Submitted to the Statistical Science

Supplement to "Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes"

Phillip J. Schulte, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian

Abstract. This supplemental article contains technical details related to material presented in the parent article. Section and equation numbers not preceded by "A" refer to those in the parent article.

Phillip J. Schulte is Biostatistician, Duke Clinical Research Institute, Durham, North Carolina 27701, USA (e-mail: phillip.schulte@duke.edu). Anastasios A. Tsiatis is Gertrude M. Cox Distinguished Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, USA (e-mail: tsiatis@ncsu.edu). Eric B. Laber is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, USA (e-mail: eblaber@ncsu.edu). Marie Davidian is William Neal Reynolds Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, USA (e-mail: davidian@ncsu.edu).

A.1. DEMONSTRATION THAT (5)–(8) DEFINE AN OPTIMAL REGIME

For $k = 1, \dots, K$ and any $d \in \mathcal{D}$, define the random variables $\alpha_k\{\bar{S}_k^*(\bar{d}_{k-1})\}$ such that

$$\alpha_k\{\bar{S}_k^*(\bar{d}_{k-1})\}(\omega) = V_k^{(1)}(\bar{s}_k, \bar{u}_{k-1}) \tag{A.1}$$

for any $\omega \in \Omega$, where $(\bar{s}_k, \bar{u}_{k-1})$ are defined by (3). Because $Vone_k$ as defined in (6) and (8) are functions of arguments $(\bar{s}_k, \bar{a}_{k-1}) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_{k-1}$, whereas the random variable $\bar{S}_k^*(\bar{d}_{k-1})$ has realizations only in $\bar{\mathcal{S}}_k$, we find it convenient to introduce (A.1) to make this distinction clear. In words, $\alpha_k\{\bar{S}_k^*(\bar{d}_{k-1})\}$ is the expected outcome conditional on a patient in the population having followed regime d through decision point $k-1$ and then following the optimal regime from that point on.

We now argue that $d^{(1)\text{opt}}$ is an optimal regime; i.e., $d^{(1)\text{opt}}$ satisfies (4). We first show that, for any $d \in \mathcal{D}$,

$$\begin{aligned} \mathrm{E}\{Y^*(d)|S_1 = s_1, S_2^*(d_1), \dots, S_K^*(\bar{d}_{K-1})\} &\leq \mathrm{E}\{Y^*(\bar{d}_{K-1}, d_K^{(1)\text{opt}})|S_1 = s_1, S_2^*(d_1), \dots, S_K^*(\bar{d}_{K-1})\} \\ &= \alpha_K\{s_1, S_2^*(d_1), \dots, S_K^*(\bar{d}_{K-1})\}. \end{aligned} \tag{A.2}$$

This follows because, for the set in Ω where $\{S_2^*(d_1) = s_2, \dots, S_K^*(\bar{d}_{K-1}) = s_K\}$, the left- and right-hand sides of the first line of (A.2) are equal to

$$\mathrm{E}\{Y^*(d)|\bar{S}_K^*(\bar{d}_{K-1}) = \bar{s}_K\} = \mathrm{E}\{Y^*(\bar{u}_{K-1}, u_K)|\bar{S}_K^*(\bar{u}_{K-1}) = \bar{s}_K\}, \tag{A.3}$$

$$\mathrm{E}\{Y^*(\bar{d}_{K-1}, d_K^{(1)\text{opt}})|\bar{S}^*(\bar{d}_{K-1}) = \bar{s}_K\} = \mathrm{E}[Y^*\{\bar{u}_{K-1}, d_K^{(1)\text{opt}}(\bar{s}_K, \bar{u}_{K-1})\}|\bar{S}_K^*(\bar{u}_{K-1}) = \bar{s}_K], \tag{A.4}$$

respectively. By the definition of $d_K^{(1)\text{opt}}$ in (5), (A.4) is greater than or equal to (A.3), and, by the definition of $V_K^{(1)}$ in (6), (A.4) equals $V_K^{(1)}(\bar{s}_K, \bar{u}_{K-1})$. Because these results hold for sets $\{S_2^*(d_1) = s_2, \dots, S_K^*(\bar{d}_{K-1}) = s_K\}$ for any $(s_2, \dots, s_K)$ such that $(s_1, \dots, s_K, u_1, \dots, u_{K-1}) \in \Gamma_k$, and by the definition of α_K in (A.1), (A.2) holds. Taking conditional expectations given $S_1 = s_1$ yields

$$\begin{aligned} &\mathrm{E}\{Y^*(d)|S_1 = s_1\} \leq \mathrm{E}\{Y^*(\bar{d}_{K-1}, d_K^{(1)\text{opt}})|S_1 = s_1\} \\ &= \mathrm{E}[\alpha_K\{s_1, S_2^*(d_1), \dots, S_K^*(\bar{d}_{K-1})\}|S_1 = s_1]. \end{aligned} \tag{A.5}$$

The equality in (A.5) holds for any $\bar{d}_{K-1} = (d_1, \dots, d_{K-1})$ with $d \in \mathcal{D}$, hence it must hold for

$(d_1, \ldots, d_{K-2}, d_{K-1}^{(1)\mathrm{opt}})$. Thus, we also have that

$$\begin{aligned} &\mathrm{E}\{Y^*(\bar{d}_{K-2}, d_{K-1}^{(1)\mathrm{opt}}, d_K^{(1)\mathrm{opt}})|S_1 = s_1\} \\ &\quad = \mathrm{E}[\alpha_K\{S_1, S_2^*(d_1), \ldots, S_{K-1}^*(\bar{d}_{K-2}), S_K^*(\bar{d}_{K-2}, d_{K-1}^{(1)\mathrm{opt}})\}|S_1 = s_1]. \end{aligned} \tag{A.6}$$

Similarly, for any $k = K-1, \ldots, 1$, we can show that $\mathrm{E}[\alpha_{k+1}\{S_1, S_2^*(d_1), \ldots, S_{k+1}^*(\bar{d}_k)\}|S_1 = s_1, S_2^*(d_1), \ldots, S_k^*(\bar{d}_{k-1})] \leq \mathrm{E}[\alpha_{k+1}\{S_1, S_2^*(d_1), \ldots, S_{k+1}^*(\bar{d}_{k-1}, d_k^{(1)\mathrm{opt}})\}|S_1 = s_1, S_2^*(d_1), \ldots, S_k^*(\bar{d}_{k-1})]$ $= \alpha_k\{s_1, S_2^*(d_1), \ldots, S_k^*(\bar{d}_{k-1})\}$, which implies for $k = K-1, \ldots, 1$,

$$\begin{aligned} &\mathrm{E}[\alpha_{k+1}\{s_1, S_2^*(d_1), \ldots, S_{k+1}^*(\bar{d}_k)\}|S_1 = s_1] \leq \mathrm{E}[\alpha_{k+1}\{s_1, S_2^*(d_1), \ldots, S_{k+1}^*(\bar{d}_{k-1}, d_k^{(1)\mathrm{opt}})\}|S_1 = s_1] \\ &\quad = \mathrm{E}[\alpha_k\{s_1, S_2^*(d_1), \ldots, S_k^*(\bar{d}_{k-1})\}|S_1 = s_1] \end{aligned} \tag{A.7}$$

Using (A.5) and (A.7) with $k = K-1$, we thus have

$$\begin{aligned} \mathrm{E}\{Y^*(d)|S_1 = s_1\} &\leq \mathrm{E}\{Y^*(\bar{d}_{K-1}, d_K^{(1)\mathrm{opt}})|S_1 = s_1\} = \mathrm{E}[\alpha_K\{s_1, S_2^*(d_1), \ldots, S_K^*(\bar{d}_{K-1})|S_1 = s_1] \\ &\leq \mathrm{E}[\alpha_K\{s_1, S_2^*(d_1), \ldots, S_K^*(\bar{d}_{K-2}, d_{K-1}^{(1)\mathrm{opt}})|S_1 = s_1] \\ &= \mathrm{E}[\alpha_{K-1}\{s_1, S_2^*(d_1), \ldots, S_{K-1}^*(\bar{d}_{K-2})|S_1 = s_1] \end{aligned} \tag{A.8}$$

Because of (A.6), the term in (A.8) is equal to $\mathrm{E}\{Y^*(\bar{d}_{K-2}, d_{K-1}^{(1)\mathrm{opt}}, d_K^{(1)\mathrm{opt}})|S_1 = s_1\}$. Hence,

$$\begin{aligned} &\mathrm{E}\{Y^*(d)|S_1 = s_1\} \leq \mathrm{E}\{(Y^*(\bar{d}_{K-1}, d_K^{(1)\mathrm{opt}})|S_1 = s_1\} \leq \mathrm{E}\{Y^*(\bar{d}_{K-2}, d_{K-1}^{(1)\mathrm{opt}}, d_K^{(1)\mathrm{opt}})|S_1 = s_1\} \\ &\quad = \mathrm{E}[\alpha_{K-1}\{s_1, S_2^*(d_1), \ldots, S_{K-1}^*(\bar{d}_{K-2})|S_1 = s_1]. \end{aligned} \tag{A.9}$$

Again, because $\bar{d}_{K-2}$ is arbitrary, if we replace it by $(\bar{d}_{K-3}, d_{K-2}^{(1)\mathrm{opt}})$, the equality in (A.9) implies

$$\mathrm{E}\{Y^*(\bar{d}_{K-3}, \underline{d}_{K-2}^{(1)\mathrm{opt}})|S_1 = s_1\} = \mathrm{E}[\alpha_{K-1}\{s_1, S_2^*(d_1), \ldots, S_{K-1}^*(\bar{d}_{K-3}, d_{K-2}^{(1)\mathrm{opt}})\}|S_1 = s_1], \tag{A.10}$$

where, for any d, $\underline{d}_k = (d_k, \ldots, d_K)$. Using (A.7) with $k = K-2$, (A.9), and (A.10), we obtain

$$\begin{aligned} &\mathrm{E}\{Y^*(\bar{d}_{K-2}, \underline{d}_{K-1}^{(1)\mathrm{opt}})|S_1 = s_1\} = \mathrm{E}[\alpha_{K-1}\{s_1, S_2^*(d_1), \ldots, S_{K-1}^*(\bar{d}_{K-2})\}|S_1 = s_1] \\ &\quad \leq \mathrm{E}[\alpha_{K-1}\{s_1, S_2^*(d_1), \ldots, S_{K-1}^*(\bar{d}_{K-3}, d_{K-2}^{(1)\mathrm{opt}})|S_1 = s_1] = \mathrm{E}\{Y^*(\bar{d}_{K-3}, \underline{d}_{K-2}^{(1)\mathrm{opt}})\}|S_1 = s_1\} \\ &\quad = \mathrm{E}[\alpha_{K-2}\{s_1, S_2^*(d_1), \ldots, S_{K-2}^*(\bar{d}_{K-3})\}|S_1 = s_1]. \end{aligned}$$

Continuing in this fashion, we may conclude that, for any $d \in \mathcal{D}$,

$$\mathrm{E}\{Y^*(d)|S_1 = s_1\} \leq \cdots \leq \mathrm{E}\{Y^*(\bar{d}_{k-1}, \underline{d}_k^{(1)\mathrm{opt}})|S_1 = s_1\} \leq \cdots \leq \mathrm{E}\{Y^*(d^{(1)\mathrm{opt}})|S_1 = s_1\},$$

showing that $d^{(1)\mathrm{opt}}$ defined in (5) and (7) is an optimal regime satisfying (4).

A.2. DEMONSTRATION OF CORRESPONDENCE IN (19)

We make the consistency, sequential randomization, and positivity (15) assumptions given in Section 3; the latter states that, for any $(\bar{s}_k, \bar{a}_{k-1})$ for which $\mathrm{pr}(\bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}) > 0$, $\mathrm{pr}(A_k = a_k | \bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}) > 0$ if $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$ and $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$, $k = 1, \ldots, K$. This ensures that the observed data contain information on the possible treatment options involved in the class of regimes under consideration. We wish to show (16)–(18), which we restate here for convenience as

$$\text{(A.11)} \qquad \mathrm{pr}(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) > 0,$$

$$\text{(A.12)} \quad \mathrm{pr}(S_{k+1} = s_{k+1} | \bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) = \mathrm{pr}\{S^*_{k+1}(\bar{a}_k) = s_{k+1} | \bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}\}$$

$$\text{(A.13)} \qquad = \mathrm{pr}\{S^*_{k+1}(\bar{a}_k) = s_{k+1} | \bar{S}_j = \bar{s}_j, \bar{A}_{j-1} = \bar{a}_{j-1}, S^*_{j+1}(\bar{a}_j) = s_{j+1}, \ldots, S^*_k(\bar{a}_{k-1}) = s_k\},$$

for any $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$ and $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$, $k = 1, \ldots, K$, for $j = 1, \ldots, k$, where we define (A.13) with $j = k$ to be the same as the expression on the right-hand side of (A.12) and take $S_{K+1} = Y$ and $S^*_{K+1}(\bar{a}_K) = Y^*(\bar{a}_K)$. If we can show (A.11), then the quantities in (9)–(14) are well-defined. If (A.12)–(A.13) also hold, then the conditional distributions of the observed data involved in the quantities in (9)–(14) are the same as the conditional distributions of the potential outcomes involved in (5)–(8), and (19) follows.

Assume for the moment that (A.11) is true. We now demonstrate (A.12) and (A.13) must follow. For any fixed k, by the consistency assumption, the left-hand expression in (A.12) is equal to $\mathrm{pr}\{S^*_{k+1}(\bar{a}_k) = s_{k+1} | \bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}, A_k = a_k\}$. It follows by the sequential randomization assumption, which implies $A_k \perp\!\!\!\perp S^*_{k+1}(\bar{a}_k) | \bar{S}_k, \bar{A}_{k-1}$, that this is equal to the right-hand side of (A.12). The equality in (A.13) follows by induction. First note that right-hand side of (A.12) is equal to (A.13) with $j = k$. The equality of (A.12) and (A.13) for all $j = 1, \ldots, k$ can be deduced if we can show that (A.13) being true for a given j implies that it is also true for $j - 1$. For a given $j = 2, \ldots, k$, by the consistency assumption, (A.13) is equal to $\mathrm{pr}\{S^*_{k+1}(\bar{a}_k) = s_{k+1} | \bar{S}_{j-1} = \bar{s}_{j-1}, \bar{A}_{j-2} = \bar{a}_{j-2}, A_{j-1} = a_{j-1}, S^*_j(\bar{a}_j) = s_j, \ldots, S^*_k(\bar{a}_{k-1}) = s_k\}$. By the sequential randomization assumption, $A_{j-1} \perp\!\!\!\perp \{S^*_j(\bar{a}_j), \ldots, S^*_{k+1}(\bar{a}_k)\} | \bar{S}_{j-1}, \bar{A}_{j-2}$, this expression is equal to $\mathrm{pr}\{S^*_{k+1}(\bar{a}_k) = s_{k+1} | \bar{S}_{j-1} = \bar{s}_{j-1}, \bar{A}_{j-2} = \bar{a}_{j-2}, S^*_j(\bar{a}_j) = s_j, \ldots, S^*_k(\bar{a}_{k-1}) = s_k\}$, which is (A.13) for $j - 1$. Note, then, that this implies that the conditional densities in (A.13), which are j-dependent, are the same as those on the left-hand side of (A.12), which are not and are also equal to $\mathrm{pr}\{S^*_{k+1}(\bar{a}_k) = s_{k+1} | \bar{S}^*_k(\bar{a}_{k-1}) = \bar{s}_k\}$; i.e., the left-hand side of (A.12) is completely in terms of the distribution of the observed data, whereas (A.13) with $j = 1$ is completely in terms of the distribution of the potential outcomes.

We now prove (A.11) by induction. Assume we have shown (A.11) for a fixed k; i.e., if $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$ and $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$, then $\mathrm{pr}(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) > 0$. Then we must show that

$$\mathrm{pr}(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_{k+1} = \bar{a}_{k+1}) > 0 \tag{A.14}$$

if $(\bar{s}_{k+1}, \bar{a}_k) \in \Gamma_{k+1}$ and $a_{k+1} \in \Psi_{k+1}(\bar{s}_{k+1}, \bar{a}_k)$. Note that

$$\mathrm{pr}(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k l = \bar{a}_k) = \mathrm{pr}(S_{k+1} = s_{k+1} | \bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k)\, \mathrm{pr} \bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k). \tag{A.15}$$

By (A.15), (A.14) will be true if $\mathrm{pr}(S_{k+1} = s_{k+1} | \bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) > 0$. But by the arguments above,

$$\mathrm{pr}(S_{k+1} = s_{k+1} | \bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) = \mathrm{pr}\{S^*_{k+1}(\bar{a}_k) = s_{k+1} | \bar{S}^*_k(\bar{a}_{k-1}) = \bar{s}_k\},$$

which is positive because $(\bar{s}_{k+1}, \bar{a}_k) \in \Gamma_{k+1}$. Next, $\mathrm{pr}(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_{k+1} = \bar{a}_{k+1}) =$ $pr(A_{k+1} = a_{k+1} | \bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k = \bar{a}_k)\mathrm{pr}(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k = \bar{a}_k)$. However, because $a_{k+1} \in \Psi_{k+1}(\bar{s}_{k+1}, \bar{a}_k)$, by the positivity assumption, $\mathrm{pr}(A_{k+1} = a_{k+1} | \bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k = \bar{a}_k) > 0$. The proof is completed by noting that $\mathrm{pr}(S_1 = s_1, A_1 = a_1) = \mathrm{pr}(A_1 = a_1 | S_1 = s_1)\mathrm{pr}(S_1 = s_1)$. If $s_1 \in \Gamma_1$, $\mathrm{pr}(S_1 = s_1) > 0$, and $\mathrm{pr}(A_1 = a_1 | S_1 = s_1) > 0$ for $a_1 \in \Psi(s_1)$ by the positivity assumption.

To demonstrate (19), consider first the definitions of $d^{(1)\mathrm{opt}}_K(\bar{s}_K, \bar{a}_{K-1})$ and $V^{(1)}_K(\bar{s}_K, \bar{a}_{K-1})$ given in (5) and (6). These quantities involve the conditional expectation of the potential outcome $Y^*(\bar{a}_K)$ given $\bar{S}^*_K(\bar{a}_{k-1})$, which by (A.12)-(A.13) is the same as the conditional expectation of Y given $\{\bar{S}_K = \bar{s}_K, \bar{A}_K = \bar{a}_K\}$. Thus, $d^{(1)\mathrm{opt}}_K(\bar{s}_K, \bar{a}_{K-1})$ and $V^{(1)}_K(\bar{s}_K, \bar{a}_{K-1})$ are the same as $d^{\mathrm{opt}}_K(\bar{s}_K, \bar{a}_{K-1})$ and $V_K(\bar{s}_K, \bar{a}_{K-1})$ defined in (10) and (11). Next, from (7) and (8), $d^{(1)\mathrm{opt}}_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2}) = \arg\max_{a_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} \mathrm{E}[V^{(1)}_K\{\bar{s}_{K-1}, S^*_K(\bar{a}_{K-2}, a_{K-1}), \bar{a}_{K-2}, a_{K-1}\} | \bar{S}^*_{K-1}(\bar{a}_{K-2}) =$ $\bar{s}_{K-1}]$. This involves the conditional expectation of $V^{(1)}_K$, a function of $S^*_K(\bar{a}_{K-1})$, given $\bar{S}^*_{K-1}(\bar{a}_{K-2}) =$ $\bar{s}_{K-1}$. Again, by (A.12)-(A.13), this is the same as the conditional expectation of the function $V^{(1)}_K$ of S_K given $\{\bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = \bar{a}_{K-1}\}$. Because we have already shown that $V^{(1)}_K$ is the same as V_K, this implies that $d^{(1)\mathrm{opt}}_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})$ is given by

$$\arg\max_{a_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} \mathrm{E}\{V_K(\bar{s}_{K-1}, S_K, \bar{a}_{K-2}, a_{K-1}) | \bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = (\bar{a}_{K-2}, a_{K-2})\},$$

which is the same as $d^{\text{opt}}_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})$ given by (13) with $k = K-1$. The argument continues in a backward iterative fashion for $k = K-2, \ldots, 1$.

A.3. JUSTIFICATION FOR $\widetilde{V}_{KI}$ IN A-LEARNING

We wish to show that

$$E\left(V_{k+1}(\bar{S}_{k+1}, \bar{A}_k) + C_k(\bar{S}_k, \bar{A}_{k-1})[\, I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k\,]\,\Big|\, \bar{S}_k, \bar{A}_{k-1}\right) = V_k(\bar{S}_k, \bar{A}_{k-1}). \tag{A.16}$$

Defining $\Gamma(\bar{S}_{k+1}, \bar{A}_k) = V_{k+1}(\bar{S}_{k+1}, \bar{A}_k) + C_k(\bar{S}_k, \bar{A}_{k-1})[\, I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k\,]$, we may write (A.16) as

$$\mathrm{E}[\,\mathrm{E}\{\Gamma(\bar{S}_{k+1}, \bar{A}_k)|\bar{S}_k, \bar{A}_k\}|\bar{S}_k, \bar{A}_{k-1}\,]. \tag{A.17}$$

The inner expectation in (A.17) may be seen to be equal to

$$\begin{aligned} &\mathrm{E}\{V_{k+1}(\bar{S}_{k+1}, \bar{A}_k)|\bar{S}_k, \bar{A}_k\} + C_k(\bar{S}_k, \bar{A}_{k-1})[\, I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k\,] \\ &\quad = Q_k(\bar{S}_k, \bar{A}_k) + C_k(\bar{S}_k, \bar{A}_{k-1})[\, I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k\,]. \end{aligned}$$

Substituting $Q_k(\bar{S}_k, \bar{A}_k) = h_k(\bar{S}_k, \bar{A}_{k-1}) + A_k C_k(\bar{S}_k, \bar{A}_{k-1})$, $h_k(\bar{S}_k, \bar{A}_{k-1}) = Q_k(\bar{S}_k, \bar{A}_{k-1}, 0)$, we obtain $\mathrm{E}\{\Gamma(\bar{S}_{k+1}, \bar{A}_k)|\bar{S}_k, \bar{A}_k\} = h_k(\bar{S}_k, \bar{A}_{k-1}) + C_k(\bar{S}_k, \bar{A}_{k-1}) I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} = V_k(\bar{S}_k, \bar{A}_{k-1})$. Substituting this in (A.17) yields the result.

A.4. DEMONSTRATION OF EQUIVALENCE OF Q- AND A-LEARNING IN A SPECIAL CASE

We take $K = 1$ and let $\mathrm{pr}(A_1 = 1|S_1 = s_1) = \pi$. Consider the A-learning estimating equations (31) with $k = 1$, and take $\lambda_1(s_1; \psi_1) = \partial/\partial\psi_1 C_1(s_1; \psi_1)$. Then the equations become

$$\sum_{i=1}^{n} \frac{\partial C_1(S_{1i}; \psi_1)}{\partial \psi_1}(A_{1i} - \pi)\{Y_i - A_{1i}C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1)\} = 0,$$

$$\sum_{i=1}^{n} \frac{\partial h_1(S_{1i}; \beta_1)}{\partial \beta_1}\{Y_i - A_{1i}C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1)\} = 0.$$

Likewise, under these conditions, taking $Q_1(s_1, a_1) = a_1 C_1(s_1; \psi_1) + h(s_1; \beta_1)$, the Q-learning equation is

$$\sum_{i=1}^{n} \frac{\partial Q_1(S_{1i}, A_{1i}; \xi_1)}{\partial \xi_1}\{Y_i - A_{1i}C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1)\} = 0,$$

where, with $\xi_1 = (\psi_1^T, \beta_1^T)^T$,

$$\frac{\partial Q_1(S_{1i}, A_{1i}; \xi_1)}{\partial \xi_1} = \begin{pmatrix} A_{1i}\dfrac{\partial C_1(S_{1i};\psi_1)}{\partial \psi_1} \\ \dfrac{\partial h_1(S_{1i};\beta_1)}{\partial \beta_1} \end{pmatrix}.$$

Thus note that, with $C_1(s_1;\psi_1)$ and $h_1(s_1;\beta_1)$ linear in functions of S_1, as long as terms of the form in $C_1(s_1;\psi_1)$ are contained in those in $h_1(s_1,\beta_1)$, the Q- and A-learning estimating equations are identical, as then

$$\sum_{i=1}^n \frac{\partial C_1(S_{1i};\psi_1)}{\partial \psi_1}\{Y_i - A_{1i}C_1(S_{1i};\psi_1) - h_1(S_{1i};\beta_1)\} = 0.$$

For example, if $C_1(s_1;\psi_1) = \psi_{10} + s_1^T\psi_{11}$ and $h_1(s_1;\beta_1) = \beta_{10} + s_1^T\beta_{11}$, then note that

$$\frac{\partial C_1(S_{1i};\psi_1)}{\partial \psi_1} = \frac{\partial h_1(S_{1i};\beta_1)}{\partial \beta_1} = \begin{pmatrix} 1 \\ S_{1i} \end{pmatrix},$$

and the result is immediate.

See Chakraborty et al. (2010) for discussion of the case $K = 2$.

A.5. EXAMPLE OF INCOMPATIBILITY OF Q-FUNCTION MODELS

To show (33), noting $\mathcal{H}_2 = (1, s_1, a_1, s_2)^T = (\mathcal{K}_1^T, s_2)^T$, we have

$$\begin{aligned} \mathrm{E}\{V_2(s_1, S_2, a_1; \xi_2)|S_1 = s_1, A_1 = a_1\} &= \mathcal{K}_1^T\beta_{21} + \beta_{22}\mathrm{E}(S_2|S_1 = s_1, A_1 = a_1) \\ &+(\mathcal{K}_1^T\psi_{21})\mathrm{E}\{I(\mathcal{K}_1^T\psi_{21} + S_2\psi_{22} > 0)|S_1 = s_1, A_1 = a_1\} \\ &+\psi_{22}\mathrm{E}\{S_2 I(\mathcal{K}_1^T\psi_{21} + S_2\psi_{22} > 0)|S_1 = s_1, A_1 = a_1\}. \end{aligned}$$

Taking $\psi_{22} > 0$, we also have $I(\mathcal{K}_1^T\psi_{21} + S_2\psi_{22} > 0) = I(S_2 > -\mathcal{K}_1^T\psi_{21}/\psi_{22})$, from which it follows that $\mathrm{E}\{I(\mathcal{K}_1^T\psi_{21} + S_2\psi_{22} > 0)|S_1 = s_1, A_1 = a_1\} = 1 - \Phi\{(-\mathcal{K}_1^T\psi_{21}/\psi_{22} - \mathcal{K}_1^T\gamma)/\sigma\} = 1 - \Phi(\eta)$ for $\eta = -\mathcal{K}_1^T(\psi_{21}/\psi_{22} + \gamma)/\sigma$. Similarly, $\mathrm{E}\{S_2 I(\mathcal{K}_1^T\psi_{21} + S_2\psi_{22} > 0)|S_1 = s_1, A_1 = a_1\} = \mathrm{E}\{S_2 I(S_2 > -\mathcal{K}_1^T\psi_{21}/\psi_{22})|S_1 = s_1, A_1 = a_1\}$. It is straightforward to deduce that this is equal to $\int_\eta^\infty (\sigma t + \mathcal{K}_1^T\gamma)\,\varphi(t)\,dt = \sigma\varphi(\eta) + (\mathcal{K}_1^T\gamma)\{1 - \Phi(\eta)\}$. Using $\mathrm{E}(S_2|S_1 = s_1, A_1 = a_1) = \mathcal{K}_1^T\gamma$ and combining yields (33).

A.6. CALCULATION OF E$\{H(\widehat{D}^{\mathrm{OPT}})\}$ AND $R(\widehat{D}^{\mathrm{OPT}})$

Calculation for $K = 1$. We consider the generative data model in Section 6.1 and treatment regimes of the form $d(s_1) = d_1(s_1) = I(\psi_{10} + \psi_{11}s_1 > 0)$ for arbitrary ψ_{10}, ψ_{11}. It is possible to derive analytically $H(d) = \mathrm{E}\{Y^*(d)\}$ in this case. Under the generative data model, $\mathrm{E}\{Y^*(d)\} = \mathrm{E}[\mathrm{E}\{Y^*(d)|S_1\}] = \mathrm{E}[\mathrm{E}\{Y|S_1, A_1 = d_1(S_1)\}] = \beta^0_{10} + \beta^0_{11}\mathrm{E}(S_1) + \beta^0_{12}\mathrm{E}(S_1^2) + \mathrm{E}\{I(\psi_{10} + \psi_{11}S_1 > 0)(\psi^0_{10} + \psi^0_{11}S_1)\}$, and $S_1 \sim \mathrm{Normal}(0, 1)$. It is straightforward to deduce that $\mathrm{E}\{I(\psi_{10} + \psi_{11}S_1 > 0)\} = \mathrm{pr}(S_1 > -\psi_{10}/\psi_{11})$ or $\mathrm{pr}(S_1 < -\psi_{10}/\psi_{11})$ as $\psi_{11} > 0$ or $\psi_{11} < 0$, which is readily obtained from the standard normal cdf. Likewise, $\mathrm{E}\{S_1 I(\psi_{10}+\psi_{11}S_1 > 0)\} = \mathrm{E}(S_1|S_1 > -\psi_{10}/\psi_{11})\mathrm{pr}(S_1 > -\psi_{10}/\psi_{11})$ if $\psi_{11} > 0$ and $\mathrm{E}\{S_1 I(\psi_{10} + \psi_{11}S_1 > 0)\} = \mathrm{E}(S_1|S_1 < -\psi_{10}/\psi_{11})\mathrm{pr}(S_1 < -\psi_{10}/\psi_{11})$ if $\psi_{11} < 0$, which are again easily calculated in a manner similar to that in Section A.5. Thus, $\mathrm{E}\{Y^*(d^{\mathrm{opt}})\}$ is obtained by substituting ψ^0_{10}, ψ^0_{11} in the resulting expression. To approximate $\mathrm{E}\{H(\widehat{d}^{\mathrm{opt}})\}$ and hence $R(\widehat{d}^{\mathrm{opt}})$ for $\widehat{d}^{\mathrm{opt}} = \widehat{d}^{\mathrm{opt}}_Q$ or $\widehat{d}^{\mathrm{opt}}_A$, we may use Monte Carlo simulation. Specifically, for the bth of B Monte Carlo data sets, substitute the estimates $\widehat{\psi}_{10,b}$, $\widehat{\psi}_{11,b}$, say, defining $\widehat{d}^{\mathrm{opt}}$ for that data set in the expression for $\mathrm{E}\{Y^*(d)\}$, and call the resulting quantity U_b. Then $\mathrm{E}\{H(\widehat{d}^{\mathrm{opt}})\}$ is approximated by $B^{-1}\sum_{b=1}^{B} U_b$. Combining yields the approximation to $R(\widehat{d}^{\mathrm{opt}})$.

Calculation for $K = 2$. The developments are analogous to those above. We consider the generative data model in Section 6.2 and treatment regimes of the form $d = (d_1, d_2)$, where $d_1(s_1) = I(\psi_{10} + \psi_{11}s_1 > 0)$ and $d_2(s_1, s_2, a_1) = I(\psi_{20} + \psi_{21}a_1 + \psi_{22}s_2 > 0)$ for arbitrary $\psi_{10}, \psi_{11}, \psi_{20}, \psi_{21}, \psi_{22}$. Here, $\mathrm{E}\{Y^*(d)\} = \mathrm{E}\Big(\mathrm{E}[\mathrm{E}\{Y^*(d)|S_2^*(d), S_1\}|S_1]\Big) = \mathrm{E}\Big\{\mathrm{E}\Big(\mathrm{E}[Y|S_2, S_1, A_1 = d_1(S_1), A_2 = d_2\{S_2, S_1, d_1(S_1)\}]\,\Big|S_1, A_1 = d_1(S_1)\Big)\Big\}$. Because S_1 is binary taking values in $\{0, 1\}$, $\mathrm{E}\{Y^*(d)\} = \mathrm{E}\Big(\mathrm{E}[Y|S_2, S_1, A_1 = d_1(0), A_2 = d_2\{S_2, 0, d_1(0)\}]\,\Big|S_1 = 0, A_1 = d_1(0)\Big)\,\mathrm{pr}(S_1 = 0) + \mathrm{E}\Big(\mathrm{E}[Y|S_2, S_1, A_1 = d_1(1), A_2 = d_2\{S_2, 1, d_1(1)\}]\,\Big|S_1 = 1, A_1 = d_1(1)\Big)\,\mathrm{pr}(S_1 = 1)$. Under the generative model, writing $a_1 = I(\psi_{10} + \psi_{11}s_1 > 0)$ for brevity, these expectations are of the form $\mathrm{E}\Big(\mathrm{E}[Y|S_2, S_1, A_1 = d_1(s_1), A_2 = d_2\{S_2, s_1, d_1(s_1)\}]\,\Big|S_1 = s_1, A_1 = d_1(s_1)\Big) = \beta_{20} + \beta^0_{21}s_1 + \beta^0_{22}a_1 + \beta^0_{23}s_1a_1 + \beta^0_{24}\mathrm{E}\{(S_2|S_1 = s_1, A_1 = d_1(s_1)\} + \beta^0_{25}\mathrm{E}\{S_2^2|S_1 = s_1, A_1 = d_1(s_1)\} + (\psi^0_{20} + \psi^0_{21}a_1)\mathrm{E}\{I(\psi_{20} + \psi_{21}a_1 + \psi_{22}S_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\} + \psi^0_{22}\mathrm{E}\{S_2 I(\psi_{20} + \psi_{21}a_1 + \psi_{22}S_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\}$, for $s_1 = 0, 1$. In the generative data model, the conditional distribution of S_2 given S_1, A_1 is normal; accordingly, it is straightforward to calculate $\mathrm{E}\{S_2|S_1 = s_1, A_1 = d_1(s_1)\}$, $\mathrm{E}\{S_2^2|S_1 = s_1, A_1 = d_1(s_1)\}$, $\mathrm{E}\{I(\psi_{20} + \psi_{21}a_1 + \psi_{22}S_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\}$, and $E\{S_2 I(\psi_{20} + \psi_{21}a_1 + \psi_{22}S_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\}$ in a manner analogous to those for the case $K = 1$. Approximation of $\mathrm{E}\{H(\widehat{d}^{\mathrm{opt}})\}$ and hence $R(\widehat{d}^{\mathrm{opt}})$ for $\widehat{d}^{\mathrm{opt}} = \widehat{d}^{\mathrm{opt}}_Q$ or $\widehat{d}^{\mathrm{opt}}_A$ may then be carried out as for the case $K = 1$.

Calculation by simulation. When an analytical expression for $H(d) = \mathrm{E}\{Y^*(d)\}$ for regimes of a certain form d is not available, $H(d)$ for a fixed d may be approximated by simulation using the g-computation algorithm of Robins (1986). We demonstrate for $K = 2$, so that $d = (d_1, d_2)$; the procedure for $K = 1$ is then immediate. For total number of simulations B, for each $b = 1, \ldots, B$, the steps are: (i) Generate s_{1b} from the true distribution of S_1; (ii) generate s_{2b} from the true conditional distribution of S_2 given $S_1 = s_{1b}$ and $A_1 = d_1(s_{1b})$; (iii) evaluate the true $\mathrm{E}(Y|\bar{S}_2 = \bar{s}_2, \bar{A}_2 = \bar{a}_2)$ at $\bar{s}_2 = \bar{s}_{2b} = (s_{1b}, s_{2b})$ and $\bar{a}_2 = [d_1(s_{1b}), d_2\{\bar{s}_{2b}, d_1(s_{1b})\}]$, and call the resulting value U_b; and (iv) estimate $H(d) = \mathrm{E}\{Y^*(d)\}$ by $B^{-1}\sum_{b=1}^{B} U_b$. When $d = \widehat{d}_Q^{\text{opt}}$ or $\widehat{d}_A^{\text{opt}}$, one would follow the above procedure for each Monte Carlo data set. In each of steps (i)–(iii), it is important to recognize that, while $\widehat{d}_Q^{\text{opt}}$ and $\widehat{d}_A^{\text{opt}}$ are determined by the estimated ψ, the distributions from which realizations are generated depend on the true β and ψ. The values of $\mathrm{E}\{H(\widehat{d}_Q^{\text{opt}})\}$ and $\mathrm{E}\{H(\widehat{d}_A^{\text{opt}})\}$ may then be approximated by the average of the estimated $H(\widehat{d}_Q^{\text{opt}})$ and $H(\widehat{d}_Q^{\text{opt}})$ across the Monte Carlo data sets, as before.

A.7. CREATING "EQUIVALENTLY MISSPECIFIED PAIRS" WHEN BOTH THE PROPENSITY MODEL AND Q-FUNCTION ARE MISSPECIFIED

Consider the $K = 2$ decision point scenario; the developments apply equally to the $K = 1$ setting. To identify pairs $(\beta_{25}^0, \phi_{25}^0)$ that are "equivalently misspecified," for each of the combinations of β_{25}^0 and ϕ_{25}^0 within a pre-specified grid, say $(\beta_{25}^0, \phi_{25}^0) \in [-1, 1] \times [-1, 1]$ with a step size of 0.05, we generate a large data set of size $n = 10,000$ from the generative data model in Section 6.2 with all other parameters fixed at their true values. This yields $41 \times 41 = 1681$ combinations and hence such data sets. For each data set, the linear regression model for the response and the logistic model for propensity of treatment assignment are then fitted, and the ratio of standard errors for $\widehat{\phi}_{25}$ and $\widehat{\beta}_{25}$, $SE(\widehat{\phi}_{25})/SE(\widehat{\beta}_{25})$, say, obtained. We then fit to these values a polynomial model in ϕ_{25}^0, $f(\phi_{25}^0)$, say, and select the polynomial degree yielding a sufficiently large adjusted R^2. Setting $\beta_{25}^0 = \phi_{25}^0/f(\phi_{25}^0)$ then yields the result that the corresponding t-statistics will be approximately equal. These were re-checked in the course of running the simulations so that the t-statistics differed by less than some reasonable value, usually at most a 5 percent difference, as it cannot be guaranteed that they will be precisely the same.

A.8. DERIVATION OF $H_1^0(S_1;\beta_1^0)$ AND $C_1^0(S_1;\psi_1^0)$ IN THE TWO DECISION POINT SCENARIO

We seek to identify the true $h_1^0(s_1)$ and $C_1^0(s_1)$, where S_1 and A_1 are Bernoulli. With $h_1^0(s_1) = \beta_{10}^0 + \beta_{11}^0 s_1$ and $C_1^0(s_1) = \psi_{10}^0 + \psi_{11}^0 s_1$, it follows that the true Q-function at the first decision is $Q_1^0(s_1, a_1) = h_1^0(s_1) + a_1 C_1^0(s_1)$. We thus calculate $Q_1^0(s_1, a_1)$ under the generative model and equate terms to determine the form of β_{10}^0, β_{11}^0, ψ_{10}^0, and ψ_{11}^0. The true value function at the second decision is $V_2^0(S_1, S_2, A_1) = h_2^0(S_1, S_2, A_1) + C_2^0(S_1, S_2, A_1)I\{C_2^0(S_1, S_2, A_1) > 0\}$. Thus, $Q_1^0(s_1, a_1) = \mathrm{E}\{V_2^0(S_1, S_2, A_1)|S_1 = s_1, A_1 = a_1\} = \beta_{20}^0 + \beta_{21}^0 s_1 + \beta_{22}^0 a_1 + \beta_{23}^0 s_1 a_1 + \beta_{24}^0 \mathrm{E}\{S_2|S_1 = s_1, A_1 = a_1\} + \beta_{25}^0 \mathrm{E}\{S_2^2|S_1 = s_1, A_1 = a_1\} + \mathrm{E}\{C_2^0(S_1, S_2, A_1)I\{C_2^0(S_1, S_2, A_1) > 0)|S_1 = s_1, A_1 = a_1\}$. The conditional expectations in this expression may be calculated in a manner analogous to that in Section A.5 to obtain the form of $Q_1^0(s_1, a_1)$. It follows that $Q_1^0(0, 0) = \beta_{10}^0$, $Q_1^0(1, 0) = \beta_{10}^0 + \beta_{11}^0$, $Q_1^0(0, 1) = \beta_{10}^0 + \psi_{10}^0$, and $Q_1^0(1, 1) = \beta_{10}^0 + \beta_{11}^0 + \psi_{10}^0 + \psi_{11}^0$, which may be solved to yield expressions for β_{10}^0, β_{11}^0, ψ_{10}^0, and ψ_{11}^0.

A.9. ADDITIONAL SIMULATION RESULTS

In this section, we present additional summaries of results of the simulations reported in Sections 6.1 and 6.2.

The quantities $R(\widehat{d}_Q^{\mathrm{opt}})$ and $R(\widehat{d}_A^{\mathrm{opt}})$ plotted in Figures 1–6 are the v-efficiencies of the estimated regimes produced by each method, as discussed at the beginning of Section 6 of the parent article. These are based on the averages $E\{H(\widehat{d}_Q^{\mathrm{opt}})\}$ and $E\{H(\widehat{d}_A^{\mathrm{opt}})\}$. Because the distributions of the $H(\widehat{d}_Q^{\mathrm{opt}})$ and $H(\widehat{d}_A^{\mathrm{opt}})$ may be left-skewed, it is instructive to also present both the distributions themselves and alternatives to $R(\widehat{d}_Q^{\mathrm{opt}})$ and $R(\widehat{d}_A^{\mathrm{opt}})$ based on the median.

The left-hand panel of Figure A.1 is the same as the right-most panel of Figure 1 in the main paper shown for convenience. For comparison, the right-hand panel shows both these distributions and the alternative measures of estimated regime efficiency based on medians rather than averages. Figures A.2 and A.3 show the same corresponding to Figures 2 and 3 in the main paper.

Likewise, Figures A.4–A.6 show the same information corresponding to the lower right-hand panels of Figures 4–6 in the main article.

In nearly all cases, the plots using the median dominate those using the mean, reflecting the expected skewness.

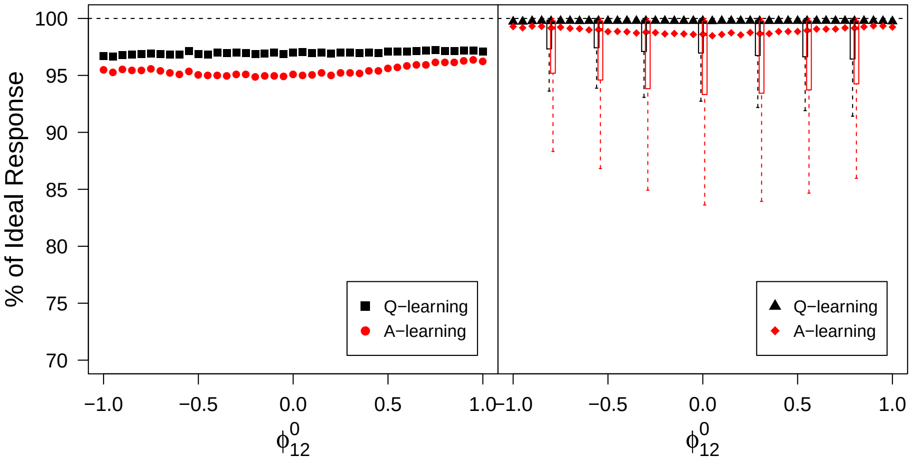

FIG A.1. *One decision point, misspecified propensity model. Left panel is same as the right-most panel of Figure 1 of the main article. Right panel shows alternative measure of efficiency based on medians and their distributions.*

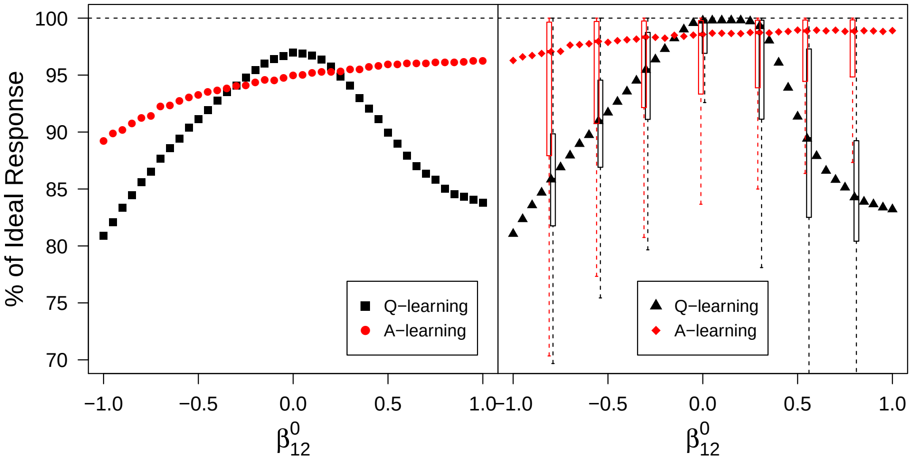

FIG A.2. *One decision point, misspecified Q-function. Left panel is same as the right-most panel of Figure 2 of the main article. Right panel shows alternative measure of efficiency based on medians and their distributions.*

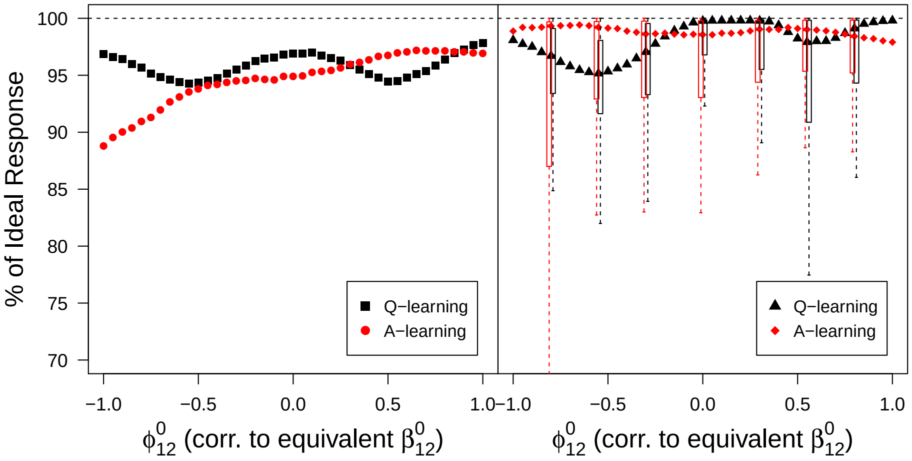

FIG A.3. *One decision point, both models misspecified. Left panel is same as the right-most panel of Figure 3 of the main article. Right panel shows alternative measure of efficiency based on medians and their distributions.*

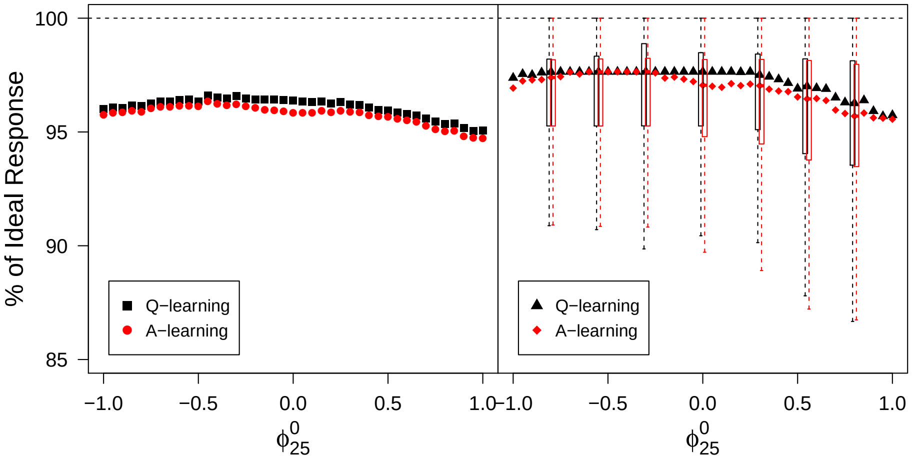

FIG A.4. *Two decision points, misspecified propensity model. Left panel is same as the lower right-most panel of Figure 4 of the main article. Right panel shows alternative measure of efficiency based on medians and their distributions.*

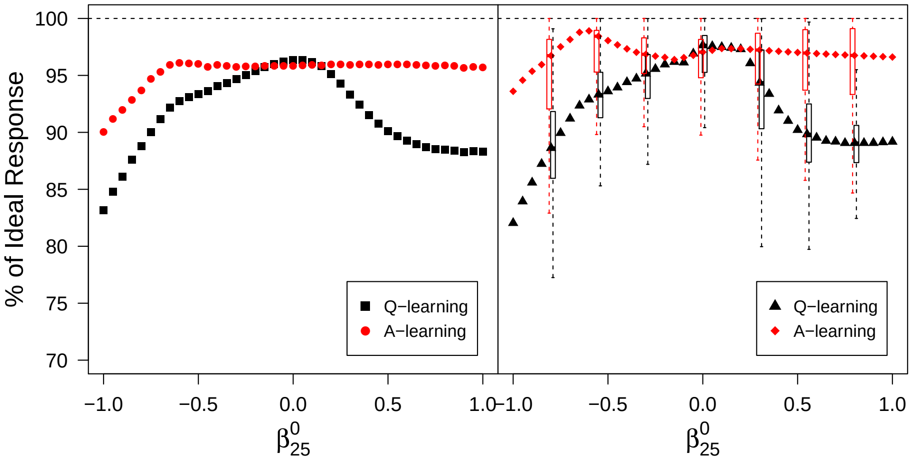

FIG A.5. *Two decision points, misspecified Q-function. Left panel is same as the lower right-most panel of Figure 5 of the main article. Right panel shows alternative measure of efficiency based on medians and their distributions.*

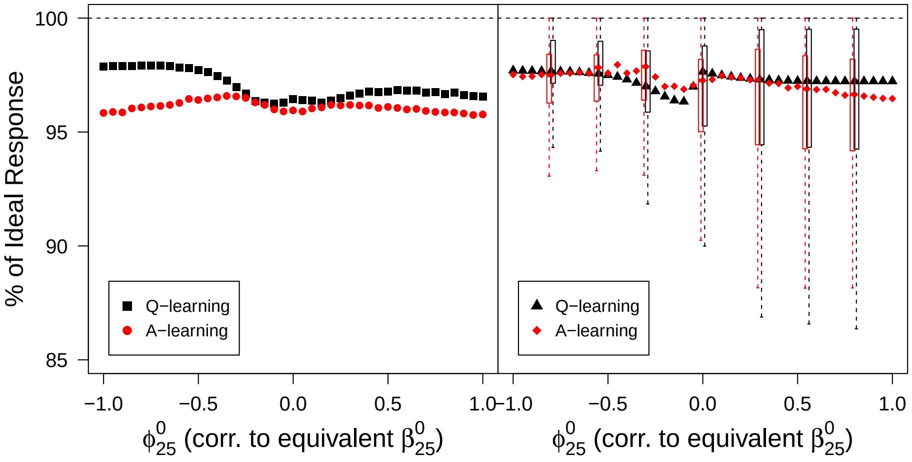

FIG A.6. *Two decision points, both models misspecified. Left panel is same as the right-most panel of Figure 6 of the main article. Right panel shows alternative measure of efficiency based on medians and their distributions.*

For each level of misspecification in the misspecified propensity score scenario for one decision, Figure A.7 plots the average across all simulated data sets of the standard deviations of the estimated propensity scores for each simulated data set. The figure shows that as the level of misspecification increases, the standard deviation decreases, as consistent with the claim in the main article.

The bottom panels of Figure A.7 show similar results for the two decision point misspecified propensity scenario. For the second decision point, shown in the bottom right panel of Figure A.7, where the propensity model is misspecified, the behavior is similar (standard deviation decreases as level of misspecification increases). For comparison, the bottom left panel of Figure A.7 shows the behavior of the estimated propensities at the first decision point, where the propensity model is correctly specified; as expected, there is no discernible pattern.

We also carried out simulations under the case of misspecified contrast function in a scenario similar to that in Section 6.1. Here, we used the class of generative models defined in (34) with the exception that

$$Y|S_1 = s_1, A_1 = a_1 \sim \text{Normal}\{\beta_{10}^0 + \beta_{11}^0 s_1 + \beta_{12}^0 s_1^2 + a_1(\psi_{10}^0 + \psi_{11}^0 s_1 + \psi_{12}^0 s_1^2), 9\},$$

so that $\psi^0 = (\psi_{10}^0, \psi_{11}^0, \psi_{12}^0)^T$, and thus $d^{\text{opt}} = d_1^{\text{opt}}$, $d_1^{\text{opt}}(s_1) = I(\psi_{10}^0 + \psi_{11}^0 s_1 + \psi_{12}^0 s_1^2 > 0)$. For A-learning, we assumed models $h_1(s_1;\beta_1) = \beta_{10} + \beta_{11}s_1$, $C_1(s_1;\psi_1) = \psi_{10} + \psi_{11}s_1$, and $\pi_1(s_1;\phi_1) = \text{expit}(\phi_{10} + \phi_{11}s_1)$, and for Q-learning used $Q_1(s_1, a_1; \xi_1) = h_1(s_1;\beta_1) + a_1 C_1(s_1;\psi_1)$. To simplify notation and distinguish between different components of the Q-function, we refer to $h_1(s_1;\beta_1)$ as the h-function.

These models involve correctly specified contrast functions only when $\psi_{12}^0 = 0$. The h-function is correctly specified when $\beta_{12}^0 = 0$, and the propensity model, $\pi_1(s_1;\phi_1)$, is correctly specified when $\phi_{12}^0 = 0$. We studied the effects of misspecification of the contrast function by systematically varying β_{12}^0, ϕ_{12}^0, and ψ_{12}^0 while keeping the others fixed, considering parameter settings of the form $\phi^0 = (0, -2, \phi_{12}^0)^T$, $\beta^0 = (1, 1, \beta_{12}^0)^T$, and $\psi^0 = (1, 0.5, \psi_{12}^0)^T$.

Three scenarios were considered:

1. misspecified contrast function alone ($\beta_{12}^0 = \phi_{12}^0 = 0$, and nonzero ψ_{12}^0)
2. misspecified contrast and h-function ($\phi_{12}^0 = 0$, and nonzero β_{12}^0 and ψ_{12}^0)
3. misspecified contrast and propensity model ($\beta_{12}^0 = 0$, and nonzero ϕ_{12}^0 and ψ_{12}^0).

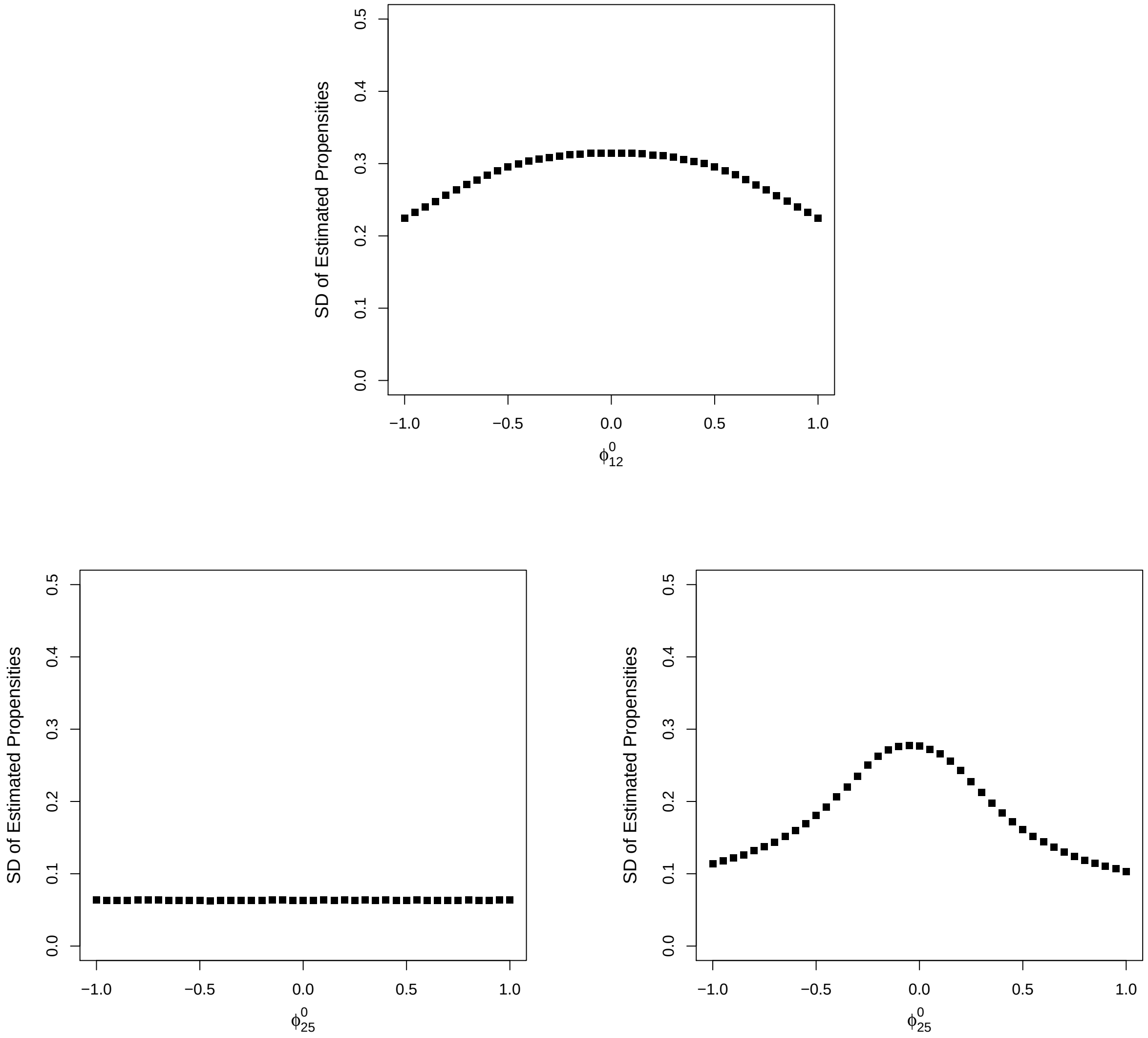

FIG A.7. *Misspecified propensity model. Each symbol represents the average across all simulated data sets of the standard deviation of the estimated propensity scores for each data set.* ***Top:*** *one decision point.* ***Lower left:*** *two decision points; first stage.* ***Lower right:*** *two decision points; second stage.*

Under scenarios 2 and 3, we identified $\beta_{12}^0 = \beta_{12}^0(\psi_{12}^0)$ and $\psi_{12}^0 = \psi_{12}^0(\phi_{12}^0)$, respectively, using the approach outlined in Section A.7, so that an analyst might be equally likely to detect either form of misspecification. Further, we observe that in the second and third scenarios, $\mathrm{sgn}(\beta_{12}^0) = \mathrm{sgn}(\psi_{12}^0)$ and $\mathrm{sgn}(\phi_{12}^0) = \mathrm{sgn}(\psi_{12}^0)$, respectively, where $\mathrm{sgn}(x) = I(x > 0)$.

Figure A.8 presents the v-efficiencies of estimators for d^{opt} under Q- and A-learning for each scenario. Here, the true d^{opt} is an indicator for whether the quadratic function $\psi_{10}^0 + \psi_{11}^0 s_1 + \psi_{12}^0 s_1^2$ is positive. For the given parameterization, when $\psi_{12}^0 > 0$ the true $d_1^{\text{opt}}(s_1) = 1$ for all s_1. However, estimators for d^{opt} under Q- and A-learning assume a linear contrast function such that $\widehat{d}_Q^{\text{opt}}(s_1) = 0$ or $\widehat{d}_A^{\text{opt}}(s_1) = 0$ for some s_1. This is illustrated in all three panels by $R(\widehat{d}_Q^{\text{opt}}) < 1$ and $R(\widehat{d}_A^{\text{opt}}) < 1$.

The top panel of Figure A.8 indicates that, when only the contrast function is misspecified (scenario 1) and when $\psi_{12}^0 \leq 0$, a small gain in v-efficiency is achieved by $\widehat{d}_Q^{\text{opt}}$ over $\widehat{d}_A^{\text{opt}}$. Alternatively, for $\psi_{12}^0 > 0$, $\widehat{d}_A^{\text{opt}}$ yields slightly better performance as ψ_{12}^0 increases.

For the second scenario, we considered misspecification of the contrast and h-function. The results are shown in the bottom left panel of Figure A.8. Similar to the prior scenario, we observe small gains in v-efficiency for $\widehat{d}_Q^{\text{opt}}$ over $\widehat{d}_A^{\text{opt}}$ when $\psi_{12}^0 \leq 0$ (and $\beta_{12}^0 = \beta_{12}^0(\psi_{12}^0) \leq 0$) and superior performance under A-learning for some values of $\psi_{12}^0 > 0$. Finally, the bottom right panel shows the results for the scenario of both misspecified contrast and misspecified propensity (scenario 3). Both Q- and A-learning yield estimators for d^{opt} that exhibit poor v-efficiency for much of the range where $\phi_{12}^0 < 0$ (and $\psi_{12}^0 < 0$), while $\widehat{d}_Q^{\text{opt}}$ shows better v-efficiency relative to $\widehat{d}_A^{\text{opt}}$ when $\phi_{12}^0 > 0$ (and $\psi_{12}^0 > 0$).

These results demonstrate that neither method need dominate the other in terms of performance as reflected by v-efficiency when the contrast function is misspecified.

A.10. DESIGN OF STAR*D

Figure A.9 presents a schematic of the STAR*D study design. At levels 2 and 3, patients/physicians expressed preference for switch or augment, and were then randomized to the options shown. Patients entering level 2A were randomized; those entering level 4 were not.

A.11. EXPLORATORY ANALYSIS AND DIAGNOSTICS

The development of the posited Q-functions used in the analysis of STAR*D data in Section 7 of the main paper relied on input from the investigators. The information included in S_1 and the

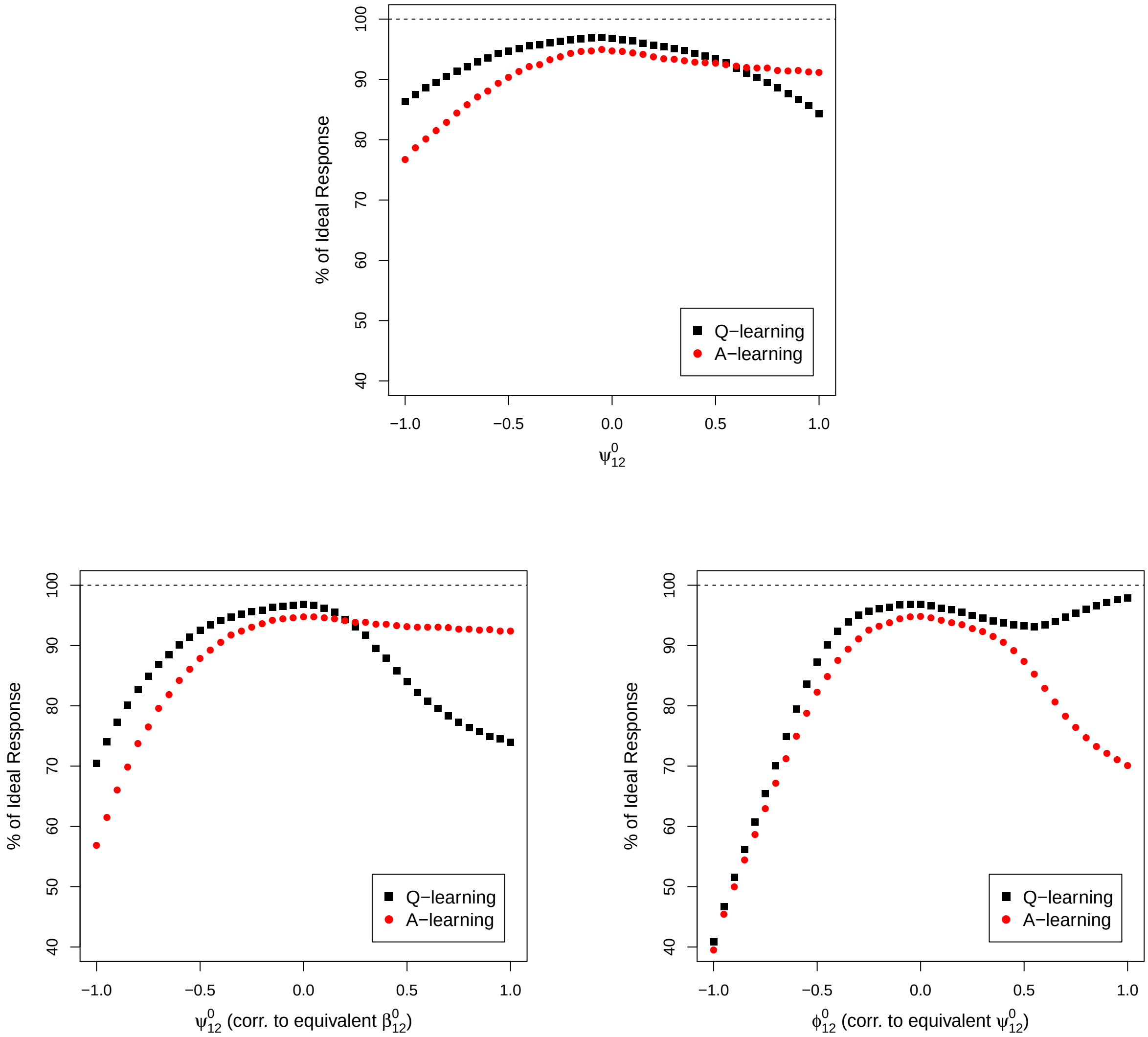

FIG A.8. *V-efficiencies $R(\widehat{d}_Q^{opt})$ and $R(\widehat{d}_A^{opt})$ for estimating the true d^{opt} (right panel) under misspecification of the contrast model, one decision point.* ***Top:*** *misspecified contrast function only.* ***Lower left:*** *misspecified contrast function and misspecified h-function.* ***Lower right:*** *misspecified contrast function and misspecified propensity model.*

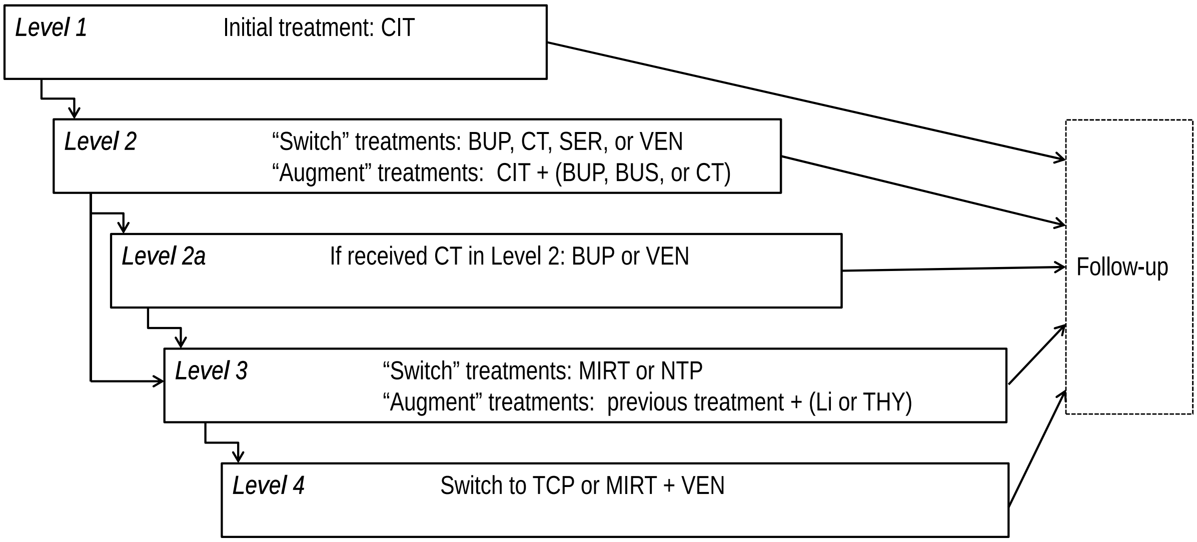
Level 1
Initial treatment: CIT
Level 2
"Switch" treatments: BUP, CT, SER, or VEN
"Augment" treatments: CIT + (BUP, BUS, or CT)
Level 2a
If received CT in Level 2: BUP or VEN
Level 3
"Switch" treatments: MIRT or NTP
"Augment" treatments: previous treatment + (Li or THY)
Level 4
Switch to TCP or MIRT + VEN
Follow-up

FIG A.9. *Schematic depiction of the STAR*D study.*

dependence of the models on QIDS slope was guided by discussions with John Rush, the principal investigator.

We present diagnostic plots for the first and second stage regressions used in the STAR*D analysis. Figure A.10 displays standard regression diagnostics for the second stage regression models used in Q- and A-learning. The figure suggests that there are no major deviations from the assumed linear model and that there is no evidence of outliers or high-influence points. Figure A.11 displays regression diagnostics for the first stage regression models used in Q- and A-learning. The figure suggests that a linear model may be a satisfactory approximation, though it appears that the linear model overestimates the first stage Q-function for non-responders and underestimates the first stage Q-function for responders. Recall that, by design, responders have higher responses at the end of the first stage. Thus, responders tend to have higher fitted values and positive residuals. There do not appear to be any outliers or high-influence points.

ACKNOWLEDGMENTS

This work was supported by NIH grants R37 AI031789, R01 CA051962, R01 CA085848, P01 CA142538, and T32 HL079896.

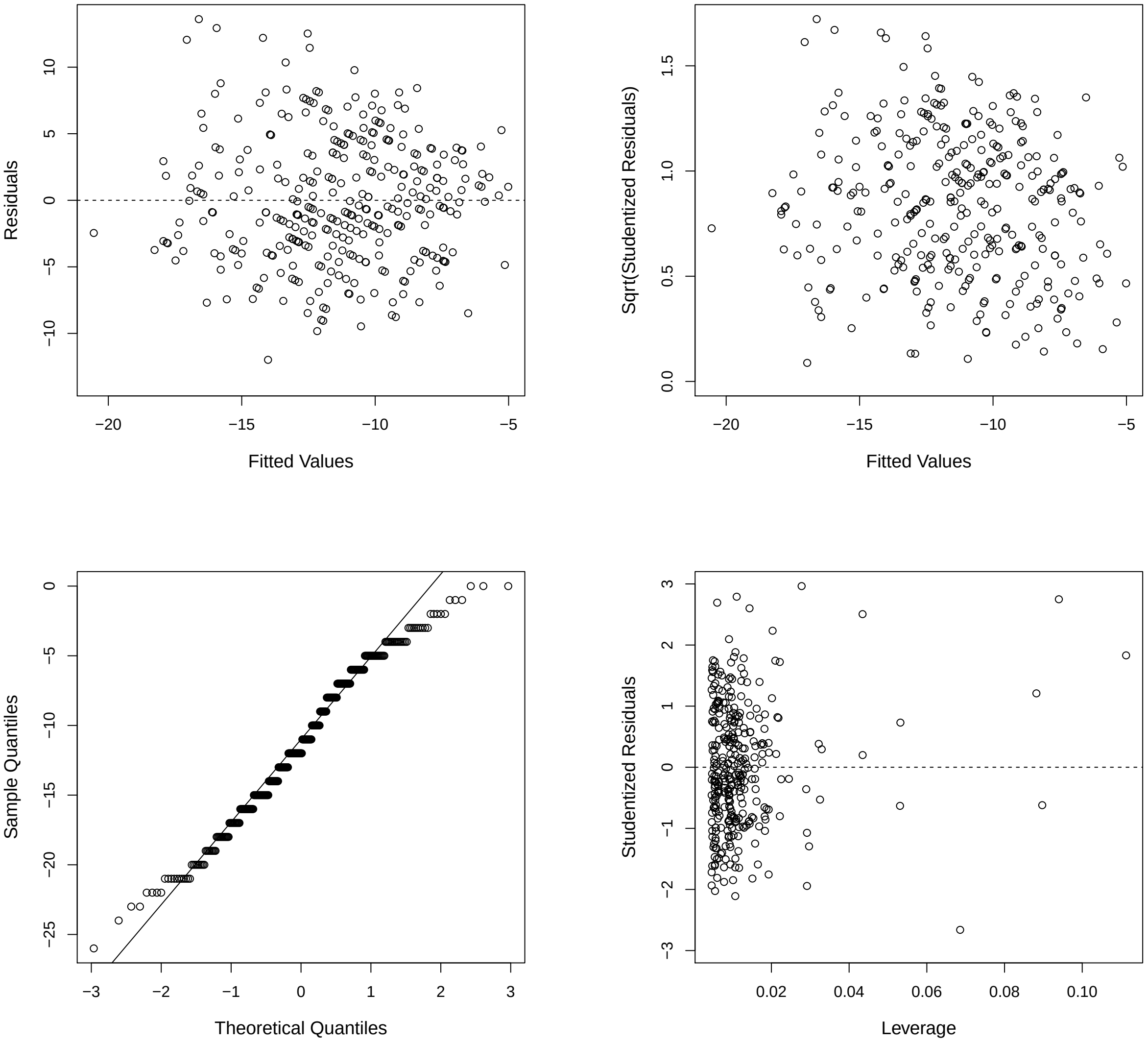

FIG A.10. *Regression diagnostics for the second stage Q-function used in Q- and A-learning.* ***Upper left:*** *displays residuals vs. fitted values; the figure does not suggest deviation from an underlying linear model with homoscedastic variance, and there do not appear to be any outliers.* ***Upper right:*** *displays the studentized residuals vs. the fitted values; the plot does not suggest deviation form an underlying linear model with homoscedastic variance, and there do not appear to be any outliers.* ***Lower left:*** *displays a QQ-plot of the residuals; note that the sample quantiles are a step-function because we have treated the discrete variable QIDS as continuous.* ***Lower right:*** *displays the studentized residuals vs. leverage; there do not appear to be any high-influence points.*

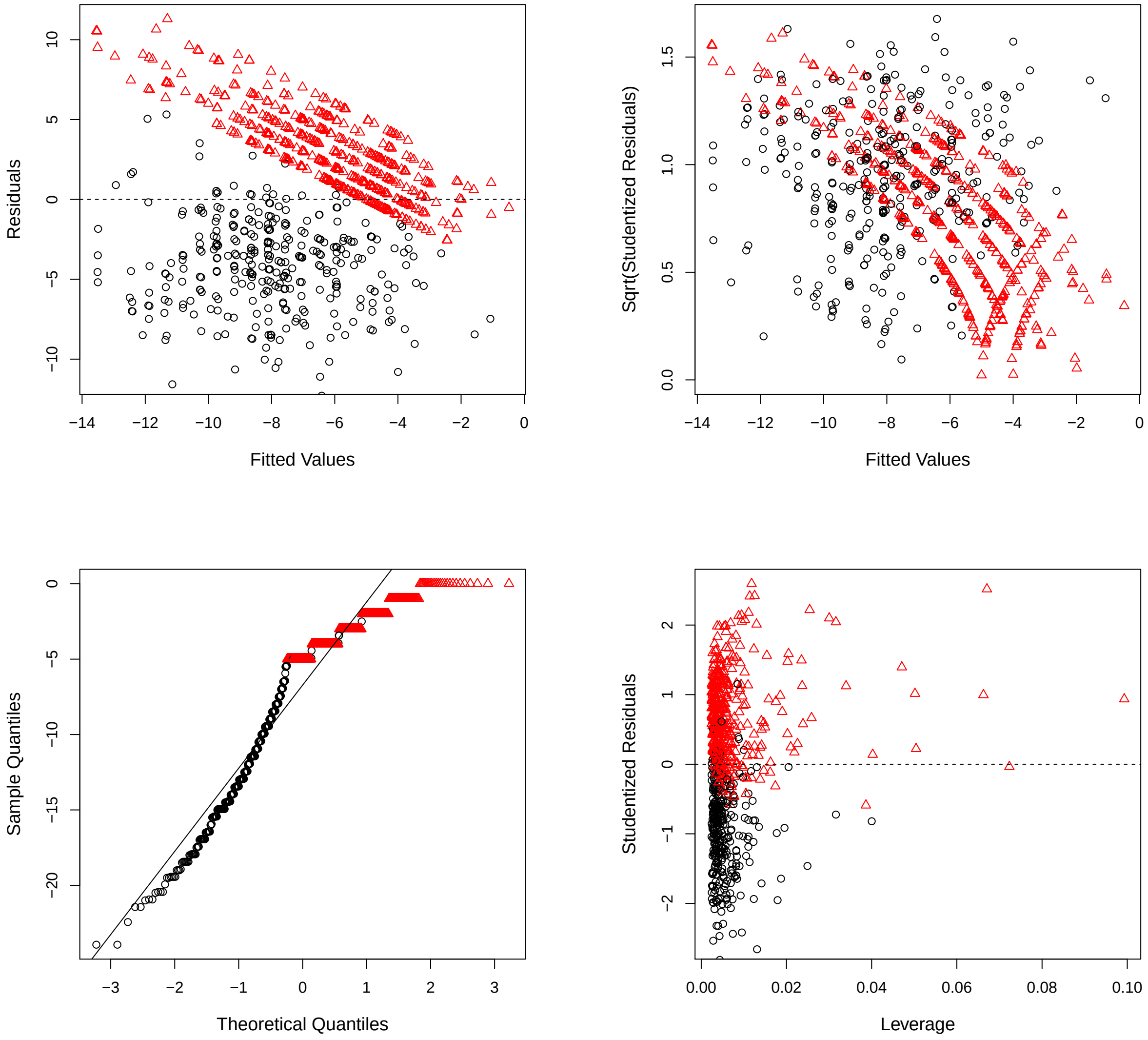

Fig A.11. *Regression diagnostics for the first stage Q-function used in Q- and A-learning. Responders are depicted with a △ and non-responders with a ○.* ***Upper left:*** *displays residuals vs. fitted values; there is a clear distinction between the responders and non-responders. Responders, by design, have higher first stage responses and thus tend to have higher fitted values and positive residuals.* ***Upper right:*** *displays the studentized residuals vs. the fitted values; again, the separation between responders and non-responders is clear.* ***Lower left:*** *displays a QQ-plot of the residuals; note that the sample quantiles are a step-function because we have treated the discrete variable QIDS as continuous.* ***Lower right:*** *displays the studentized residuals vs. leverage; there do not appear to be any high-influence points.*

REFERENCES

Chakraborty, B., Murphy, S. A. and Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research* **19** 317–343.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods: Applications to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512.