Appendix for "Data Integration with Oracle Use of External Information from Heterogeneous Populations"

Yuqi Zhai*and Peisong Han†

PROOFS

For ease of notation, let $\hat{F}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\log f_i(\boldsymbol{\beta})$, $F(\boldsymbol{\beta}) = \mathbb{E}\left[\log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})\right]$, $\hat{Q}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\rho}) = n^{-1}\sum_{i=1}^{n}\log\left\{1-\boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma})]\right\}$, $\hat{\boldsymbol{r}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^{n}[\boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma}]$, $\hat{H}_n(\boldsymbol{\beta},\boldsymbol{\gamma}) = \{\boldsymbol{\rho}:\boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma})]<1, i=1,\cdots,n\}$, $\mathcal{K}_{=0} = \{k:\boldsymbol{\gamma}_{0(k)}=\mathbf{0}, k=1,\cdots,K\}$, $\mathcal{K}_{\neq 0} = \{k:\boldsymbol{\gamma}_{0(k)}\neq\mathbf{0}, k=1,\cdots,K\}$, and $C>0$ a generic positive constant whose value varies from one place to another.

To facilitate the proofs of all theorems we first present three lemmas. Lemmas 1 and 2 are Lemmas A1 and A2 in Newey and Smith (2004), and Lemma 3 is part of Inequality (A.5) in Newey and Smith (2004). Refer to Newey and Smith (2004) for proofs of these lemmas.

Lemma 1. *If Assumption 1 is satisfied, then for any ζ with $1/\alpha < \zeta \leq 1/2$ and $H_n = \left\{\boldsymbol{\rho}:\|\boldsymbol{\rho}\|\leq n^{-\zeta}\right\}$, $\sup_{(\boldsymbol{\beta},\boldsymbol{\gamma})\in\mathcal{B}\times\mathcal{T},\boldsymbol{\rho}\in H_n,1\leq i\leq n}|\boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta})-\boldsymbol{\gamma}]| \xrightarrow{p} 0$ and, with probability approaching one, $H_n \subseteq \hat{H}_n(\boldsymbol{\beta},\boldsymbol{\gamma})$ for all $(\boldsymbol{\beta},\boldsymbol{\gamma})\in\mathcal{B}\times\mathcal{T}$.*

*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA (Email: *yqzhai@umich.edu*).

†Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA (Email: *peisong@umich.edu*).

Lemma 2. *If Assumption 1 is satisfied,* $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}) \in \mathcal{B} \times \mathcal{T}$, $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}) \xrightarrow{p} (\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$, *and* $\hat{\boldsymbol{r}}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}) = O_p(n^{-1/2})$, *then* $\bar{\boldsymbol{\rho}} = \arg\max_{\boldsymbol{\rho} \in \hat{H}_n(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}})} \hat{Q}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}, \boldsymbol{\rho})$ *exists with probability approaching one,* $\bar{\boldsymbol{\rho}} = O_p(n^{-1/2})$, *and* $\sup_{\boldsymbol{\rho} \in \hat{H}_n(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}})} \hat{Q}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\gamma}}, \boldsymbol{\rho}) \leq O_p(n^{-1})$.

Lemma 3. *If Assumption 1 is satisfied, then for* ζ *in Lemma 1 we have* $n^{-\zeta} \|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\| - Cn^{-2\zeta} \leq \hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}})$.

Proof of Theorem 1

Proof. By the definition of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ we have

$$\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}) + \sum_{k=1}^{K} \hat{P}_{\lambda_n}\left(\hat{\boldsymbol{\gamma}}_{(k)}\right) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq \sup_{\boldsymbol{\rho} \in \hat{H}_n(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\rho}) + \sum_{k=1}^{K} \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right) - \hat{F}(\boldsymbol{\beta}_0). \tag{S1}$$

Also by definition we have $\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}) \geq \hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \mathbf{0}) = 0$ and the penalty function is non-negative. Therefore, from (S1) we have

$$\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq \sup_{\boldsymbol{\rho} \in \hat{H}_n(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\rho}) + \sum_{k=1}^{K} \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right). \tag{S2}$$

On the other hand, by Assumption 1(iv) and the Central Limit Theorem, we have $\|\hat{r}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\| = O_p(n^{-1/2})$, which leads to $\sup_{\boldsymbol{\rho} \in \hat{H}_n(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\rho}) \leq O_p(n^{-1})$ based on Lemma 2. For $k \in \mathcal{K}_{=0}$ we have $\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right) = 0$, and for $k \in \mathcal{K}_{\neq 0}$ we have $\hat{P}_{\lambda_n}(\boldsymbol{\gamma}_{0(k)}) = O_p(\lambda_n) = O_p(n^{-\xi})$ from Assumption 1(vii). Therefore from (S2) we have $\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) \leq O_p(n^{-1}) + O_p(n^{-\xi}) = O_p(n^{-\xi})$. In addition, from Assumption 1(vi) we have $\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) = F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) + O_p(n^{-1/2})$, and thus $F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) \leq O_p(n^{-\xi})$. On the other hand Assumption 1(ii) implies that $F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}}) \geq 0$. Hence, we must have

$$|F(\boldsymbol{\beta}_0) - F(\hat{\boldsymbol{\beta}})| = O_p(n^{-\xi}), \tag{S3}$$

which then implies $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$ in probability based on Assumptions 1(ii) and (iii).

Take ζ such that $1/\alpha < \zeta < \xi$. From Lemma 3, Equations (S1) and (S3), Assumptions 1(vi) and 1(vii) we have

$$\begin{aligned}
& n^{-\zeta}\|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\| - Cn^{-2\zeta} \\
\leq \quad & \sup_{\boldsymbol{\rho}\in\hat{H}_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)} \hat{Q}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\boldsymbol{\rho}) + \sum_{k=1}^{K} \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right) + \hat{F}(\hat{\boldsymbol{\beta}}) - \hat{F}(\boldsymbol{\beta}_0) \\
\leq \quad & O_p(n^{-1}) + O_p(\lambda_n) + |\hat{F}(\hat{\boldsymbol{\beta}}) - F(\hat{\boldsymbol{\beta}})| + |F(\hat{\boldsymbol{\beta}}) - F(\boldsymbol{\beta}_0)| + |F(\boldsymbol{\beta}_0) - \hat{F}(\boldsymbol{\beta}_0)| \\
= \quad & O_p(n^{-1}) + O_p(n^{-\xi}) + O_p(n^{-1/2}) + O_p(n^{-\xi}) + O_p(n^{-1/2}) \\
= \quad & O_p(n^{-\xi}),
\end{aligned}$$

which leads to $\|\hat{r}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\| \leq O_p(n^{\zeta-\xi}) + Cn^{-\zeta} = o_p(1)$. Thus, by Assumption 1(vi) and the consistency of $\hat{\boldsymbol{\beta}}$ we have

$$\begin{aligned}
\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| &= \left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)] \right\| \\
&\leq \left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\hat{\boldsymbol{\beta}})] \right\| + \left\| \mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\hat{\boldsymbol{\beta}})] - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0)] \right\| + \|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\| \\
&= o_p(1),
\end{aligned}$$

which implies $\hat{\boldsymbol{\gamma}} \to \boldsymbol{\gamma}_0$ in probability as $n \to \infty$. □

Proof of Theorem 2

Proof. From (S1) and the proof of Theorem 1 we have

$$\hat{Q}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\rho}}) + \hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) + \left\{ \sum_{k\in\mathcal{K}_{\neq 0}} \left[\hat{P}_{\lambda_n}(\hat{\boldsymbol{\gamma}}_{(k)}) - \hat{P}_{\lambda_n}(\boldsymbol{\gamma}_{0(k)}) \right] \right\} \leq O_p(n^{-1}). \tag{S4}$$

It is easy to check that, at any $\boldsymbol{\gamma}_{(k)} \neq \mathbf{0}$,

$$\frac{\partial \hat{P}_{\lambda_n}(\boldsymbol{\gamma}_{(k)})}{\partial \boldsymbol{\gamma}_{(k)}} = \frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w} \frac{\boldsymbol{\gamma}_{(k)}}{\|\boldsymbol{\gamma}_{(k)}\|}.$$

Therefore, by the mean value theorem, Cauchy-Schwarz inequality and Assumption 2(v) we have

$$\begin{aligned}
\left| \sum_{k \in \mathcal{K}_{\neq 0}} \left[\hat{P}_{\lambda_n}(\hat{\boldsymbol{\gamma}}_{(k)}) - \hat{P}_{\lambda_n}(\boldsymbol{\gamma}_{0(k)}) \right] \right| &= \left| \sum_{k \in \mathcal{K}_{\neq 0}} \left[\frac{\partial \hat{P}_{\lambda_n}(\dot{\boldsymbol{\gamma}}_{(k)})}{\partial \boldsymbol{\gamma}_{(k)}^T} (\hat{\boldsymbol{\gamma}}_{(k)} - \boldsymbol{\gamma}_{0(k)}) \right] \right| \\
&\leq K \max_{k \in \mathcal{K}_{\neq 0}} \left\| \frac{\partial \hat{P}_{\lambda_n}(\dot{\boldsymbol{\gamma}}_{(k)})}{\partial \boldsymbol{\gamma}_{(k)}} \right\| \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \\
&\leq K |\lambda_n| \max_{k \in \mathcal{K}_{\neq 0}} \left\{ \frac{1}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w} \right\} \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \\
&= o_p(n^{-1/2}) \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|, \qquad \text{(S5)}
\end{aligned}$$

where $\dot{\boldsymbol{\gamma}}_{(k)}$ is some value between $\hat{\boldsymbol{\gamma}}_{(k)}$ and $\boldsymbol{\gamma}_{0(k)}$. By the mean value theorem, Assumptions 1(ii) 2(iii) and the central limit theorem we have

$$\begin{aligned}
\hat{F}(\hat{\boldsymbol{\beta}}) &= \hat{F}(\boldsymbol{\beta}_0) + \frac{\partial \hat{F}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{1}{2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \frac{\partial^2 \hat{F}(\dot{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&= \hat{F}(\boldsymbol{\beta}_0) + O_p(n^{-1/2}) \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \frac{1}{2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \frac{\partial^2 \hat{F}(\dot{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \qquad \text{(S6)}
\end{aligned}$$

where $\dot{\boldsymbol{\beta}}$ is some value between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$. Then by Assumptions 1(ii) 2(iii)(iv) and the consistency of $\hat{\boldsymbol{\beta}}$ we have

$$\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}}) \geq C[1 + o_p(1)] \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 + O_p(n^{-1/2}) \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|. \qquad \text{(S7)}$$

Taking $\zeta = 1/2$ in Lemma 3 leads to

$$\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}) \geq n^{-1/2} \|\hat{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\| - C n^{-1}. \qquad \text{(S8)}$$

Then by Assumptions 1(vi), 2(ii) and the triangle inequality we have

$$\begin{aligned}
& \hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}) \\
\geq\ & n^{-1/2} \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}})] + \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \hat{\boldsymbol{\beta}})] - \hat{\boldsymbol{\gamma}} + \boldsymbol{\gamma}_0 - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)] \right\| - Cn^{-1} \\
\geq\ & n^{-1/2} \left\{ \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| - |O_p(n^{-1/2})| - C[1 + o_p(1)] \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \right\} - Cn^{-1}. \qquad \text{(S9)}
\end{aligned}$$

From (S4), (S5), (S7) and (S9) we have

$$C[1 + o_p(1)] \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 + O_p(n^{-\frac{1}{2}}) \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + n^{-\frac{1}{2}} [1 + o_p(1)] \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \leq O_p(n^{-1}). \quad \text{(S10)}$$

If $\hat{\boldsymbol{\beta}}$ has a faster convergence rate than $\hat{\boldsymbol{\gamma}}$, then (S10) becomes $C[1 + o_p(1)] \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 + n^{-\frac{1}{2}} [1 + o_p(1)] \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \leq O_p(n^{-1})$, which implies that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$ and $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = O_p(n^{-1/2})$. If $\hat{\boldsymbol{\beta}}$ has the same or slower convergence rate than $\hat{\boldsymbol{\gamma}}$, then (S10) becomes $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 + O_p(n^{-\frac{1}{2}}) \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq O_p(n^{-1})$, which implies that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq O_p(n^{-1/2})$ from the property of quadratic functions, and thus $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$. Since $\hat{\boldsymbol{\beta}}$ has the same or slower convergence rate than $\hat{\boldsymbol{\gamma}}$, we must also have $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = O_p(n^{-1/2})$. This proves results (i) and (ii).

Based on (i), from (S6) we have $|\hat{F}(\boldsymbol{\beta}_0) - \hat{F}(\hat{\boldsymbol{\beta}})| = O_p(n^{-1})$. Then (S4) and (S5) imply that $\hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}}) \leq O_p(n^{-1})$, and then from (S8) we have $\|\hat{r}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\| = O_p(n^{-1/2})$. Therefore result (iii) directly follows from Lemma 2. □

Proof of Theorem 3

Proof. On the event $\{\hat{\boldsymbol{\gamma}}_{(k)} \neq \mathbf{0}\}$ for some $k \in \mathcal{K}_{=0}$, the KKT optimality condition

$$\frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w} \frac{\hat{\boldsymbol{\gamma}}_{(k)}}{\|\hat{\boldsymbol{\gamma}}_{(k)}\|} + \frac{\hat{\boldsymbol{\rho}}_{(k)}}{n} \sum_{i=1}^{n} \frac{1}{1 - \hat{\boldsymbol{\rho}}^T [\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\gamma}}]} = \mathbf{0}$$

implies that

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{1-\hat{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{\gamma}}]}\right|\|\hat{\boldsymbol{\rho}}_{(k)}\| = \left\|\sqrt{n}\frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}\frac{\hat{\boldsymbol{\gamma}}_{(k)}}{\|\hat{\boldsymbol{\gamma}}_{(k)}\|}\right\|.$$

Based on $\hat{\boldsymbol{\rho}} = O_p(n^{-1/2})$ from Theorem 2 and Assumption 1(iv) we have

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{1-\hat{\boldsymbol{\rho}}^T[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{\gamma}}]}\right|\|\hat{\boldsymbol{\rho}}_{(k)}\| = O_p(1).$$

On the other hand, by Assumption 3 and the $\sqrt{n}$-consistency of $\tilde{\boldsymbol{\gamma}}_{(k)}$, we have

$$\lim_{n\to\infty}\left\|\sqrt{n}\frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}\frac{\hat{\boldsymbol{\gamma}}_{(k)}}{\|\hat{\boldsymbol{\gamma}}_{(k)}\|}\right\| = \lim_{n\to\infty}\left|\sqrt{n}\frac{\lambda_n}{\|\tilde{\boldsymbol{\gamma}}_{(k)}\|^w}\right| = \infty$$

for any $k \in \mathcal{K}_{=0}$. Therefore, we must have $P(\hat{\boldsymbol{\gamma}}_{(k)} = \mathbf{0}) \to 1$ as $n \to \infty$ for any $k \in \mathcal{K}_{=0}$. This, together with the consistency of $\hat{\boldsymbol{\gamma}}$, implies the desired result. □

Proof of Theorem 4

Proof. For any compact set $\mathcal{H} \subset \mathbb{R}^{\dim(\boldsymbol{\eta})}$, denote $\boldsymbol{u}_{\boldsymbol{\eta}} \in \mathcal{H}$ as $\boldsymbol{u}_{\boldsymbol{\eta}}^T = (\boldsymbol{u}_{\boldsymbol{\beta}}^T, \boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T)$, where $\boldsymbol{u}_{\boldsymbol{\beta}}$ contains the first q elements in $\boldsymbol{u}_{\boldsymbol{\eta}}$ and $\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}$ contains the rest elements in $\boldsymbol{u}_{\boldsymbol{\eta}}$. On this compact set $\mathcal{H}$ define

$$\begin{aligned}
L(\boldsymbol{u}_{\boldsymbol{\eta}}) = &-\sum_{i=1}^{n}\log f_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right) + \sum_{i=1}^{n}\log f_i(\boldsymbol{\beta}_0) \\
&+ \max_{\boldsymbol{\rho}}\sum_{i=1}^{n}\log\left\{1 - \boldsymbol{\rho}^T\left[\boldsymbol{g}_i\left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right) - \left(\boldsymbol{\gamma}_{0,=0}^T, \boldsymbol{\gamma}_{0,\neq 0}^T + \frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}}\right)^T\right]\right\} \\
&+ n\sum_{k\in\mathcal{K}_{\neq 0}}\left[\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)} + \frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0,(k)}}{\sqrt{n}}\right) - \hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right)\right].
\end{aligned}$$

From Theorem 3, we know that $\hat{\boldsymbol{\gamma}}_{=0} = \mathbf{0}$ with probability approaching one. Thus, $\sqrt{n}(\hat{\boldsymbol{\eta}} -$

$\boldsymbol{\eta}_0$) is the minimizer of $L(\boldsymbol{u}_{\boldsymbol{\eta}})$ on $\mathcal{H}$ with probability approaching one.

From (S5) we have

$$n\left|\sum_{k\in\mathcal{K}_{\neq 0}}\left[\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0,(k)}}{\sqrt{n}}\right)-\hat{P}_{\lambda_n}\left(\boldsymbol{\gamma}_{0(k)}\right)\right]\right|\leq n|o_p(n^{-1/2})|\left\|\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}}{\sqrt{n}}\right\|=o_p(1). \quad \text{(S11)}$$

uniformly on $\mathcal{H}$.

By Assumptions 1(vi) and 2(ii), we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)=&\left\{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)\right]\right\}\\&+\left\{\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)\right]-\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0\right)\right]\right\}+\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0\right)\right]\\=&O_p(n^{-1/2})+\boldsymbol{\gamma}_{\mathbf{0}},\end{aligned}$$

uniformly on $\mathcal{H}$, which implies that

$$\begin{aligned}\hat{\boldsymbol{r}}\left[\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}},\left(\boldsymbol{\gamma}_{0,=0}^T,\boldsymbol{\gamma}_{0,\neq 0}^T+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}}\right)^T\right]&=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\left(\boldsymbol{\gamma}_{0,=0}^T,\boldsymbol{\gamma}_{0,\neq 0}^T+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}}\right)^T\\&=O_p(n^{-1/2})\end{aligned}$$

uniformly on $\mathcal{H}$. Thus, by Lemma 2,

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}=\arg\max_{\boldsymbol{\rho}}\sum_{i=1}^{n}\log\left\{1-\boldsymbol{\rho}^T\left[\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\left(\boldsymbol{\gamma}_{0,=0}^T,\boldsymbol{\gamma}_{0,\neq 0}^T+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}}\right)^T\right]\right\}$$

exists with probability approaching one and $\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}=O_p(n^{-1/2})$, uniformly on $\mathcal{H}$. Denote

$$\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)=\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\left(\boldsymbol{\gamma}_{0,=0}^T,\boldsymbol{\gamma}_{0,\neq 0}^T+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}}\right)^T.$$

It is clear that $\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}$ must satisfy

$$\sum_{i=1}^{n} \frac{\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)}{1-\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T \boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)} = \mathbf{0}.$$

Then the mean value theorem leads to

$$\mathbf{0} = \sum_{i=1}^{n} \boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right) + \sum_{i=1}^{n} \frac{\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)^T}{\left\{1-\dot{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T \boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)\right\}^2}\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}},$$

where $\dot{\boldsymbol{\rho}}_{\boldsymbol{\eta}}$ is some value between $\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}$ and $\mathbf{0}$. Then we have

$$\sqrt{n}\hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}} = -\boldsymbol{\Omega}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)\right\} + o_p(1),$$

uniformly on $\mathcal{H}$. On the other hand, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{r}_i\left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}}\right)\\
&=\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{g}_i\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)-\sqrt{n}\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0+\frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}}\right)\right]\right\}+\frac{\partial\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0\right)\right]}{\partial\boldsymbol{\beta}}\boldsymbol{u}_{\boldsymbol{\beta}}+o_p(1)\\
&\quad+\sqrt{n}\boldsymbol{\gamma}_0-\sqrt{n}\left(\boldsymbol{\gamma}_{0,=0}^T,\boldsymbol{\gamma}_{0,\neq 0}^T+\frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}}\right)^T\\
&=\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{g}_i\left(\boldsymbol{\beta}_0\right)-\sqrt{n}\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{\beta}_0\right)\right]\right\}+\boldsymbol{G}_{\boldsymbol{\eta}}\boldsymbol{u}_{\boldsymbol{\eta}}+o_p(1)\\
&\xrightarrow{d}\boldsymbol{\psi}+\boldsymbol{G}_{\boldsymbol{\eta}}\boldsymbol{u}_{\boldsymbol{\eta}},
\end{aligned}$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\eta}} \in \mathcal{H}$, where $\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$. Then the mean value theorem gives

$$
\begin{aligned}
&\sum_{i=1}^{n} \log \left\{ 1 - \hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T \left[\boldsymbol{g}_i \left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}} \right) - \left(\boldsymbol{\gamma}_{0,=0}^T, \boldsymbol{\gamma}_{0,\neq 0}^T + \frac{\boldsymbol{u}_{\boldsymbol{\gamma},\neq 0}^T}{\sqrt{n}} \right)^T \right] \right\} \\
&= - \sqrt{n} \hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{r}_i \left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}} \right) - \frac{1}{2} \sqrt{n} \hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{r}_i \left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}} \right) \boldsymbol{r}_i \left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}} \right)^T}{\left[1 - \dot{\boldsymbol{\rho}}_{\boldsymbol{\eta}}^T \boldsymbol{r}_i \left(\frac{\boldsymbol{u}_{\boldsymbol{\eta}}}{\sqrt{n}} \right) \right]^2} \right\} \sqrt{n} \hat{\boldsymbol{\rho}}_{\boldsymbol{\eta}} \\
&\xrightarrow{d} \frac{1}{2} \left\{ \boldsymbol{\psi} + \boldsymbol{G}_{\boldsymbol{\eta}} \boldsymbol{u}_{\boldsymbol{\eta}} \right\}^T \boldsymbol{\Omega}^{-1} \left\{ \boldsymbol{\psi} + \boldsymbol{G}_{\boldsymbol{\eta}} \boldsymbol{u}_{\boldsymbol{\eta}} \right\}
\end{aligned} \tag{S12}
$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\eta}} \in \mathcal{H}$.

By the mean value theorem we have

$$
\begin{aligned}
&\sum_{i=1}^{n} \log f_i \left(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}_{\boldsymbol{\beta}}}{\sqrt{n}} \right) - \sum_{i=1}^{n} \log f_i(\boldsymbol{\beta}_0) \\
&= \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta}_0)^T \right\} \boldsymbol{u}_{\boldsymbol{\beta}} + \frac{1}{2} \boldsymbol{u}_{\boldsymbol{\beta}}^T \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \boldsymbol{s}_i(\dot{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \right\} \boldsymbol{u}_{\boldsymbol{\beta}} \xrightarrow{d} \boldsymbol{u}_{\boldsymbol{\beta}}^T \boldsymbol{\phi} - \frac{1}{2} \boldsymbol{u}_{\boldsymbol{\beta}}^T \boldsymbol{S}_0 \boldsymbol{u}_{\boldsymbol{\beta}},
\end{aligned} \tag{S13}
$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\eta}} \in \mathcal{H}$, where $\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}_0)$.

Therefore, from (S11)-(S13) we have

$$
\begin{aligned}
L(\boldsymbol{u}_{\boldsymbol{\eta}}) \xrightarrow{d} L^*(\boldsymbol{u}_{\boldsymbol{\eta}}) &\equiv -\boldsymbol{u}_{\boldsymbol{\beta}}^T \boldsymbol{\phi} + \frac{1}{2} \boldsymbol{u}_{\boldsymbol{\beta}}^T \boldsymbol{S}_0 \boldsymbol{u}_{\boldsymbol{\beta}} + \frac{1}{2} (\boldsymbol{\psi} + \boldsymbol{G}_{\boldsymbol{\eta}} \boldsymbol{u}_{\boldsymbol{\eta}})^T \boldsymbol{\Omega}^{-1} (\boldsymbol{\psi} + \boldsymbol{G}_{\boldsymbol{\eta}} \boldsymbol{u}_{\boldsymbol{\eta}}) \\
&= \frac{1}{2} \boldsymbol{u}_{\boldsymbol{\eta}}^T (\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}^{-1} \boldsymbol{G}_{\boldsymbol{\eta}}) \boldsymbol{u}_{\boldsymbol{\eta}} + \boldsymbol{u}_{\boldsymbol{\eta}}^T \left\{ \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\psi} - \begin{bmatrix} \boldsymbol{\phi} \\ \boldsymbol{0} \end{bmatrix} \right\} + \frac{1}{2} \boldsymbol{\psi}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\psi},
\end{aligned}
$$

uniformly over $\boldsymbol{u}_{\boldsymbol{\eta}} \in \mathcal{H}$, and $L^*(\boldsymbol{u}_{\boldsymbol{\eta}})$ is uniquely minimized at

$$
\boldsymbol{u}_{\boldsymbol{\eta}}^* = -(\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}^{-1} \boldsymbol{G}_{\boldsymbol{\eta}})^{-1} \left\{ \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\psi} - \begin{bmatrix} \boldsymbol{\phi} \\ \boldsymbol{0} \end{bmatrix} \right\}.
$$

It is easy to see that $\boldsymbol{u}_{\boldsymbol{\eta}}^* \sim \mathcal{N}\left(\mathbf{0}, (\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}^{-1} \boldsymbol{G}_{\boldsymbol{\eta}})^{-1}\right)$, based on the fact that $\mathbb{E}(\boldsymbol{\psi}\boldsymbol{\phi}^T) = \mathbb{E}\{\mathbb{E}[(\boldsymbol{g}(\boldsymbol{\beta_0}) - \boldsymbol{\gamma}_0)\boldsymbol{s}(\boldsymbol{\beta_0})^T \mid \boldsymbol{X}, \boldsymbol{Z}]\} = \mathbf{0}$. Then from the Continuous Mapping Theorem we have $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{d} \boldsymbol{u}_{\boldsymbol{\eta}}^*$, which completes the proof. □

Proof of Theorem 5

Proof. Denote

$$\boldsymbol{G}_{\boldsymbol{\eta}} = \begin{bmatrix} \boldsymbol{G}_0 & \mathbf{0} \\ \boldsymbol{G}_{\neq 0} & -\mathcal{I} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\Omega}^{-1} = \begin{bmatrix} \boldsymbol{W}_{11} & \boldsymbol{W}_{12} \\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22} \end{bmatrix},$$

where $\boldsymbol{G}_{\neq 0} = \mathbb{E}[\partial \boldsymbol{g}_{\neq 0}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}]$, $\boldsymbol{W}_{11}$ is the leading $\dim(\boldsymbol{\gamma}_{\neq 0}) \times \dim(\boldsymbol{\gamma}_{\neq 0})$ sub-matrix of $\boldsymbol{\Omega}^{-1}$, and $\boldsymbol{W}_{12}$, $\boldsymbol{W}_{21}$, and $\boldsymbol{W}_{22}$ are the other corresponding sub-matrices of $\boldsymbol{\Omega}^{-1}$. Then we have

$$\boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}^{-1} \boldsymbol{G}_{\boldsymbol{\eta}} = \begin{bmatrix} \boldsymbol{G}_0^T \boldsymbol{W}_{11} \boldsymbol{G}_0 + \boldsymbol{G}_{\neq 0}^T \boldsymbol{W}_{21} \boldsymbol{G}_0 + \boldsymbol{G}_0^T \boldsymbol{W}_{12} \boldsymbol{G}_{\neq 0} + \boldsymbol{G}_{\neq 0}^T \boldsymbol{W}_{22} \boldsymbol{G}_{\neq 0} & -\boldsymbol{G}_0^T \boldsymbol{W}_{12} - \boldsymbol{G}_{\neq 0}^T \boldsymbol{W}_{22} \\ -\boldsymbol{W}_{21} \boldsymbol{G}_0 - \boldsymbol{W}_{22} \boldsymbol{G}_{\neq 0} & \boldsymbol{W}_{22} \end{bmatrix},$$

Therefore, the inverse of the leading $q \times q$ sub-matrix of $(\boldsymbol{S} + \boldsymbol{G}_{\boldsymbol{\eta}}^T \boldsymbol{\Omega}^{-1} \boldsymbol{G}_{\boldsymbol{\eta}})^{-1}$ is

$$\begin{aligned} &(\boldsymbol{S}_0 + \boldsymbol{G}_0^T \boldsymbol{W}_{11} \boldsymbol{G}_0 + \boldsymbol{G}_{\neq 0}^T \boldsymbol{W}_{21} \boldsymbol{G}_0 + \boldsymbol{G}_0^T \boldsymbol{W}_{12} \boldsymbol{G}_{\neq 0} + \boldsymbol{G}_{\neq 0}^T \boldsymbol{W}_{22} \boldsymbol{G}_{\neq 0}) \\ &- (\boldsymbol{G}_0^T \boldsymbol{W}_{12} + \boldsymbol{G}_{\neq 0}^T \boldsymbol{W}_{22}) \boldsymbol{W}_{22}^{-1} (\boldsymbol{W}_{21} \boldsymbol{G}_0 + \boldsymbol{W}_{22} \boldsymbol{G}_{\neq 0}) \\ =& \boldsymbol{S}_0 + \boldsymbol{G}_0^T \boldsymbol{W}_{11} \boldsymbol{G}_0 - \boldsymbol{G}_0^T \boldsymbol{W}_{12} \boldsymbol{W}_{22}^{-1} \boldsymbol{W}_{21} \boldsymbol{G}_0 \\ =& \boldsymbol{S}_0 + \boldsymbol{G}_0^T \boldsymbol{W}_{11} \boldsymbol{G}_0 - \boldsymbol{G}_0^T (\boldsymbol{W}_{11} - \boldsymbol{\Omega}_0^{-1}) \boldsymbol{G}_0 \\ =& \boldsymbol{S}_0 + \boldsymbol{G}_0^T \boldsymbol{\Omega}_0^{-1} \boldsymbol{G}_0, \end{aligned}$$

where the second equality follows from the fact that the block matrix inverse of $\boldsymbol{\Omega}^{-1}$ leads to $\boldsymbol{\Omega}_0 = (\boldsymbol{W}_{11} - \boldsymbol{W}_{12} \boldsymbol{W}_{22}^{-1} \boldsymbol{W}_{21})^{-1}$. This completes the proof. □

EQUATION (4) UNDER THE SIMULATION SETTING

We provide the specific expressions for Equation (4) under the simulation setting in Section 5 of the article. For the logistic regression model for the internal study, we have

$$f(Y|\boldsymbol{X}, Z; \boldsymbol{\beta}) = P(Y = 1|\boldsymbol{X}, Z; \boldsymbol{\beta})^Y [1 - P(Y = 1|\boldsymbol{X}, Z; \boldsymbol{\beta})]^{1-Y}$$

with $P(Y = 1|\boldsymbol{X}, Z; \boldsymbol{\beta}) = \text{expit}[(1, X_1, X_2, X_3, Z, X_1Z)\boldsymbol{\beta}]$, where $\text{expit}(x) = e^x/(1 + e^x)$. For the two external studies, since both studies fitted logistic regression models, we have the corresponding score functions

$$\begin{aligned}
\boldsymbol{h}_{(1)}(Y, \boldsymbol{X}_{(1)}; \boldsymbol{\theta}_{(1)}) &= (1, X_2, X_3)^T [Y - \text{expit}(\theta_{(1)c} + \theta_{(1)1}X_2 + \theta_{(1)2}X_3)] \\
\boldsymbol{h}_{(2)}(Y, \boldsymbol{X}_{(2)}; \boldsymbol{\theta}_{(2)}) &= (1, X_1, X_2)^T [Y - \text{expit}(\theta_{(2)c} + \theta_{(2)1}X_1 + \theta_{(2)2}X_2)].
\end{aligned}$$

Then the $\boldsymbol{g}(\boldsymbol{X}, Z; \boldsymbol{\beta})$ in Equation (4) is given by $[\boldsymbol{U}_{(1)}(\boldsymbol{X}, Z; \boldsymbol{\beta}, \boldsymbol{\theta}^*_{(1)})^T, \boldsymbol{U}_{(2)}(\boldsymbol{X}, Z; \boldsymbol{\beta}, \boldsymbol{\theta}^*_{(2)})^T]^T$, where

$$\begin{aligned}
& \boldsymbol{U}_{(1)}(\boldsymbol{X}, Z; \boldsymbol{\beta}, \boldsymbol{\theta}^*_{(1)}) \\
= \; & \mathbb{E}[\boldsymbol{h}_{(1)}(Y, \boldsymbol{X}_{(1)}; \boldsymbol{\theta}^*_{(1)})|\boldsymbol{X}, Z] \\
= \; & (1, X_2, X_3)^T [\mathbb{E}(Y|\boldsymbol{X}, Z) - \text{expit}(\theta_{(1)c} + \theta_{(1)1}X_2 + \theta_{(1)2}X_3)] \\
= \; & (1, X_2, X_3)^T \{\text{expit}[(1, X_1, X_2, X_3, Z, X_1Z)\boldsymbol{\beta}] - \text{expit}(\theta_{(1)c} + \theta_{(1)1}X_2 + \theta_{(1)2}X_3)\},
\end{aligned}$$

and similarly $\boldsymbol{U}_{(2)}(\boldsymbol{X}, Z; \boldsymbol{\beta}, \boldsymbol{\theta}^*_{(2)}) = (1, X_1, X_2)^T \{\text{expit}[(1, X_1, X_2, X_3, Z, X_1Z)\boldsymbol{\beta}] - \text{expit}(\theta_{(2)c} + \theta_{(2)1}X_1 + \theta_{(2)2}X_2)\}$.

SOME ADDITIONAL STIMULATION RESULTS

Table S1: Simulation results summarized based on 1000 replications with external study sample size 3000.

		Internal sample size $n = 300$						Internal sample size $n = 800$					
		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_Z	β_{X_1Z}	β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_Z	β_{X_1Z}
	Bias	-0.024	0.022	-0.037	0.036	0.021	-0.022	-0.008	0.011	-0.012	0.010	0.007	-0.007
MLE	ESE	0.268	0.259	0.333	0.254	0.166	0.117	0.160	0.148	0.193	0.153	0.093	0.070
	RMSE	0.269	0.260	0.335	0.256	0.167	0.119	0.161	0.148	0.193	0.153	0.093	0.071
	Bias	-0.002	0.023	-0.028	0.011	0.021	-0.023	-0.001	0.011	-0.011	0.007	0.007	-0.008
CML-1	ESE	0.138	0.258	0.179	0.175	0.166	0.118	0.102	0.148	0.125	0.113	0.093	0.070
	RMSE	0.138	0.259	0.181	0.175	0.167	0.120	0.102	0.148	0.125	0.113	0.093	0.071
	Bias	0.001	0.023	-0.033	0.011	0.021	-0.022	-0.001	0.011	-0.008	0.005	0.007	-0.008
PCMLg-1	ESE	0.156	0.258	0.211	0.187	0.166	0.117	0.106	0.148	0.130	0.116	0.093	0.070
	RMSE	0.156	0.259	0.214	0.187	0.167	0.120	0.106	0.148	0.130	0.116	0.093	0.071
	Bias	-0.000	0.023	-0.029	0.010	0.021	-0.022	0.000	0.011	-0.010	0.005	0.007	-0.008
PCMLc-1	ESE	0.142	0.258	0.182	0.177	0.166	0.118	0.105	0.148	0.125	0.115	0.093	0.070
	RMSE	0.142	0.259	0.184	0.177	0.167	0.120	0.105	0.148	0.125	0.115	0.093	0.071
	Bias	0.597	-0.005	0.141	0.037	0.022	-0.023	0.618	-0.009	0.158	0.010	0.007	-0.009
CML-2	ESE	0.204	0.220	0.215	0.254	0.166	0.117	0.140	0.129	0.145	0.153	0.093	0.070
	RMSE	0.631	0.220	0.257	0.256	0.167	0.119	0.634	0.129	0.214	0.153	0.093	0.071
	Bias	-0.026	0.030	-0.035	0.036	0.021	-0.023	-0.009	0.023	-0.011	0.010	0.007	-0.008
CML-2o	ESE	0.267	0.231	0.333	0.254	0.166	0.117	0.160	0.134	0.193	0.153	0.093	0.070
	RMSE	0.269	0.233	0.335	0.256	0.167	0.120	0.161	0.136	0.193	0.153	0.093	0.071
	Bias	-0.024	0.022	-0.037	0.036	0.021	-0.022	-0.008	0.011	-0.012	0.010	0.007	-0.007
PCMLg-2	ESE	0.268	0.259	0.333	0.254	0.166	0.117	0.160	0.148	0.193	0.153	0.093	0.070
	RMSE	0.269	0.260	0.335	0.256	0.167	0.119	0.161	0.148	0.193	0.153	0.093	0.071
	Bias	-0.023	0.011	0.105	0.036	0.021	-0.023	-0.009	0.022	-0.009	0.010	0.007	-0.008
PCMLc-2	ESE	0.272	0.238	0.480	0.254	0.166	0.117	0.160	0.135	0.198	0.153	0.093	0.070
	RMSE	0.273	0.238	0.491	0.256	0.167	0.120	0.161	0.136	0.198	0.153	0.093	0.071
	Bias	0.123	0.151	-0.170	-0.002	0.024	-0.015	0.124	0.150	-0.178	0.015	0.011	-0.003
CML-12	ESE	0.165	0.245	0.215	0.187	0.166	0.118	0.121	0.146	0.155	0.127	0.093	0.070
	RMSE	0.206	0.288	0.275	0.187	0.168	0.119	0.173	0.209	0.236	0.127	0.094	0.070
	Bias	-0.001	0.030	-0.031	0.011	0.021	-0.023	0.002	0.024	-0.019	0.007	0.007	-0.008
CML-12o	ESE	0.133	0.231	0.163	0.175	0.166	0.118	0.101	0.134	0.123	0.113	0.093	0.070
	RMSE	0.133	0.233	0.166	0.175	0.167	0.120	0.101	0.136	0.125	0.113	0.093	0.071
	Bias	0.005	0.023	-0.033	0.011	0.021	-0.022	-0.001	0.011	-0.007	0.005	0.007	-0.008
PCMLg-12	ESE	0.154	0.258	0.212	0.185	0.166	0.117	0.106	0.148	0.129	0.116	0.093	0.070
	RMSE	0.154	0.259	0.215	0.185	0.167	0.120	0.106	0.148	0.130	0.116	0.093	0.071
	Bias	-0.001	0.030	-0.021	0.007	0.021	-0.023	0.004	0.024	-0.017	0.005	0.007	-0.008
PCMLc-12	ESE	0.145	0.232	0.210	0.178	0.166	0.118	0.105	0.134	0.125	0.115	0.093	0.070
	RMSE	0.145	0.234	0.211	0.178	0.167	0.120	0.105	0.136	0.126	0.115	0.093	0.071

[1] ESE: empirical standard error. RMSE: root mean squared error. MLE: maximum likelihood estimator using internal study data alone. CML: constrained maximum likelihood. PCMLg, PCMLc: penalized constrained maximum likelihood estimators with group-wise and component-wise shrinkage, respectively. -1, -2, -2o, -12, -12o: with external study 1 only, with external study 2 only, with the second moment constraint from external study 2 only, with both external studies, with external study 1 and the second moment constraint from external study 2, respectively.

REFERENCES

Newey, W. K., & Smith, R. J. (2004), "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255.