsymbol{\psi}'_{[j]}(\mathbf{W},\boldsymbol{\theta})\|^2 \mid S=1\} = O_p(u_n^2) = o_p(1).\end{aligned}$$

Here we use Assumptions 1 and 2, as well as the boundedness of $\widehat{\mu}_m(\cdot)$. The above derivations have verified the conditions of Lemma 19.24 in Van der Vaart (2000), which ensures

$$\mathbb{P}(\mathbb{G}_{\mathcal{V}_m}[\widehat{\mu}_m(\mathbf{Z})\{\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\rm IN}) - \boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\}] \ge c \mid \mathcal{V}_m^-) = o_p(1)$$

for any constant $c > 0$. Due to Lemma 6.1 of Chernozhukov et al. (2018), it follows that

$$\mathbb{G}_{\mathcal{V}_m}[\widehat{\mu}_m(\mathbf{Z})\{\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\rm IN}) - \boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\}] = o_p(1). \tag{S47}$$

Also, we know from Chebyshev's inequality that for any positive sequence $s_n \to \infty$,

$$\mathbb{P}[\mathbb{G}_{\mathcal{V}_m}\{\widehat{\mu}_m(\mathbf{Z})\boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\} > s_n \mid \mathcal{V}_m^-] = o_p(1)$$

because $\widehat{\mu}_m(\cdot)$ is bounded. Hence, Lemma 6.1 of Chernozhukov et al. (2018) indicates

$$\mathbb{G}_{\mathcal{V}_m}\{\widehat{\mu}_m(\mathbf{Z})\boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\} = O_p(1). \tag{S48}$$

Putting (S47) and (S48) together yields

$$
\begin{aligned}
&\mathbb{G}_{\mathcal{V}_m}\{\widehat{\mu}_m(\mathbf{Z})\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\} \\
&= \mathbb{G}_{\mathcal{V}_m}\{\widehat{\mu}_m(\mathbf{Z})\boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\} + \mathbb{G}_{\mathcal{V}_m}[\widehat{\mu}_m(\mathbf{Z})\{\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}_{[j]}(\mathbf{W},\boldsymbol{\theta}_0)\}] = O_p(1). \quad \text{(S49)}
\end{aligned}
$$

Similarly, we can show

$$
\mathbb{G}_{\mathcal{N}}\{\widehat{\mu}_m(\mathbf{Z})\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\} = O_p(1). \tag{S50}
$$

Therefore, we have

$$
\begin{aligned}
|\widetilde{\mathbf{T}}_{2[j]}| &\leq |\textstyle\sum_{m=1}^{M}\{(N_1-n)^{-1/2}\mathbb{G}_{\mathcal{N}} - (n/M)^{-1/2}\mathbb{G}_{\mathcal{V}_m}\}\{\widehat{\mu}_m(\mathbf{Z})\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\}| \\
&\leq \textstyle\sum_{m=1}^{M}|(N_1-n)^{-1/2}\mathbb{G}_{\mathcal{N}}\{\widehat{\mu}_m(\mathbf{Z})\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\}| + \\
&\quad \textstyle\sum_{m=1}^{M}|(n/M)^{-1/2}\mathbb{G}_{\mathcal{V}_m}\{\widehat{\mu}_m(\mathbf{Z})\boldsymbol{\psi}_{[j]}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\}| = O_p(n^{-1/2}).
\end{aligned}
$$

Considering the fact that $\widehat{\boldsymbol{\Omega}} = \boldsymbol{\Omega} + \widehat{\mathbf{D}} = O_p(1+v_n) = O_p(1)$ from Assumption 2, we have

$$
\widehat{\boldsymbol{\Omega}}\widetilde{\mathbf{T}}_2 = O_p(n^{-1/2}). \tag{S51}
$$

Finally, combining (S45), (S46) and (S51) yields $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O_p(n^{-1/2} + u_n v_n + u_n^2)$.

Now, we turn to proving the properties of $\widetilde{\boldsymbol{\theta}}$. Write

$$
\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \equiv \widehat{\mathbf{R}}_1 + \widehat{\boldsymbol{\Omega}}\widehat{\mathbf{R}}_2, \text{ where} \tag{S52}
$$

$$
\widehat{\mathbf{R}}_1 := \widehat{\boldsymbol{\theta}}_{\text{IN}} - \boldsymbol{\theta}_0 + \widehat{\boldsymbol{\Omega}}[\alpha\,\mathbb{E}_{\widetilde{\mathcal{V}}_2}\{D\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W},\widehat{\boldsymbol{\theta}}_{\text{IN}})\}],
$$

$$\widehat{\mathbf{R}}_2 := \alpha(\mathbb{E}_{\widetilde{\mathcal{V}}_2 \cup \mathcal{N}} - \mathbb{E}_{\widetilde{\mathcal{V}}_2})\{\widetilde{\mu}_1(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\}.$$

Similarly to (S31), we can show

$$\widehat{\mathbf{R}}_1 = \boldsymbol{\Omega}[\alpha\, \mathbb{E}_{\widetilde{\mathcal{V}}_2}\{D\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} + (1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\}] + O_p(u_n v_n + u_n^2) + o_p(n_2^{-1/2}). \tag{S53}$$

Further, we know from (S47) that $\mathbb{G}_{\widetilde{\mathcal{V}}_2}[\widetilde{\mu}_1(\mathbf{Z})\{\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\}] = o_p(1)$. Similarly, we have $\mathbb{G}_{\widetilde{\mathcal{V}}_2 \cup \mathcal{N}}[\widetilde{\mu}_1(\mathbf{Z})\{\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\}] = o_p(1)$. It follows that

$$\begin{aligned}\widehat{\mathbf{R}}_2 &= \alpha(\mathbb{E}_{\widetilde{\mathcal{V}}_2 \cup \mathcal{N}} - \mathbb{E}_{\widetilde{\mathcal{V}}_2})\{\widetilde{\mu}_1(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} + \\ &\quad \alpha\{(N_1 - n_1)^{-1/2}\mathbb{G}_{\widetilde{\mathcal{V}}_2 \cup \mathcal{N}} - n_2^{-1/2}\mathbb{G}_{\widetilde{\mathcal{V}}_2}\}[\widetilde{\mu}_1(\mathbf{Z})\{\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}}) - \boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\}] \\ &= \alpha(\mathbb{E}_{\widetilde{\mathcal{V}}_2 \cup \mathcal{N}} - \mathbb{E}_{\widetilde{\mathcal{V}}_2})\{\widetilde{\mu}_1(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} + o_p(n_2^{-1/2}).\end{aligned} \tag{S54}$$

Moreover, we know from (S49) and (S50) that

$$\widehat{\mathbf{R}}_2 = \alpha\{(N_1 - n_1)^{-1/2}\mathbb{G}_{\widetilde{\mathcal{V}}_2 \cup \mathcal{N}} - n_2^{-1/2}\mathbb{G}_{\widetilde{\mathcal{V}}_2}\}\{\widetilde{\mu}_1(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \widehat{\boldsymbol{\theta}}_{\text{IN}})\} = O_p(n_2^{-1/2}),$$

which implies $\widehat{\mathbf{D}}\widehat{\mathbf{R}}_2 = O_p(v_n/n_2^{1/2}) = o_p(n_2^{-1/2})$ due to Assumption 2. This combined with (S54) gives

$$\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{R}}_2 = \boldsymbol{\Omega}\widehat{\mathbf{R}}_2 + \widehat{\mathbf{D}}\widehat{\mathbf{R}}_2 = \alpha\, \boldsymbol{\Omega}(\mathbb{E}_{\widetilde{\mathcal{V}}_2 \cup \mathcal{N}} - \mathbb{E}_{\widetilde{\mathcal{V}}_2})\{\widetilde{\mu}_1(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} + o_p(n_2^{-1/2}). \tag{S55}$$

Putting (S52), (S53) and (S55) together yields $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 =$

$$\boldsymbol{\Omega}[\alpha\, n_2^{-1}\textstyle\sum_{i=n_1+1}^{n}\{D_i - \widetilde{\mu}_1(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0) + \alpha(N_1 - n_1)^{-1}\sum_{i=n_1+1}^{N_1}\widetilde{\mu}_1(\mathbf{Z}_i)\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0) +$$
$$(1-\alpha)N_0^{-1}\textstyle\sum_{i=N_1+1}^{N}\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0)] + O_p(u_n v_n + u_n^2) + o_p(n_2^{-1/2}).$$

which can be rewritten as

$$n_2^{1/2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \boldsymbol{\Omega}\widehat{\mathbf{H}}_n + o_p(1) \qquad \text{(S56)}$$

under the condition $u_n v_n + u_n^2 = o_p(n_2^{-1/2})$, where

$$\widehat{\mathbf{H}}_n := n_2^{1/2}(\alpha\, \mathbb{E}_{\widetilde{\mathcal{V}}_2}[\{D - (1-\widetilde{\delta}_n)\widetilde{\mu}_1(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)] + \alpha\, \mathbb{E}_{\mathcal{N}}\{(1-\widetilde{\delta}_n)\widetilde{\mu}_1(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\} +$$
$$(1-\alpha)\mathbb{E}_{\mathcal{C}}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\}).$$

Then, we have $\mathrm{cov}(\widehat{\mathbf{H}}_n \mid \widetilde{\mathcal{V}}_1) \equiv \widetilde{\mathbf{A}}_n$ and

$$\mathbb{E}(\widehat{\mathbf{H}}_n / n_2^{1/2} \mid \widetilde{\mathcal{V}}_1) = \alpha\, \mathbb{E}[\{D - (1-\widetilde{\delta}_n)\widetilde{\mu}_1(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid \widetilde{\mathcal{V}}_1, S=1] +$$
$$\alpha\, \mathbb{E}\{(1-\widetilde{\delta}_n)\widetilde{\mu}_1(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid \widetilde{\mathcal{V}}_1, S=1\} +$$
$$(1-\alpha)\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S=0\} = \mathbf{0}.$$

Further, the boundedness of $\widetilde{\mu}_1(\cdot)$, the condition that $\mathbb{E}\{\|\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\|^{2(1+c)}\} < \infty$ for some $c > 0$, and the fact that $\|\boldsymbol{\Omega}\| < \infty$ together ensure

$$n_2\, \mathbb{E}[\|\boldsymbol{\Omega}\{D - (1-\widetilde{\delta}_n)\widetilde{\mu}_1(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\|^{2(1+c)} \mid S=1]/n_2^{1+c} = o(1),$$
$$(N_1 - n)\mathbb{E}[\|\boldsymbol{\Omega}\{(1-\widetilde{\delta}_n)\widetilde{\mu}_1(\mathbf{Z})\}\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\|^{2(1+c)} \mid S=1]/\{(N_1 - n)/n_2^{1/2}\}^{2(1+c)} = o(1),$$

$$N_0\,\mathbb{E}\{\|\boldsymbol{\Omega}\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\|^{2(1+c)} \mid S=0\}/(N_0/n_2^{1/2})^{2(1+c)} = o(1),$$

which, combined with the fact that

$$\lambda_{\min}^{-1}\{\mathrm{cov}(\boldsymbol{\Omega}\widehat{\mathbf{H}}_n \mid \widetilde{\mathcal{V}}_1)\} = \lambda_{\min}^{-1}(\boldsymbol{\Omega}\widetilde{\mathbf{A}}_n\boldsymbol{\Omega}) \leq \lambda_{\min}^{-1}(\widetilde{\mathbf{A}}_n)\lambda_{\min}^{-2}(\boldsymbol{\Omega}) = O_p(1),$$

indicate that with probability tending to one, the Lyapunov condition holds and therefore

$$(\boldsymbol{\Omega}\widetilde{\mathbf{A}}_n\boldsymbol{\Omega})^{-1/2}\widehat{\mathbf{H}}_n \mid \widetilde{\mathcal{V}}_1 \xrightarrow{d} \mathbf{N}(\mathbf{0},\mathbf{I}). \tag{S57}$$

It follows that for any $\mathbf{t} \in \mathbb{R}^p$,

$$\mathbb{E}[\exp\{i\mathbf{t}^{\mathrm{T}}(\boldsymbol{\Omega}\widetilde{\mathbf{A}}_n\boldsymbol{\Omega})^{-1/2}\widehat{\mathbf{H}}_n\} \mid \widetilde{\mathcal{V}}_1] - \exp(-\|\mathbf{t}\|^2/2) = o_p(1)$$

with i the imaginary unit. Then, the dominant convergence theorem ensures

$$\mathbb{E}[\exp\{i\mathbf{t}^{\mathrm{T}}(\boldsymbol{\Omega}\widetilde{\mathbf{A}}_n\boldsymbol{\Omega})^{-1/2}\widehat{\mathbf{H}}_n\}] \to \exp(-\|\mathbf{t}\|^2/2) \text{ for any } \mathbf{t} \in \mathbb{R}^p,$$

which means $(\boldsymbol{\Omega}\widetilde{\mathbf{A}}_n\boldsymbol{\Omega})^{-1/2}\widehat{\mathbf{H}}_n \xrightarrow{d} \mathbf{N}(\mathbf{0},\mathbf{I})$. This, coupled with (S56), concludes

$$n_2^{1/2}(\boldsymbol{\Omega}\widetilde{\mathbf{A}}_n\boldsymbol{\Omega})^{-1/2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{N}(\mathbf{0},\mathbf{I}) \text{ as } n_2 \to \infty. \tag{S58}$$

Proof of Corollary S1: Since $\mathbb{E}\{\|\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\|^2\} < \infty$, $\mu^*(\cdot)$ is bounded and $n = o(N)$, we have by Chebyshev's inequality that

$$(\mathbb{E}_{\mathcal{V}\cup\mathcal{N}} - \mathbb{E}_{\mathbf{Z},\mathbf{W}|S=1})\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W},\boldsymbol{\theta}_0)\} = O_p(N_1^{-1/2}) = o_p(n^{-1/2}) \text{ and}$$

$$(\mathbb{E}_{\mathcal{C}} - \mathbb{E}_{\mathbf{W}|S=0})\{\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0)\} = O_p(N_0^{-1/2}) = o_p(n^{-1/2}),$$

which, combined with (10), imply (S5). Further, under the condition that $\lim_{n\to\infty} \delta_n = 0$, we know $\mathbf{A}_0^{-1}\mathbf{A}_n(\mu^*) \to \mathbf{I}$, which indicates

$$(\boldsymbol{\Omega}\mathbf{A}_0\boldsymbol{\Omega})^{-1/2}\{\boldsymbol{\Omega}\mathbf{A}_n(\mu^*)\boldsymbol{\Omega}\}^{1/2} = \boldsymbol{\Omega}^{-1/2}\mathbf{A}_0^{-1/2}\mathbf{A}_n(\mu^*)^{1/2}\boldsymbol{\Omega}^{1/2} \to \mathbf{I}. \quad \text{(S59)}$$

Based on (13) and (S59), applying Slutsky's theorem gives (S6).

Proofs of Equations (S8), (S11), (S14) and (S15) in Section S3: We first show (S8). Since

$$\mathbb{E}\{\|\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0)\|^2\} < \infty$$

and $\mu(\cdot)$ is bounded, Chebyshev's inequality gives

$$n^{-1}\textstyle\sum_{i=1}^{n}\alpha\{D_i - \mu(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0) = O_p(n^{-1/2}),$$

which, combined with the fact that $\lim_{n\to\infty}\delta_n = \delta > 0$, implies

$$\begin{aligned}
&(\delta^{-1} - \delta_n^{-1})N^{-1}\textstyle\sum_{i=1}^{N}\alpha\,\tau^{-1}R_iS_i\{D_i - \mu(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0)\\
&= (\delta_n/\delta - 1)N^{-1}\textstyle\sum_{i=1}^{N}\alpha(\delta_n\tau)^{-1}R_iS_i\{D_i - \mu(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0)\\
&= (\delta_n/\delta - 1)n^{-1}\textstyle\sum_{i=1}^{n}\alpha\{D_i - \mu(\mathbf{Z}_i)\}\boldsymbol{\psi}(\mathbf{W}_i, \boldsymbol{\theta}_0) = o_p(n^{-1/2}). \quad \text{(S60)}
\end{aligned}$$

Because $\mu^*(\cdot) = \mu(\cdot)$ and $r_n = o_p(n^{-1/2})$, it follows from (10) and (S60) that

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = N^{-1}\textstyle\sum_{i=1}^{N}[\alpha(\delta_n\tau)^{-1}R_iS_i\{D_i - \mu(\mathbf{Z}_i)\} + \alpha\,\tau^{-1}S_i\mu(\mathbf{Z}_i) +$$

$$(1-\alpha)(1-\tau)^{-1}(1-S_i)]\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}_0)+o_p(n^{-1/2})$$

$$= N^{-1}\textstyle\sum_{i=1}^N[\alpha(\delta\tau)^{-1}R_iS_i\{D_i-\mu(\mathbf{Z}_i)\}+\alpha\,\tau^{-1}S_i\mu(\mathbf{Z}_i)+$$

$$(1-\alpha)(1-\tau)^{-1}(1-S_i)]\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}_0)+o_p(n^{-1/2}), \tag{S61}$$

In addition, we have

$$N^{-1}\textstyle\sum_{i=1}^N[\alpha\,\tau^{-1}S_i\,\mathbb{E}\{\mu(\mathbf{Z}_i)\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}_0)\mid S=1\}+$$

$$(1-\alpha)(1-\tau)^{-1}(1-S_i)\mathbb{E}\{\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}_0)\mid S=0\})]$$

$$=\alpha\,\mathbb{E}\{\mu(\mathbf{Z}_i)\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}_0)\mid S=1\}+(1-\alpha)\mathbb{E}\{\mu(\mathbf{Z}_i)\boldsymbol{\psi}(\mathbf{W}_i,\boldsymbol{\theta}_0)\mid S=0\}=\mathbf{0}, \tag{S62}$$

where the first step uses the fact that $N^{-1}\sum_{i=1}^N S_i\equiv\tau$ from the case-control sampling, and the last step holds by the definition (1) of $\boldsymbol{\theta}_0$. Combining (S61) and (S62) yields (S8).

Next, we prove (S11). According to Theorem 4.5 of Tsiatis (2007), the orthogonal complement of the tangent space Λ corresponding to the semiparametric model $\mathcal{M}$ is given by $\Lambda^\perp := \Lambda_{\widetilde{S}}\oplus\Lambda_{R\mid\widetilde{S},\mathbf{Z}}$ with

$$\Lambda_{\widetilde{S}} := \{\mathbf{g}(\widetilde{S})\in\mathbb{R}^p : \mathbb{E}\{\mathbf{g}(\widetilde{S})\}=\mathbf{0},\operatorname{cov}\{\mathbf{g}(\widetilde{S})\}<\infty\}\ \text{ and}$$

$$\Lambda_{R\mid\widetilde{S},\mathbf{Z}} := \{\mathbf{g}(R,\widetilde{S},\mathbf{Z})\in\mathbb{R}^p : \mathbb{E}\{\mathbf{g}(R,\widetilde{S},\mathbf{Z})\mid\widetilde{S},\mathbf{Z}\}=\mathbf{0},\operatorname{cov}\{\mathbf{g}(R,\widetilde{S},\mathbf{Z})\}<\infty\}.$$

Denote $\widetilde{\mathbf{W}} := (\widetilde{S},Y,\mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$. Since $\mu(\mathbf{Z})\equiv\mathbb{E}(D\mid\mathbf{Z},S=1)=\mathbb{E}(D\mid\mathbf{Z},\widetilde{S}=$

1) and $R \perp\!\!\!\perp D \mid \widetilde{S}$ from (S9), we have

$$\mathbb{E}[R\widetilde{S}\{D - \mu(\mathbf{Z})\}\boldsymbol{\psi}(\widetilde{\mathbf{W}}, \boldsymbol{\theta}_0) \mid R, \widetilde{S}, \mathbf{Z}] = \mathbf{0}.$$

Therefore, it follows from Theorem 4.5 of Tsiatis (2007) that the projections of the influence function $\boldsymbol{\varphi}(\mathbf{V})$ onto $\Lambda_{\widetilde{S}}$ and $\Lambda_{R|\widetilde{S},\mathbf{Z}}$ are

$$\Pi\{\boldsymbol{\varphi}(\mathbf{V}) \mid \Lambda_{\widetilde{S}}\} = \mathbb{E}\{\boldsymbol{\varphi}(\mathbf{V}) \mid \widetilde{S}\} - \mathbb{E}\{\boldsymbol{\varphi}(\mathbf{V})\} = \mathbf{0} \quad \text{and}$$

$$\Pi\{\boldsymbol{\varphi}(\mathbf{V}) \mid \Lambda_{R|\widetilde{S},\mathbf{Z}}\} = \mathbb{E}\{\boldsymbol{\varphi}(\mathbf{V}) \mid R, \widetilde{S}, \mathbf{Z}\} - \mathbb{E}\{\boldsymbol{\varphi}(\mathbf{V}) \mid \widetilde{S}, \mathbf{Z}\} = \mathbf{0}.$$

Hence, Theorem 4.3 of Tsiatis (2007) implies

$$\begin{aligned} \boldsymbol{\varphi}_{\text{EFF}}(\mathbf{V}) &= \Pi\{\boldsymbol{\varphi}(\mathbf{V}) \mid \Lambda\} = \boldsymbol{\varphi}(\mathbf{V}) - \Pi\{\boldsymbol{\varphi}(\mathbf{V}) \mid \Lambda^{\perp}\} \\ &= \boldsymbol{\varphi}(\mathbf{V}) - \Pi\{\boldsymbol{\varphi}(\mathbf{V}) \mid \Lambda_{\widetilde{S}}\} - \Pi\{\boldsymbol{\varphi}(\mathbf{V}) \mid \Lambda_{R|\widetilde{S},\mathbf{Z}}\} = \boldsymbol{\varphi}(\mathbf{V}). \end{aligned}$$

Now we turn to (S14). Because $\mu^*(\mathbf{Z}) = \mu(\mathbf{Z}) \equiv \mathbb{E}(D \mid \mathbf{Z}, S = 1)$, we have

$$\begin{aligned} &\alpha\, \mathbb{E}\{\mu^*(\mathbf{Z})\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1\} + (1 - \alpha)\, \mathbb{E}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 0\} \\ &= \alpha\, \mathbb{E}\{D\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 1\} + (1 - \alpha)\, \mathbb{E}\{\boldsymbol{\psi}(\mathbf{W}, \boldsymbol{\theta}_0) \mid S = 0\} = \mathbf{0}, \end{aligned}$$

where the last step holds by the definition (1) of $\boldsymbol{\theta}_0$. This, combined with (S5), gives (S14).

Regarding (S15), we know from Theorem 4.5 of Tsiatis (2007) that the tangent space according to $\mathcal{M}_0$ is

$$\Lambda_0 := \{\mathbf{g}(D, \mathbf{Z}) \in \mathbb{R}^p : \mathbb{E}\{\mathbf{g}(D, \mathbf{Z}) \mid \mathbf{Z}, S = 1\} = \mathbf{0}, \text{cov}\{\mathbf{g}(D, \mathbf{Z})\} < \infty\},$$

and that the projection of the influence function $\boldsymbol{\varphi}_0(D,\mathbf{Z})$ onto Λ_0 is

$$\begin{aligned}
\Pi\{\boldsymbol{\varphi}_0(D,\mathbf{Z}) \mid \Lambda_0\} &= \boldsymbol{\varphi}_0(D,\mathbf{Z}) - \mathbb{E}\{\boldsymbol{\varphi}_0(D,\mathbf{Z}) \mid \mathbf{Z}, S=1\} \\
&= \boldsymbol{\varphi}_0(D,\mathbf{Z}) - \alpha\{\mathbb{E}(D \mid \mathbf{Z}, S=1) - \mu(\mathbf{Z})\}\boldsymbol{\Omega}\boldsymbol{\psi}\{(S=1,Y,\mathbf{X}^{\mathrm{T}})^{\mathrm{T}},\boldsymbol{\theta}_0\} \\
&= \boldsymbol{\varphi}_0(D,\mathbf{Z}).
\end{aligned}$$

Then, it follows from Theorem 4.3 of Tsiatis (2007) that

$$\boldsymbol{\varphi}_{\text{EFF}}^{(0)}(D,\mathbf{Z}) = \Pi\{\boldsymbol{\varphi}_0(D,\mathbf{Z}) \mid \Lambda_0\} = \boldsymbol{\varphi}_0(D,\mathbf{Z}).$$

<u>Proof of Theorem S1</u>: For $\widehat{\theta}$, write

$$\widehat{\theta} - \theta_0 \equiv T_1 + (1-n/N)\widehat{T}_2, \text{ where} \tag{S63}$$

$$T_1 := \mathbb{E}_{\mathcal{L}}(Y-\theta_0) \text{ and } \widehat{T}_2 := (\mathbb{E}_{\mathcal{U}} - \mathbb{E}_{\mathcal{L}})\{\widehat{\phi}(\mathbf{X})\} \equiv M^{-1}\textstyle\sum_{m=1}^{M}(\mathbb{E}_{\mathcal{U}} - \mathbb{E}_{\mathcal{L}_m})\{\widehat{\phi}_m(\mathbf{X})\}.$$

Since $\mathbb{E}(Y^2) = O(1)$, we know $T_1 = O_p(n^{-1/2})$. Further, considering the assumption that $\mathbb{E}_{\mathbf{X}}\{\widehat{\phi}_m(\mathbf{X})\} = O_p(1)$ for $m = 1,\ldots,M$ and the fact that $\{\widehat{\phi}_m(\mathbf{X}_i) : i = n+1,\ldots,N\}$ are conditionally independent given $\mathcal{L}_m^-$, we have from Chebyshev's inequality that

$$\mathbb{P}[\mathbb{G}_{\mathcal{U}}\{\widehat{\phi}_m(\mathbf{X})\} > s_n \mid \mathcal{L}_m^-] = o_p(1)$$

for any positive sequence $s_n \to \infty$. This, combined with Lemma 6.1 of Chernozhukov et al. (2018), implies $\mathbb{G}_{\mathcal{U}}\{\widehat{\phi}_m(\mathbf{X})\} = O_p(1)$. Similarly, we can show $\mathbb{G}_{\mathcal{L}_m}\{\widehat{\phi}_m(\mathbf{X})\} = O_p(1)$ since $\{\widehat{\phi}_m(\mathbf{X}_i) : i \in \mathcal{I}_m\}$ are also conditionally independent given $\mathcal{L}_m^-$. Therefore, we have

$$
\begin{aligned}
|(1-n/N)\widehat{T}_2| &\leq \textstyle\sum_{m=1}^M [|(N-n)^{-1/2}\mathbb{G}_{\mathcal{U}}\{\widehat{\phi}_m(\mathbf{X})\}| + |(n/M)^{-1/2}\mathbb{G}_{\mathcal{L}_m}\{\widehat{\phi}_m(\mathbf{X})\}|] \\
&= O_p(n^{-1/2}).
\end{aligned}
$$

Putting pieces together yields $\widehat{\theta} - \theta_0 = O_p(n^{-1/2})$.

For $\widetilde{\theta}$, we have $\mathbb{E}(\widetilde{\theta} - \theta_0 \mid \widetilde{\mathcal{L}}_1) = 0$. Notice also that $n_2^{1/2}(\widetilde{\theta} - \theta_0) =$

$$
\textstyle\sum_{i=n_1+1}^n [n_2^{-1/2}\{Y_i - \widetilde{\phi}_1(\mathbf{X}_i)\} + n_2^{1/2}(N-n_1)^{-1}\widetilde{\phi}_1(\mathbf{X}_i)] + \sum_{i=n+1}^N n_2^{1/2}(N-n_1)^{-1}\widetilde{\phi}_1(\mathbf{X}_i). \tag{S64}
$$

Since $\widetilde{\phi}_1(\cdot)$ involves $\widetilde{\mathcal{L}}_1$ only, we know

$$
\begin{aligned}
&\{n_2^{-1/2}\{Y_i - \widetilde{\phi}_1(\mathbf{X}_i)\} + n_2^{1/2}(N-n_1)^{-1}\widetilde{\phi}_1(\mathbf{X}_i) : i = n_1+1, \ldots, n\} \cup \\
&\{n_2^{1/2}(N-n_1)^{-1}\widetilde{\phi}_1(\mathbf{X}_i) : i = n+1, \ldots, N\}
\end{aligned}
$$

are conditionally independent given $\widetilde{\mathcal{L}}_1$. Similaryly to (S57), applying the Lyapunov central limit theorem to (S64) yields $n_2^{1/2}\widetilde{\sigma}_n^{-1/2}(\widetilde{\theta} - \theta_0) \mid \widetilde{\mathcal{L}}_1 \xrightarrow{d} \mathbf{N}(0,1)$ as $n_2 \to \infty$. By the argument used to obtained (S58), we have $n_2^{1/2}\widetilde{\sigma}_n^{-1/2}(\widetilde{\theta} - \theta_0) \xrightarrow{d} \mathbf{N}(0,1)$ as $n_2 \to \infty$.

References

Beesley, L. J. and B. Mukherjee (2022). Statistical inference for association studies using electronic health records: Handling both selection bias and outcome misclassification. *Biometrics 78*(1), 214–226.

Bloniarz, A., H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences 113*(27), 7383–7390.

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

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

Jiang, Y., A. J. Scott, and C. J. Wild (2006). Secondary analysis of case-control data. *Statistics in medicine 25*(8), 1323–1339.

Lancaster, T. and G. Imbens (1996). Case-control studies with contaminated controls. *Journal of Econometrics 71*(1-2), 145–160.

Liang, L., R. J. Carroll, and Y. Ma (2018). Dimension reduction and estimation in the secondary analysis of case-control studies. *Electronic Journal of Statistics 12*(1), 1782–1821.

Ma, Y. (2010). A semiparametric efficient estimator in case-control studies. *Bernoulli 16*(2), 585–603.

Ma, Y. and R. J. Carroll (2016). Semiparametric estimation in the secondary analysis of case–control studies. *Journal of the Royal Statistical Society: Series B 78*(1), 127–151.

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

Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *The Annals of Statistics 27*(4), 1368–1384.

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

Tian, L., T. Cai, E. Goetghebeur, and L. Wei (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika 94*(2), 297–311.

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.

Van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and*

Empirical Processes: With Applications to Statistics. Springer Series in Statistics. Springer.

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.

Wei, J., R. J. Carroll, U. U. Müller, I. V. Keilegom, and N. Chatterjee (2013). Robust estimation for homoscedastic regression in the secondary analysis of case–control data. *Journal of the Royal Statistical Society: Series B 75*(1), 185–206.

Wei, Y., X. Song, M. Liu, I. Ionita-Laza, and J. Reibman (2016). Quantile regression in the secondary analysis of case–control data. *Journal of the American Statistical Association 111*(513), 344–354.

Zhang, A., L. D. Brown, and T. T. Cai (2019). Semi-supervised inference: General theory and estimation of means. *Annals of Statistics 47*(5), 2538–2566.

Zhang, X. and Z. Tan (2020). Semi-supervised logistic learning based on exponential tilt mixture models. *Stat 9*(1), e312.

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