BIOMETRICS 64, 1–6 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Supporting Information for "A Robust Approach for Electronic Health Record-Based Case-Control Studies with Contaminated Case Pools" by Guorong Dai, Yanyuan Ma, Jill Hasler, Jinbo Chen and Raymond J. Carroll

Guorong Dai[1,*], Yanyuan Ma[2], Jill Hasler[3], Jinbo Chen[3] and Raymond J. Carroll[4]

[1]Department of Statistics and Data Science, School of Management, Fudan University, Shanghai 200433, China

[2]Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA

[3]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[4]Department of Statistics, Texas A&M University, College Station, TX 77843, USA

**email:* guorongdai@fudan.edu.cn

SUMMARY: We collect in the following (i) all the technical details that cannot be accommodated in the main manuscript, including the regularity conditions, an auxiliary lemma and the proofs of all the theoretical results in Section 3, and (ii) some necessary supplements for the numerical studies in Sections 4–5

Web Appendix A: Regularity Conditions

ASSUMPTION 1: The population-level equation $\mathbf{\Phi}(\mathbf{b}) = \mathbf{0}$ has a unique solution. The matrices $\mathbf{A}_{\mathrm{V}}$, $\mathbf{A}_{\mathrm{U}}$, $\mathbf{A}_0$, $\mathbf{B}_{\mathrm{V}}$, $\mathbf{B}_{\mathrm{U}}$ and $\mathbf{B}_0$ exist, while $\mathbf{A}$ is invertible. The function $\mathbf{\Psi}_n(\cdot,\cdot)$ satisfies

$$\sup_{\mathbf{b}\in\mathcal{B}}\|n^{-1}\textstyle\sum_{i=1}^{N}\mathbf{\Psi}_n(\mathbf{W}_i,\mathbf{b}) - \mathbf{\Phi}(\mathbf{b})\| = o_p(1). \tag{S.1}$$

In addition, the partial derivative $\mathbf{\Omega}'_{n,j}(\mathbf{b}) \equiv \partial\mathbf{\Omega}_n(\mathbf{b})/\partial b_j$ is such that

$$\sup_{\mathbf{b}\in\mathcal{B}_0} \lambda_{\max}\{\mathbf{\Omega}'_{n,j}(\mathbf{b})\} = O_p(n) \quad (j = 1,\ldots,2p), \tag{S.2}$$

where $\lambda_{\max}(\cdot)$ denotes the largest absolute eigenvalue of a matrix, $\mathbf{b} \equiv (b_1,\ldots,b_{2p})^{\mathrm{T}}$ and $\mathcal{B}_0$ is some open neighborhood of $\boldsymbol{\beta}$.

Assumption 1 is quite standard for an estimator defined as the solution to an equation. In our setting, considering the fact that $n^{-1}\sum_{i=1}^{N}E\{\mathbf{\Psi}_n(\mathbf{W}_i,\mathbf{b})\} = \mathbf{\Phi}(\mathbf{b})$ because of the case-control sampling, the high-level condition (S.1) of uniform convergence holds, for example, if the parameter space $\mathcal{B}$ is compact; see Example 19.7 and the discussion after Theorem 5.9 in Van der Vaart (2000). The condition (S.2) is needed to control the residual in Taylor's expansion of (16). In general, Assumption 1 imposes some fairly mild requirements to regulate the behavior of the estimating equation (16).

Web Appendix B: An Auxiliary Lemma

LEMMA 1: *For some fixed integer $M > 0$, suppose $\mathbf{U}_{n,1},\ldots,\mathbf{U}_{n,M} \in \mathbb{R}^p$ are mutually independent sequences of random vectors satisfying that, for some p-dimensional vector $\boldsymbol{\mu}_m$ and $p\times p$ covariance matrix $\mathbf{\Sigma}_m$,*

$$\mathbf{U}_{n,m} \to \mathbf{N}(\boldsymbol{\mu}_m, \mathbf{\Sigma}_m) \quad (n\to\infty,\ m = 1,\ldots,M) \tag{S.3}$$

in distribution. Then $\sum_{m=1}^{M}\mathbf{U}_{n,m} \to \mathbf{N}(\sum_{m=1}^{M}\boldsymbol{\mu}_m, \sum_{m=1}^{M}\mathbf{\Sigma}_m)$ in distribution as $n\to\infty$.

Web Appendix C: Proof of Lemma 1

We have that, for any $\mathbf{t} \in \mathbb{R}^p$,

$$\begin{aligned} E\{\exp(i\mathbf{t}^{\mathrm{T}}\textstyle\sum_{m=1}^{M}\mathbf{U}_{n,m})\} &= \textstyle\prod_{m=1}^{M} E\{\exp(i\mathbf{t}^{\mathrm{T}}\mathbf{U}_{n,m})\} \\ &\to \textstyle\prod_{m=1}^{M} \exp(i\boldsymbol{\mu}_m^{\mathrm{T}}\mathbf{t} - \mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}_m\mathbf{t}/2) \\ &= \exp\{i(\textstyle\sum_{m=1}^{M}\boldsymbol{\mu}_m^{\mathrm{T}})\mathbf{t} - \mathbf{t}^{\mathrm{T}}(\textstyle\sum_{m=1}^{M}\boldsymbol{\Sigma}_m)\mathbf{t}/2\} \qquad \text{(S.4)} \end{aligned}$$

where i is the imaginary unit. In the above, the first step uses the mutual independence of $\mathbf{U}_{n,1}, \ldots, \mathbf{U}_{n,M}$, while the second step is due to (S.3) and Levy's continuity theorem. The fact that (S.4) is the characteristic function of $\mathbf{N}(\sum_{m=1}^{M}\boldsymbol{\mu}_m, \sum_{m=1}^{M}\boldsymbol{\Sigma}_m)$ implies the conclusion.

Web Appendix D: Proof of Proposition 1

We have

$$\begin{aligned} \textstyle\sum_{i=1}^{N} E[S_i\{g(\mathbf{X}_i) - D_i\}\mathbf{X}_i\overline{h}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i)] &= \tau N E[\{g(\mathbf{X}) - D\}\mathbf{X}\overline{h}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}) \mid S = 1] \\ &= \eta^{-1}\tau N E[S\{g(\mathbf{X}) - D\}\mathbf{X}\overline{h}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X})] = \mathbf{0}, \quad \text{(S.5)} \end{aligned}$$

where the first step is due to case-control sampling and the last step holds by the condition (12). It follows that

$$\begin{aligned} &\textstyle\sum_{i=1}^{N} E\{S_i g(\mathbf{X}_i)\mathbf{X}_i\overline{h}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i) - (1 - S_i)\mathbf{X}_i h(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i)\} \\ &\quad = \textstyle\sum_{i=1}^{N} E\{S_i D_i\mathbf{X}_i\overline{h}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i) - (1 - S_i)\mathbf{X}_i h(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i)\} \\ &\quad = N[\tau E\{D\mathbf{X}\overline{h}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}) \mid S = 1\} - (1 - \tau)E\{\mathbf{X}h(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}) \mid S = 0\}] = \mathbf{0}, \end{aligned}$$

where the second step is due to the case-control sampling and the last step is because of the definition (3) of $\boldsymbol{\theta}$. This verifies (13). Then (14) is from (S.5) and the fact that the validation set $\{\mathbf{W}_i : R_i S_i = 1, i = 1, \ldots, N\}$ can be viewed as a random sample drawn from the conditional distribution of $(\mathbf{X}^{\mathrm{T}}, D, S)^{\mathrm{T}}$ given $S = 1$.

Web Appendix E: Proof of Theorem 1

Write $N_{\rm U} \equiv \sum_{i=1}^{N}(1-R_i)S_i$, $\mathcal{V} \equiv \{i : R_iS_i = 1, i = 1,\ldots,N\}$, $\mathcal{U} \equiv \{i : (1-R_i)S_i = 1, i = 1,\ldots,N\}$ and $\mathcal{C} \equiv \{i : S_i = 0, i = 1,\ldots,N\}$. Due to case-control sampling and the fact that the validation set can be viewed as a random sample drawn from the candidate case pool, the sets $\{\mathbf{W}_i : i \in \mathcal{V}\}$, $\{\mathbf{W}_i : i \in \mathcal{U}\}$ and $\{\mathbf{W}_i : i \in \mathcal{C}\}$ can actually be treated as mutually independent samples randomly selected from $P(\mathbf{X}, D, S \mid S = 1)$, $P(\mathbf{X}, S \mid S = 1)$ and $P(\mathbf{X}, S \mid S = 0)$, respectively. Here $P(\cdot \mid S = s)$ refers to a conditional distribution given $S = s$ $(s \in \{0, 1\})$.

Denote $\mathbf{\Psi}'_n(\mathbf{W}, \mathbf{b}) \equiv \partial \mathbf{\Psi}_n(\mathbf{W}, \mathbf{b})/\partial \mathbf{b}$. The law of large numbers, Slutsky's Theorem and Assumption 1 give that, as $n, N \to \infty$,

$$
\begin{aligned}
&n^{-1}\textstyle\sum_{i=1}^{N}\mathbf{\Psi}'_n(\mathbf{W}_i,\boldsymbol{\beta})\\
&\equiv n^{-1}\{\textstyle\sum_{i\in\mathcal{V}}\mathbf{\Psi}'_n(\mathbf{W}_i,\boldsymbol{\beta}) + \sum_{i\in\mathcal{U}}\mathbf{\Psi}'_n(\mathbf{W}_i,\boldsymbol{\beta}) + \sum_{i\in\mathcal{C}}\mathbf{\Psi}'_n(\mathbf{W}_i,\boldsymbol{\beta})\}\\
&= n^{-1}\textstyle\sum_{i\in\mathcal{V}}\mathbf{\Psi}'_n(\mathbf{W}_i,\boldsymbol{\beta}) + (\tau-\delta_{n,N})N_{\rm U}^{-1}\sum_{i\in\mathcal{U}}\mathbf{\Psi}'_{\rm U}(\mathbf{W}_i,\boldsymbol{\beta}) + (1-\tau)N_0^{-1}\sum_{i\in\mathcal{C}}\mathbf{\Psi}'_{\rm U}(\mathbf{W}_i,\boldsymbol{\beta})\\
&\to \mathbf{A}_{\rm V} + (\tau-\delta)\mathbf{A}_{\rm U} + (1-\tau)\mathbf{A}_0 = \mathbf{A} \qquad \text{(S.6)}
\end{aligned}
$$

in probability.

Then denote $\boldsymbol{\zeta}_{\rm V} \equiv E\{\mathbf{\Psi}_{\rm V}(\mathbf{W},\boldsymbol{\beta}) \mid S = 1\}$, $\boldsymbol{\zeta}_{\rm U} \equiv E\{\mathbf{\Psi}_{\rm U}(\mathbf{W},\boldsymbol{\beta}) \mid S = 1\}$ and $\boldsymbol{\zeta}_0 \equiv E\{\mathbf{\Psi}_{\rm U}(\mathbf{W},\boldsymbol{\beta}) \mid S = 0\}$. By the central limit theorem, Slutsky's theorem and Assumption 1, we have

$$
n^{1/2}\{n^{-1}\textstyle\sum_{i\in\mathcal{V}}\mathbf{\Psi}_n(\mathbf{W}_i,\boldsymbol{\beta}) - \boldsymbol{\zeta}_{\rm V}\} \to \mathbf{N}(\mathbf{0},\mathbf{B}_{\rm V}) \quad (n, N \to \infty). \qquad \text{(S.7)}
$$

Analogously, we can obtain that, as $n, N \to \infty$,

$$
n^{1/2}\{n^{-1}\textstyle\sum_{i\in\mathcal{U}}\mathbf{\Psi}_n(\mathbf{W}_i,\boldsymbol{\beta}) - (\tau-\delta)\boldsymbol{\zeta}_{\rm U}\} \to \mathbf{N}\{\mathbf{0},\delta(\tau-\delta)\mathbf{B}_{\rm U}\}, \qquad \text{(S.8)}
$$

$$
n^{1/2}\{n^{-1}\textstyle\sum_{i\in\mathcal{C}}\mathbf{\Psi}_n(\mathbf{W}_i,\boldsymbol{\beta}) - (1-\tau)\boldsymbol{\zeta}_0\} \to \mathbf{N}\{\mathbf{0},\delta(1-\tau)\mathbf{B}_0\}. \qquad \text{(S.9)}
$$

Considering the facts that $\{\mathbf{W}_i : i \in \mathcal{V}\}$, $\{\mathbf{W}_i : i \in \mathcal{U}\}$ and $\{\mathbf{W}_i : i \in \mathcal{C}\}$ are mutually independent, and that

$$\boldsymbol{\zeta}_{\mathrm{V}} + (\tau - \delta)\boldsymbol{\zeta}_{\mathrm{U}} + (1-\tau)\boldsymbol{\zeta}_0 = \boldsymbol{\Phi}(\boldsymbol{\beta}) = n^{-1}\textstyle\sum_{i=1}^N E\{\boldsymbol{\Psi}_n(\mathbf{W}_i, \boldsymbol{\beta})\} = \mathbf{0}$$

due to Proposition 1 and the case-control sampling, Lemma 1 and the convergences in (S.7)–(S.9) imply that, as $n, N \to \infty$,

$$\begin{aligned} n^{-1/2}\textstyle\sum_{i=1}^N \boldsymbol{\Psi}_n(\mathbf{W}_i, \boldsymbol{\beta}) &\equiv n^{-1/2}\{\textstyle\sum_{i\in\mathcal{V}} \boldsymbol{\Psi}_n(\mathbf{W}_i, \boldsymbol{\beta}) + \textstyle\sum_{i\in\mathcal{U}} \boldsymbol{\Psi}_n(\mathbf{W}_i, \boldsymbol{\beta}) + \textstyle\sum_{i\in\mathcal{C}} \boldsymbol{\Psi}_n(\mathbf{W}_i, \boldsymbol{\beta})\} \\ &\to \mathbf{N}(\mathbf{0}, \mathbf{B}). \end{aligned} \tag{S.10}$$

The case-control sampling and Proposition 1 guarantee that

$$\boldsymbol{\Phi}(\boldsymbol{\beta}) = n^{-1}\textstyle\sum_{i=1}^N E\{\boldsymbol{\Psi}_n(\mathbf{W}_i, \boldsymbol{\beta})\} = \mathbf{0}. \tag{S.11}$$

Then, considering the condition (S.1) and the fact that $\boldsymbol{\beta}$ is the unique solution to $\boldsymbol{\Phi}(\mathbf{b}) = \mathbf{0}$ from Assumption 1, Theorem 5.9 of Van der Vaart (2000) gives

$$\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta} \quad (n, N \to \infty) \tag{S.12}$$

in probability. By Taylor's expansion and (S.6), we have

$$\begin{aligned} \mathbf{0} = n^{-1}\textstyle\sum_{i=1}^N \boldsymbol{\Psi}_n(\mathbf{W}_i, \widehat{\boldsymbol{\beta}}) &= n^{-1}\textstyle\sum_{i=1}^N \boldsymbol{\Psi}_n(\mathbf{W}_i, \boldsymbol{\beta}) + n^{-1}\textstyle\sum_{i=1}^N \boldsymbol{\Psi}'_n(\mathbf{W}_i, \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{r}_n \\ &= n^{-1}\textstyle\sum_{i=1}^N \boldsymbol{\Psi}_n(\mathbf{W}_i, \boldsymbol{\beta}) + \mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{u}_n + \mathbf{r}_n. \end{aligned} \tag{S.13}$$

In (S.13), the term $\mathbf{u}_n \equiv \{n^{-1}\sum_{i=1}^N \boldsymbol{\Psi}'_n(\mathbf{W}_i, \boldsymbol{\beta}) - \mathbf{A}\}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, while $\mathbf{r}_n$ is a $2p$-dimensional vector, whose jth component is $r_{n,j} \equiv n^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Omega}'_{n,j}(\bar{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ with $\bar{\boldsymbol{\beta}} \equiv \boldsymbol{\beta} + \boldsymbol{\Lambda}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ for some diagonal matrix $\boldsymbol{\Lambda} \equiv \mathrm{diag}(\lambda_1, \ldots, \lambda_{2p})$ and some $\lambda_j \in (0,1)$ $(j = 1, \ldots, 2p)$. The convergence in (S.6) and Assumption 1 give

$$\|\mathbf{A}^{-1}\mathbf{u}_n\| = o_p(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|). \tag{S.14}$$

Further, by the condition (S.2) of Assumption 1 and the convergence (S.12), we know that

$$\begin{aligned} |r_{n,j}| = n^{-1}|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Omega}'_{n,j}(\bar{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})| &\leqslant n^{-1}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\lambda_{\max}\{\boldsymbol{\Omega}'_{n,j}(\bar{\boldsymbol{\beta}})\} \\ &= o_p(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|) \quad (j = 1, \ldots, 2p), \end{aligned}$$

which implies that

$$\|\mathbf{A}^{-1}\mathbf{r}_n\| = o_p(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|). \tag{S.15}$$

Considering the fact that $n^{-1}\|\sum_{i=1}^{N}\boldsymbol{\Psi}_n(\mathbf{W}_i,\boldsymbol{\beta})\| = O_p(n^{-1/2})$ from (S.10), (S.10) through (S.15) conclude that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_p(n^{-1/2})$ and that

$$\begin{aligned} n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= -\mathbf{A}^{-1}\{n^{-1/2}\textstyle\sum_{i=1}^{N}\boldsymbol{\Psi}_n(\mathbf{W}_i,\boldsymbol{\beta}) + n^{1/2}(\mathbf{u}_n + \mathbf{r}_n)\} \\ &= -n^{-1/2}\mathbf{A}^{-1}\textstyle\sum_{i=1}^{N}\boldsymbol{\Psi}_n(\mathbf{W}_i,\boldsymbol{\beta}) + o_p(1) \to \mathbf{N}\{\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}\} \end{aligned}$$

in distribution as $n, N \to \infty$.

Web Appendix F: Proof of Corollary 1

We have

$$\begin{aligned} N^{-1}\textstyle\sum_{i=1}^{N}\boldsymbol{\Psi}_1(\mathbf{W}_i,\boldsymbol{\beta}) &\equiv N^{-1}\textstyle\sum_{i\in\mathcal{V}\cup\mathcal{U}}\boldsymbol{\Psi}_1(\mathbf{W}_i,\boldsymbol{\beta}) + N^{-1}\textstyle\sum_{i\in\mathcal{C}}\boldsymbol{\Psi}_1(\mathbf{W}_i,\boldsymbol{\beta}) \\ &= \tau E\{\boldsymbol{\Psi}_1(\mathbf{W},\boldsymbol{\beta}) \mid S=1\} + (1-\tau)E\{\boldsymbol{\Psi}_1(\mathbf{W},\boldsymbol{\beta}) \mid S=0\} + O_p(N^{-1/2}) \\ &= o_p(n^{-1/2}). \end{aligned} \tag{S.16}$$

In the above, the second step is due to the case-control sampling and the central limit theorem, while the third step holds by the facts that $\delta \equiv \lim_{n,N\to\infty} n/N = 0$ and that

$$\begin{aligned} &\tau E\{\boldsymbol{\Psi}_1(\mathbf{W},\boldsymbol{\beta}) \mid S=1\} + (1-\tau)E\{\boldsymbol{\Psi}_1(\mathbf{W},\boldsymbol{\beta}) \mid S=0\} \\ &\equiv \tau E\{h(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X})\mathbf{X}\overline{h}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}) \mid S=1\} - (1-\tau)E\{\mathbf{X}h(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}) \mid S=0\} \\ &= N^{-1}\textstyle\sum_{i=1}^{N}E\{S_i g(\mathbf{X}_i)\mathbf{X}_i\overline{h}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i) - (1-S_i)\mathbf{X}_i h(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i)\} = \mathbf{0} \end{aligned}$$

from (13) in Proposition 1. The result (S.16), combined with (17) and the fact that $R_iS_i = I(i \leqslant n)$ $(i = 1,\ldots,N)$, implies (19). Then the asymptotic normality can be shown by setting $\delta \equiv 0$ in (18).

Web Appendix G: Proof of the Equation (22)

Since $E(D \mid \mathbf{X}, S=1) = h(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X})$, we have

$$E\{\boldsymbol{\Psi}_2(\mathbf{W},\boldsymbol{\beta}) \mid \mathbf{X}, S=1\} \equiv E[\{h(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}) - D\}\mathbf{X}\overline{h}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}) \mid \mathbf{X}, S=1] = \mathbf{0},$$

which indicates

$$E\{\boldsymbol{\varphi}(\mathbf{W},\boldsymbol{\beta}) \mid \mathbf{X}, S=1\} \equiv E[\mathbf{A}^{-1}\{\mathbf{0}_p^{\mathrm{T}}, \boldsymbol{\Psi}_2^{\mathrm{T}}(\mathbf{W},\boldsymbol{\beta})\}^{\mathrm{T}} \mid \mathbf{X}, S=1] = \mathbf{0}. \tag{S.17}$$

According to Theorem 4.5 of Tsiatis (2007), the tangent space corresponding to the semiparametric model $\mathcal{M}$, given in (20), is

$$\mathcal{T} \equiv \{\mathbf{g}(D,S,\mathbf{X}) \in \mathbb{R}^{2p} : E\{\mathbf{g}(D,S,\mathbf{X}) \mid \mathbf{X}, S=1\} = \mathbf{0}, \mathrm{cov}\{\mathbf{g}(D,S,\mathbf{X})\} < \infty\},$$

and the projection of the influence function $\boldsymbol{\varphi}(\mathbf{W},\boldsymbol{\beta})$ of $\widehat{\boldsymbol{\beta}}$ onto $\mathcal{T}$ is

$$\Pi\{\boldsymbol{\varphi}(\mathbf{W},\boldsymbol{\beta}) \mid \mathcal{T}\} = \boldsymbol{\varphi}(\mathbf{W},\boldsymbol{\beta}) - E\{\boldsymbol{\varphi}(\mathbf{W},\boldsymbol{\beta}) \mid \mathbf{X}, S=1\} = \boldsymbol{\varphi}(\mathbf{W},\boldsymbol{\beta})$$

due to (S.17). Then, using Theorem 4.3 of Tsiatis (2007), we know

$$\boldsymbol{\varphi}_{\mathrm{EFF}}(\mathbf{W},\boldsymbol{\beta}) \equiv \Pi\{\boldsymbol{\varphi}(\mathbf{W},\boldsymbol{\beta}) \mid \mathcal{T}\} = \boldsymbol{\varphi}(\mathbf{W},\boldsymbol{\beta}) \equiv \mathbf{A}^{-1}\{\mathbf{0}_p^{\mathrm{T}}, \boldsymbol{\Psi}_2^{\mathrm{T}}(\mathbf{W},\boldsymbol{\beta})\}.$$

Web Appendix H: Supplements for the Numerical Studies in Sections 4 and 5

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

References

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.

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

Table S1

Proportions $E(D \mid S = 1)$ of true cases in the candidate case pool under the simulation settings (23) and (24) used in Section 4. Here p is the covariate dimension while (a)–(e) are forms of the function $\rho(\mathbf{x})$ in (23)–(24).

	Model (23)					Model (24)				
p	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)
7	0.39	0.55	0.45	0.61	0.43	0.65	0.56	0.61	0.50	0.65
13	0.33	0.57	0.47	0.62	0.49	0.71	0.54	0.61	0.49	0.58

Table S2

Results of the simulations in Section 4: biases, measured by the criterion (25)*, of the estimators for* $\boldsymbol{\theta}_{-1}$ *from the validation-only estimating equation (VEE), from the naive estimating equation (NEE), from the parametrically imputed estimating equation (PIEE) and from the unbiasedly imputed estimating equation* (**UIEE**) *under the models* (23) *and* (24) *with the covariate dimension* $p = 7$*. Here* n *is the validation set size,* N *is the whole sample size and* $\rho(\mathbf{x})$ *is the function in* (23)–(24)*.*

			$N = 2,500$				$N = 5,000$				$N = 25,000$			
Model	n	$\rho(\mathbf{x})$	VEE	NEE	PIEE	**UIEE**	VEE	NEE	PIEE	**UIEE**	VEE	NEE	PIEE	**UIEE**
		(a)	0.01	0.41	0.01	**0.01**	0.01	0.40	0.01	**0.01**	0.01	0.40	0.01	**0.01**
		(b)	0.01	0.22	0.07	**0.01**	0.00	0.22	0.07	**0.01**	0.01	0.22	0.07	**0.01**
	100	(c)	0.01	0.34	0.07	**0.00**	0.00	0.34	0.07	**0.01**	0.01	0.34	0.07	**0.00**
		(d)	0.01	0.21	0.05	**0.01**	0.01	0.21	0.06	**0.01**	0.00	0.21	0.06	**0.00**
		(e)	0.01	0.33	0.09	**0.01**	0.01	0.32	0.09	**0.01**	0.01	0.32	0.09	**0.00**
		(a)	0.01	0.41	0.00	**0.00**	0.01	0.40	0.00	**0.00**	0.01	0.40	0.00	**0.00**
		(b)	0.01	0.22	0.06	**0.00**	0.01	0.22	0.07	**0.00**	0.01	0.22	0.07	**0.00**
(23)	200	(c)	0.01	0.34	0.06	**0.00**	0.00	0.34	0.07	**0.00**	0.00	0.34	0.08	**0.00**
		(d)	0.01	0.21	0.05	**0.00**	0.01	0.21	0.06	**0.00**	0.00	0.21	0.06	**0.00**
		(e)	0.01	0.33	0.08	**0.00**	0.00	0.32	0.09	**0.00**	0.00	0.32	0.09	**0.00**
		(a)	0.01	0.41	0.00	**0.00**	0.01	0.40	0.00	**0.00**	0.00	0.40	0.00	**0.00**
		(b)	0.01	0.22	0.05	**0.00**	0.00	0.22	0.06	**0.00**	0.00	0.22	0.07	**0.00**
	400	(c)	0.01	0.34	0.05	**0.00**	0.00	0.34	0.06	**0.00**	0.00	0.34	0.08	**0.00**
		(d)	0.01	0.21	0.04	**0.00**	0.00	0.21	0.05	**0.00**	0.00	0.21	0.06	**0.00**
		(e)	0.01	0.33	0.06	**0.00**	0.01	0.32	0.08	**0.00**	0.00	0.32	0.09	**0.00**
		(a)	0.15	0.47	0.02	**0.02**	0.18	0.47	0.02	**0.02**	0.22	0.47	0.02	**0.02**
		(b)	0.15	0.43	0.15	**0.02**	0.19	0.43	0.15	**0.02**	0.22	0.43	0.16	**0.02**
	100	(c)	0.17	0.48	0.10	**0.02**	0.20	0.49	0.11	**0.02**	0.24	0.49	0.11	**0.01**
		(d)	0.20	0.56	0.17	**0.02**	0.24	0.56	0.18	**0.02**	0.29	0.56	0.18	**0.02**
		(e)	0.13	0.40	0.10	**0.02**	0.16	0.40	0.11	**0.02**	0.20	0.40	0.11	**0.02**
		(a)	0.11	0.47	0.02	**0.02**	0.16	0.47	0.01	**0.01**	0.21	0.47	0.01	**0.01**
		(b)	0.12	0.43	0.14	**0.02**	0.16	0.43	0.15	**0.01**	0.21	0.43	0.16	**0.01**
(24)	200	(c)	0.13	0.48	0.09	**0.01**	0.18	0.49	0.11	**0.01**	0.23	0.49	0.11	**0.01**
		(d)	0.16	0.56	0.16	**0.02**	0.21	0.56	0.18	**0.01**	0.27	0.56	0.19	**0.01**
		(e)	0.10	0.40	0.09	**0.01**	0.14	0.40	0.10	**0.01**	0.19	0.40	0.11	**0.01**
		(a)	0.07	0.47	0.01	**0.01**	0.12	0.47	0.01	**0.01**	0.19	0.47	0.01	**0.01**
		(b)	0.08	0.43	0.11	**0.01**	0.13	0.43	0.14	**0.01**	0.20	0.43	0.16	**0.01**
	400	(c)	0.08	0.48	0.07	**0.01**	0.14	0.49	0.10	**0.01**	0.22	0.49	0.11	**0.01**
		(d)	0.10	0.56	0.13	**0.02**	0.16	0.56	0.16	**0.01**	0.25	0.56	0.19	**0.01**
		(e)	0.06	0.40	0.07	**0.01**	0.11	0.40	0.09	**0.01**	0.17	0.40	0.11	**0.01**

Table S3

*Results of the simulations in Section 4: mean squared error ratios of the estimators for $\boldsymbol{\theta}_{-1}$ from the validation-only estimating equation to those from the naive estimating equation (NEE), from the parametrically imputed estimating equation (PIEE) and from the unbiasedly imputed estimating equation (**UIEE**) under the models (23) and (24) with the covariate dimension $p = 7$. Here n is the validation set size, N is the whole sample size and $\rho(\mathbf{x})$ is the function in (23)–(24).*

			$N = 2,500$			$N = 5,000$			$N = 25,000$		
Model	n	$\rho(\mathbf{x})$	NEE	PIEE	**UIEE**	NEE	PIEE	**UIEE**	NEE	PIEE	**UIEE**
		(a)	0.22	3.28	**3.10**	0.21	3.71	**3.49**	0.22	4.21	**3.97**
		(b)	0.43	1.59	**2.45**	0.43	1.68	**2.86**	0.42	1.69	**3.19**
	100	(c)	0.24	1.68	**2.66**	0.24	1.77	**3.03**	0.21	1.60	**3.21**
		(d)	0.46	2.10	**3.23**	0.45	2.23	**3.86**	0.41	2.23	**4.37**
		(e)	0.26	1.32	**2.25**	0.25	1.30	**2.45**	0.23	1.21	**2.50**
		(a)	0.11	2.57	**2.47**	0.10	3.25	**3.10**	0.10	4.33	**4.14**
		(b)	0.23	1.22	**2.05**	0.23	1.23	**2.60**	0.21	1.19	**3.34**
(23)	200	(c)	0.12	1.26	**2.17**	0.12	1.29	**2.76**	0.11	1.18	**3.39**
		(d)	0.24	1.50	**2.45**	0.23	1.56	**3.27**	0.21	1.52	**4.70**
		(e)	0.14	1.05	**1.97**	0.13	0.96	**2.30**	0.12	0.85	**2.62**
		(a)	0.06	1.73	**1.70**	0.05	2.46	**2.40**	0.05	4.22	**4.06**
		(b)	0.13	1.02	**1.55**	0.12	0.91	**2.14**	0.11	0.75	**3.32**
	400	(c)	0.07	1.04	**1.60**	0.06	0.96	**2.21**	0.06	0.79	**3.40**
		(d)	0.13	1.10	**1.58**	0.12	1.05	**2.27**	0.11	0.93	**4.29**
		(e)	0.08	0.93	**1.57**	0.07	0.74	**1.99**	0.06	0.59	**2.62**
		(a)	0.21	2.88	**2.82**	0.24	3.85	**3.78**	0.28	5.38	**5.19**
		(b)	0.27	1.24	**2.58**	0.31	1.38	**3.30**	0.36	1.60	**4.45**
	100	(c)	0.23	1.83	**3.14**	0.26	2.15	**4.29**	0.32	2.69	**5.81**
		(d)	0.22	1.29	**3.04**	0.27	1.48	**4.19**	0.33	1.82	**5.39**
		(e)	0.25	1.52	**2.47**	0.29	1.72	**3.19**	0.34	2.06	**4.33**
		(a)	0.12	2.84	**2.78**	0.16	4.85	**4.71**	0.23	9.60	**9.34**
		(b)	0.16	1.02	**2.66**	0.21	1.20	**4.29**	0.30	1.62	**8.19**
(24)	200	(c)	0.14	1.67	**3.20**	0.18	2.14	**5.55**	0.27	3.15	**10.81**
		(d)	0.14	1.07	**3.18**	0.19	1.32	**5.57**	0.28	1.84	**10.14**
		(e)	0.15	1.38	**2.48**	0.19	1.66	**3.99**	0.27	2.31	**7.40**
		(a)	0.06	2.00	**1.99**	0.10	4.61	**4.53**	0.19	14.57	**14.16**
		(b)	0.08	0.82	**1.98**	0.13	0.97	**4.32**	0.24	1.50	**12.71**
	400	(c)	0.07	1.31	**2.23**	0.11	1.84	**5.45**	0.22	3.15	**16.82**
		(d)	0.07	0.85	**2.16**	0.11	1.06	**5.23**	0.23	1.72	**17.50**
		(e)	0.08	1.10	**1.80**	0.11	1.38	**3.81**	0.21	2.24	**11.10**

Table S4

Names, descriptions and summary statistics of the covariates considered in the data analysis of Section 5. Here "SD" stands for "standard deviation".

Variable name	**Description**	**Mean**	**SD**
urineoutput	Urine output (ml/kg/hr)	1532.4	1384.1
sysbp_min	Minimum systolic blood pressure (mmHg)	82.0	17.3
spo2_mean	Mean fraction of oxygen-saturated hemoglobin relative to total hemoglobin (unsaturated+saturated) in the blood	96.5	3.1
sodium_max	Maximum sodium concentration in blood (mmol/L)	140.7	6.48
metastatic_cancer	Having metastatic cancer (1) or not (0)	0.1	0.3
lactate_min	Minimum lactate concentration in blood (mmol/L)	2.0	1.6
is_male	Being male (1) or not (0)	0.6	0.5
cretinine_min	Minimum cretinine concentration in blood (mmol/L)	1.6	1.3
bun_mean	Mean level of blood urea nitrogen (mmol/L)	37.3	26.5
aniongap_max	Maximum anion gap (mmol/L)	18.3	5.6
age	Age (year)	68.5	16.9