Electronic Journal of Statistics
Vol. 8 (2014) 1273–1289
ISSN: 1935-7524
DOI: 10.1214/14-EJS908

Discussion of "Dynamic treatment regimes: Technical challenges and applications"*

James Robins†

Dep. of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA, 02115, USA
e-mail: robins@hsph.harvard.edu

and

Andrea Rotnitzky†

Dep. of Economics, Di Tella University, Buenos Aires, 1425, Argentina
e-mail: arotnitzky@utdt.edu

Received May 2014.

We thank the editor for organizing this discussion of the article by Laber et al. (2014) (throughout referred to as LLQPM). The authors offer an elegant solution to the inferential problem caused by nonregularity. Our discussion will to a large extent focus on conceptual rather than technical issues, in part because the authors handled the technical matters so decisively and well. In so doing, we recognize that discussion of conceptual issues was not the authors' goal and that the authors have written elsewhere about many of the issues we raise. Indeed, in our own writing, we have often either ignored the issues we raise or were unable to offer coherent solutions to them. We hope our discussion makes for an interesting and lively interchange.

We first address the following conceptual issue. The author's target of inference is the stage one nonregular parameter β_{11}^* that determines the optimal treatment strategy π_1^{dp} at stage one of their two-stage trial. Robins (2004, Sec. 5) first recognized that β_{11}^* was nonregular and offered a method for obtaining a valid (necessarily conservative) confidence interval. However, in that section, Robins also noted that β_{11}^* only determines the optimal treatment decision at stage one for patients who will follow the optimal strategy at stage two; hence, because of uncertainty, it is not possible to know that the optimal strategy π_2^{dp} will in fact be followed at time two (even when we assume all the uncertainty is attributable to sampling variability), and therefore it is unclear that β_{11}^* should

*Main article 10.1214/14-EJS920.

†James Robins and Andrea Rotnitzky were partially supported by grant 2 R37 AI032475-16A 1 from the US from the National Institutes of Health.

be the target of inference. In light of these facts, he justified the study of inference for β_{11}^{*} because of its mathematical rather than its clinical or public health interest. Below we will argue that, in many settings, the first stage parameter of clinical interest is not β_{11}^{*} but rather the random parameter that determines the optimal stage one strategy conditional on the treatment regime that one will indeed follow at stage two. We show the estimation of this random parameter is a regular problem if one (i) splits the sample and chooses the regime at stage two to be the one estimated as optimal from one half of the sample and (ii) then estimates the aforementioned random parameter from the other half of the data. However, we also show that if one estimates the stage one and two regimes from the same sample, i.e. without sample splitting, then the problem becomes irregular. We will show that the conservative interval estimator we are led to is, interestingly, precisely the ACI interval estimator of the authors, even though our targets of inference differ.

We then address a second conceptual issue. How should univariate stage-specific confidence intervals be used in clinical decision making? The authors suggest that if a confidence interval (say a 90% interval) for $\beta_{t,1}^{*T} h_{t,1}$ includes zero, then trialists should make no recommendation for the treatment that should be administered at stage t to a patient with features $h_{t,1}$; rather the choice of which of the two stage t treatments under study to administer should be left to the discretion of the treating physician. We show that this strategy can lead to the implementation, for a sizable subset of the study population, of a treatment regime that is apriori known, i.e. even before the data are collected, to be nonoptimal.

In our final section, we discuss the issue of the sensitivity of indirect methods and the insensitivity of most direct methods to model misspecification. Further we discuss the question of whether direct methods that are insensitive to model misspecification also suffer from nonregularity. We show that the evidence presented for the non regularity of a direct method by the authors in their appendix A does not bear on this question, because the direct method they discuss is not insensitive to model misspecification. None the less, the authors' Theorem 3.3 and their subsequent toy example shows that estimation of the optimal value function is an irregular problem, even when estimated with direct methods insensitive to model misspecification.

Target of inference The authors' target of inference is the stage one non-regular parameter β_{11}^{*} that determines the optimal treatment strategy π_1^{dp} at stage one if the optimal strategy π_2^{dp} were followed at stage two. In this section we argue that the parameter of clinical interest may not be β_{11}^{*} but rather the (random) parameter determining the optimal stage one regime given the regime that was actually followed at stage two. Interestingly, we shall see that the authors' adaptive confidence interval also provides valid inference for this latter parameter.

To avoid unrelated modeling issues, we will assume for now that X_1, X_2, A_1, A_2 are each Bernoulli, Y is continuous and the models for the Q-functions are saturated, and hence correctly specified. To this end we define $H_1^T \equiv$

$(H_{1,0}^T, H_{1,1}^T), H_{1,0}^T \equiv H_{1,1}^T \equiv (1, X_1)$ and $H_2^T \equiv (H_{2,0}^T, H_{2,1}^T)$ with $H_{2,0}^T \equiv H_{2,1}^T \equiv (1_{X_1=i, X_2=j, A_1=k})_{i,j,k \in \{0,1\}}$. We then consider the model $h_{2,0}^T \beta_{2,0} + a_2 h_{2,1}^T \beta_{2,1}$ for $Q_2(h_2, a_2) \equiv E[Y|H_2 = h_2, A_2 = a_2]$ with β_2^* the true value of β_2. Recall that among all stage two regimes π_2 the optimal one is given by the dynamic programing solution $\pi_2^{dp}(h_2) = \arg\max_{a_2} Q_2(h_2, a_2) = 1_{Q_2(h_2,1) - Q_2(h_2,0) > 0} = 1_{h_{2,1}^T \beta_{2,1}^* > 0}$.

Given that a particular, possibly non-optimal, regime π_2 is to be followed at stage two, the optimal stage one regime $\pi_1^{\pi_2}(h_1)$ is given by $\arg\max_{a_1} Q_1^{\pi_2}(h_1, a_1)$ where $Q_1^{\pi_2}(h_1, a_1) = E[Q_2(H_2, \pi_2(H_2))|H_1 = h_1, A_1 = a_1]$. Note that $Q_1^{\pi_2^{dp}}(h_1, a_1)$ is therefore the authors' stage one value function $Q_1(h_1, a_1)$ and $\pi_1^{\pi_2^{dp}}(h_1)$ is the authors' dynamic programming solution $\pi_1^{d_p}(h_1)$. Assuming, as we do, saturated models $h_{1,0}^T \beta_{1,0}^{\pi_2} + a_1 h_{1,1}^T \beta_{1,1}^{\pi_2}$ for $Q_1^{\pi_2}(h_1, a_1)$ the optimal stage one regime when π_2 will be followed at stage two is $\pi_1^{\pi_2}(h_1) = 1_{h_{1,1}^T \beta_{1,1}^{\pi_2,*} > 0}$ where $\beta_{1,1}^{\pi_2,*}$ solves the population moment equation

$$P\left[H_{1,1}\left(A_1 - \frac{1}{2}\right)\left\{Y + \left(\pi_2\left(H_2\right) - A_2\right) H_{2,1}^T \beta_{2,1}^* - A_1 H_{1,1}^T \beta_{1,1}^{\pi_2}\right\}\right] = 0 \quad (0.1)$$

Note that the stage one parameters must be labelled by the stage two regime π_2 as the function $Q_1^{\pi_2}(h_1, a_1)$, and thus $\beta_1^{\pi_2,*}$, varies with π_2. The authors' parameter $\beta_{1,1}^*$ is equal to $\beta_{1,1}^{\pi_2^{dp},*}$.

Consider the following parable presented by one of us at a conference in Bristol England in 2012 attended by Susan Murphy and ourselves. You are told the regime π_2 that will be used to assign treatment at stage two. Your job is to decide, from the data on the n subjects available to you, the best stage one treatment for a patient with features h_1. To do so you will estimate $\beta_{1,1}^{\pi_2,*}$ with a semiparametric efficient estimator $\widehat{\beta}_{1,1}^{\pi_2}$ solving

$$\mathbb{P}_n\left[H_{1,1}\left\{A_1 - \mathbb{P}_n\left(A_1|X_1\right)\right\}\left[Y + \left\{\pi_2\left(H_2\right) - A_2\right\} H_{2,1}^T \beta_{2,1}^* - A_1 H_{1,1}^T \beta_{1,1}^{\pi_2}\right]\right] = 0$$

and then use this estimator to construct a point and interval estimator for $h_{1,1}^T \beta_{1,1}^{\pi_2,*}$ [here $\mathbb{P}_n(A_1|x_1)$ is the empirical conditional mean of A_1 given $X_1 = x_1$]. You will then provide the treatment $1_{h_{1,1}^T \widehat{\beta}_{1,1}^{\pi_2} > 0}$ if the 90% CI for $h_{1,1}^T \beta_{1,1}^{\pi_2,*}$ does not contain zero; otherwise, following the authors, you leave the treatment decision up to the patient's physician.

Suppose, one day you suddenly learn that, unknown to you, the original data set had been of size $2n$ but had been randomly divided in two and only the first half-sample had been made available to you. You are told that the second half-sample had been used to estimate the stage two regime π_2 that you were told was to be used in the future. Specifically, π_2 was the estimate $\widetilde{\pi}_2^{dp}(h_2) = 1_{h_{2,1}^T \widetilde{\beta}_{2,1} > 0}$ of the optimal second stage regime $\pi_2^{dp}(h_2)$ based on the second half of the data. Here we use the symbol $\widetilde{}$ to denote an estimator calculated from the second half of the data. If in the second half sample, $\widetilde{\pi}_2^{dp}(h_2)$ differed from $\pi_2^{dp}(h_2)$ due to sampling variability, then your targets of inference would be $\beta_{1,1}^{\pi_2,*}|_{\pi_2 = \widetilde{\pi}_2^{dp}}$

and $\pi_1^{\widetilde{\pi}_2^{dp}}(h_1)$ rather than $\beta_{1,1}^{\pi_2^{dp},*}$and $\pi_1^{dp}(h_1)$ since it is $\widetilde{\pi}_2^{dp}$, rather than $\pi_2^{dp}(h_2)$, which would actually be followed in the second stage. As far as inference is concerned, conditional on the second half of the sample, inference based on $\widehat{\beta}_{1,1}^{\widetilde{\pi}_2^{dp}}$ is regular as $\widetilde{\pi}_2^{dp}(h_2) = 1_{h_{2,1}^T \widetilde{\beta}_{2,1} > 0}$ is fixed, so no random parameter value occurs inside an indicator. In particular, $n^{1/2}\{\widehat{\beta}_{1,1}^{\widetilde{\pi}_2^{dp}} - \beta_{1,1}^{\pi_2,*}|_{\pi_2 = \widetilde{\pi}_2^{dp}}\}$ will be uniformly asymptotically normal with mean zero conditional on the second sample data. Furthermore standard $1-\alpha$ Wald type confidence intervals for the (conditionally fixed) parameter $\beta_{1,1}^{\pi_2,*}|_{\pi_2 = \widetilde{\pi}_2^{dp}}$ centered at $\widehat{\beta}_{1,1}^{\widetilde{\pi}_2^{dp}}$ will have asymptotically correct conditional coverage uniformly over the parameter space. This implies that this Wald interval will provide correct unconditional coverage of the random parameter $\beta_{1,1}^{\pi_2,*}|_{\pi_2 = \widetilde{\pi}_2^{dp}}$. We emphasize that unconditionally, it is the random variable $\beta_{1,1}^{\pi_2,*}|_{\pi_2 = \widetilde{\pi}_2^{dp}}$, rather than the fixed irregular parameter $\beta_{1,1}^{\pi_2^{dp},*} = \beta_{1,1}^*$ for which these intervals would provide uniformly correct coverage.

Consider the implication of this parable for the more typical setting in which both the first and second stage regime are estimated from the same data. Obviously, our philosophical conclusion that we should optimize our stage one regime in light of the regime that will actually be followed at stage two cannot depend on whether or not we split the sample. Thus, our target of inference should remain the random regime $\pi_1^{\pi_2}(h_1)|_{\pi_2 = \widehat{\pi}_2^{dp}}$ just as in the split sample case, except now $\widehat{\pi}_2^{dp}$ is the stage two regime estimated from the entire sample.

Conditional inference given $\widehat{\pi}_2^{dp}$ and, in particular, the construction of confidence intervals that cover $h_{1,1}^T \beta_{1,1}^{\pi_2,*}|_{\pi_2 = \widehat{\pi}_2^{dp}}$ with probability $(1-\alpha)$ in the subset of hypothetical repetitions with the same $\widehat{\pi}_2^{dp}$ does not appear to be doable. This difficulty arises because of the dependence between $\widehat{\pi}_2^{dp}$ and $\widehat{\beta}_{1,1}^{\widehat{\pi}_2^{dp}}$; in fact, the difference $\widehat{\beta}_{1,1}^{\widehat{\pi}_2^{dp}} - \beta_{1,1}^{\pi_2,*}|_{\pi_2 = \widehat{\pi}_2^{dp}}$ does not even converge to 0 in probability conditional on $\widehat{\pi}_2^{dp}$.

Although less desirable, unconditional inference about the random parameter $\beta_{1,1}^{\pi_2,*}|_{\pi_2 = \widehat{\pi}_2^{dp}}$ is feasible. Interestingly, we will now show that the adaptive confidence intervals proposed by the authors cover the random parameter $\pi_1^{\pi_2}(h_1)|_{\pi_2 = \widehat{\pi}_2^{dp}}$ with probability at least $(1-\alpha)$. Specifically, it follows from Eq. (0.1) that

$$\begin{aligned}
&\sqrt{n}\left\{\beta_{1,1}^{\pi_2,*}|_{\pi_2 = \widehat{\pi}_2^{dp}} - \beta_{1,1}^{\pi_2^{dp},*}\right\} \\
&= \Sigma_{1,1\infty}^{-1}\sqrt{n} P\left[H_{1,1}\left(A_1 - \frac{1}{2}\right)\left(1_{H_{2,1}^T \beta_{2,1} > 0} - 1_{H_{2,1}^T \beta_{2,1}^* > 0}\right) H_{2,1}^T \beta_{2,1}^*\right]\Bigg|_{\beta_{2,1} = \widehat{\beta}_{2,1}}
\end{aligned}$$

where $\Sigma_{1,1\infty} = P[H_{1,1}H_{1,1}^T(A_1 - \frac{1}{2})^2]$. On the other hand,

$$\sqrt{n}\left\{\widehat{\beta}_{1,1}^{\widehat{\pi}_2} - \beta_{1,1}^{\pi_2^{dp},*}\right\} = \mathbb{S}_{1,n} + \widehat{\Sigma}_{1,1}^{-1}\mathbb{P}_n\left[H_{1,1}\left\{A_1 - \mathbb{P}_n\left(A_1|X_1\right)\right\}\mathbb{U}_n\right]$$

where $\widehat{\Sigma}_{1,1} = \mathbb{P}_n[H_{1,1}H_{1,1}^T\{A_1 - \mathbb{P}_n(A_1|X_1)\}^2]$,

$$\mathbb{S}_{1,n} = \widehat{\Sigma}_{1,1}^{-1}\sqrt{n}\mathbb{P}_n\Big[H_{1,1}\left\{A_1 - \mathbb{P}_n\left(A_1|X_1\right)\right\} \\ \times\left\{Y + \left(1_{H_{2,1}^T\beta_{2,1}^*>0} - A_2\right)H_{2,1}^T\beta_{2,1}^* - A_1H_{1,1}^T\beta_{1,1}^{\pi_2^{dp},*}\right\}\Big]$$

and $\mathbb{U}_n = \sqrt{n}([H_{2,1}^T\widehat{\beta}_{2,1}]_+ - [H_{2,1}^T\beta_{2,1}^*]_+)$. Note that because the model for $Q_1^{\pi_2^{dp}}(h_1, a_1)$ is saturated, the estimator $\widehat{\beta_{1,1}^{\widehat{\pi}_2}}$ coincides with the Q-learning estimator $\widehat{\beta}_{1,1}$ of LLQPM, and $\mathbb{S}_{1,n}$ is the second component of the 2×1 vector $\mathbb{S}_n$ defined in section 2.1 of the article. We therefore have that

$$\sqrt{n}\left\{\widehat{\beta}_{1,1} - \beta_{1,1}^{\pi_2,*}|_{\pi_2=\widehat{\pi}_2^{dp}}\right\} \\ = \mathbb{S}_{1,n} + \widehat{\Sigma}_{1,1}^{-1}\mathbb{P}_n\left[H_{1,1}\left\{A_1 - \mathbb{P}_n\left(A_1|X_1\right)\right\}\mathbb{U}_n\right] \\ - \Sigma_{1,1\infty}^{-1}\sqrt{n}P\left[H_{1,1}\left(A_1 - \frac{1}{2}\right)\left(1_{H_{2,1}^T\beta_{2,1}>0} - 1_{H_{2,1}^T\beta_{2,1}^*>0}\right)H_{2,1}^T\beta_{2,1}^*\right]\Big|_{\beta_{2,1}=\widehat{\beta}_{2,1}}$$

The following Lemma (proved in the Appendix) extends Theorem 4.1 of LLQPM for the special case considered here: i.e. X_1 and X_2 binary and the models for the Q-functions saturated. We conjecture however that, under the conditions of Theorem 4.1., the Lemma is valid for arbitrary covariates X_1 and X_2 and linear, not necessarily saturated, models for the Q-functions.

Lemma. *Suppose X_1, X_2, A_1 and A_2 are binary and $P(A_1|X_1) = 1/2$. Suppose for each n, the underlying generative distribution P_n satisfies $(A3)$ of LLQPM. For any stage two regime, let $\beta_{1,1,n}^{\pi_2,*}$ denote the solution to (0.1) with $\beta_{2,1}^*$ replaced with $\beta_{2,1,n}^* = \beta_{2,1}^* + s/\sqrt{n}$. Let $(\mathbb{S}_\infty^T, \mathbb{V}_\infty^T)^T$ be the random vector defined in Theorem 4.1 of LLQPM. Write $\mathbb{S}_\infty^T = (\mathbb{S}_{0,\infty}^T, \mathbb{S}_{1,\infty}^T)$ where $\mathbb{S}_{1,\infty}$ is of dimension 2 and for any given fixed $c \in \mathbb{R}^4$, let $\mathcal{U}(c)$ and $\mathcal{L}(c)$ be as defined in section 4.2 of LLQPM,*

1) the limiting distribution of $\sqrt{n}\{\widehat{\beta}_{1,1} - \beta_{1,1,n}^{\pi_2,*}|_{\pi_2=\widehat{\pi}_2^{dp}}\}$ is equal to

$$\mathbb{S}_{1,\infty} + \Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1 - \frac{1}{2}\right)H_{2,1}^T\mathbb{V}_\infty 1_{H_{2,1}^T\beta_{2,1}^*>0}\right] \\ + \Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1 - \frac{1}{2}\right)1_{H_{2,1}^T(\mathbb{V}_\infty+s)>0}H_{2,1}^T\mathbb{V}_\infty 1_{H_{2,1}^T\beta_{2,1}^*=0}\right]$$

2) for any fixed $c_1 \in \mathbb{R}^2$

$$sup_{\gamma\in\mathbb{R}^8}c_1^T\Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1 - \frac{1}{2}\right)1_{H_{2,1}^T(\mathbb{V}_\infty+\gamma)>0}H_{2,1}^T\mathbb{V}_\infty 1_{H_{2,1}^T\beta_{2,1}^*=0}\right] \\ = sup_{\gamma\in\mathbb{R}^8}c_1^T\Sigma_{1,1\infty}^{-1}P\Big[H_{1,1}\left(A_1 - \frac{1}{2}\right)\Big\{\left[H_{2,1}^T\left(\mathbb{V}_\infty + \gamma\right)\right]_+ \\ - \left[H_{2,1}^T\gamma\right]_+\Big\}1_{H_{2,1}^T\beta_{2,1}^*=0}\Big]$$

and

$$\begin{aligned}
&inf_{\gamma\in\mathbb{R}^s} c_1^T\Sigma_{1,1\infty}^{-1} P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)1_{H_{2,1}^T(\mathbb{V}_\infty+\gamma)>0}H_{2,1}^T\mathbb{V}_\infty 1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\\
&= inf_{\gamma\in\mathbb{R}^s} c_1^T\Sigma_{1,1\infty}^{-1} P\Bigg[H_{1,1}\left(A_1-\frac{1}{2}\right)\Big\{\left[H_{2,1}^T(\mathbb{V}_\infty+\gamma)\right]_+\\
&\quad-\left[H_{2,1}^T\gamma\right]_+\Big\}1_{H_{2,1}^T\beta_{2,1}^*=0}\Bigg]
\end{aligned}$$

3) Let $c_1^*=(0,0,c_1^T)$*. The limiting distribution of* $\sqrt{n}c_1^T\{\widehat{\beta}_{1,1}-\beta_{1,1,n}^{\pi_2,*}|_{\pi_2=\widehat{\pi}_2^{dp}}\}-\mathcal{U}(c_1^*)$ *is*

$$\begin{aligned}
&c_1^T\Sigma_{1,1\infty}^{-1} P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)1_{H_{2,1}^T(\mathbb{V}_\infty+s)>0}H_{2,1}^T\mathbb{V}_\infty 1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\\
&\quad- sup_{\gamma\in\mathbb{R}^s} c_1^T\Sigma_{1,1\infty}^{-1} P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)1_{H_{2,1}^T(\mathbb{V}_\infty+\gamma)>0}H_{2,1}^T\mathbb{V}_\infty 1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\\
&\le 0
\end{aligned}$$

and the limiting distribution of $\sqrt{n}c_1^T\{\widehat{\beta}_{1,1}-\beta_{1,1,n}^{\pi_2,*}|_{\pi_2=\widehat{\pi}_2^{dp}}\}-\mathcal{L}(c_1^*)$ *is*

$$\begin{aligned}
&c_1^T\Sigma_{1,1\infty}^{-1} P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)1_{H_{2,1}^T(\mathbb{V}_\infty+s)>0}H_{2,1}^T\mathbb{V}_\infty 1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\\
&\quad- inf_{\gamma\in\mathbb{R}^s} c_1^T\Sigma_{1,1\infty}^{-1} P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)1_{H_{2,1}^T(\mathbb{V}_\infty+\gamma)>0}H_{2,1}^T\mathbb{V}_\infty 1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\\
&\ge 0
\end{aligned}$$

The Lemma establishes that under the local generative process P_n the limiting law of $\sqrt{n}\{\widehat{\beta}_{1,1}-\beta_{1,1,n}^{\pi_2,*}|_{\pi_2=\widehat{\pi}_2^{dp}}\}$ is different from that of $\sqrt{n}\{\widehat{\beta}_{1,1}-\beta_{1,1,n}^{*}\}$. None the less the sharp bounds for the limit random variables are the same.

By part (3) of the Lemma,

$$P_n\left(c_1^T\widehat{\beta}_{1,1}-\mathcal{U}\left(c_1^*\right)/\sqrt{n}\le c_1^T\beta_{1,1,n}^{\pi_2,*}|_{\pi_2=\widehat{\pi}_2^{dp}}\le c_1^T\widehat{\beta}_{1,1}-\mathcal{L}\left(c_1^*\right)/\sqrt{n}\right)=1-o_p\left(1\right)$$

Arguing as in section 4.2 of LLQPM, we thus conclude that if $\widehat{u}_1$ is the $(1-\alpha/2)\times 100$ percentile of the bootstrap distribution of $\mathcal{U}(c_1^*)$ and $\widehat{l}_1$ is the $\alpha/2\times 100$ percentile of the bootstrap distribution of $\mathcal{L}(c_1^*)$, then under assumptions (A1), (A2) and (A4) of LLQPM,

$$P_M\left(c_1^T\widehat{\beta}_{1,1}-\widehat{u}_1/\sqrt{n}\le c_1^T\beta_{1,1,n}^{\pi_2,*}|_{\pi_2=\widehat{\pi}_2^{dp}}\le c_1^T\widehat{\beta}_{1,1}-\widehat{l}_1/\sqrt{n}\right)\ge 1-\alpha+o_P\left(1\right)$$

where the inequality is an equality if $P(H_{2,1}^T\beta_{2,1}^*=0)=0$.

We therefore see that the confidence interval $(c_1^T\widehat{\beta}_{1,1}-\widehat{u}_1/\sqrt{n}, c_1^T\widehat{\beta}_{1,1}-\widehat{l}_1/\sqrt{n})$ serves just as much as a $(1-\alpha)\times 100\%$ confidence interval for the fixed parameter $\beta_{1,1,n}^{\pi_2^{dp}}$ as for the random parameter $\beta_{1,1,n}^{\pi_2,*}|_{\pi_2=\widehat{\pi}_2^{dp}}$.

Incidentally, aside from the issue of whether the target parameter should be $\beta_{1,1}^{\pi_2^{dp}}$ or $\beta_{1,1,n}^{\pi_2,*}|_{\pi_2=\widehat{\pi}_2^{dp}}$, the availability of the formula for the limit laws of $\mathcal{U}(c_1^*)$ and $\mathcal{L}(c_1^*)$ offers the opportunity to use a semi-parametric, rather than the non-parametric, bootstrap to approximate the quantiles of $\mathcal{U}(c_1^*)$ and $\mathcal{L}(c_1^*)$. Specifically, one might consider replacing $\widehat{u}_1$ with $\widetilde{u}_1$, the $(1-\alpha/2)\times 100$ percentile of the distribution of $u_n^{(sb)}(c_1^*;\mathbb{S}_{1,\infty},\mathbb{V}_\infty)$ and replacing $\widehat{l}_1$ with $\widetilde{l}_1$, the $\alpha/2\times 100$ percentile of the distribution of $l_n^{(sb)}(c_1^*;\mathbb{S}_{1,\infty},\mathbb{V}_\infty)$ where

$$\begin{aligned}&u_n^{(sb)}\left(c_1^*;s,v\right)\\ &\equiv s+c_1^T\widehat{\Sigma}_{1,1}^{-1}\mathbb{P}_n\left[H_{1,1}\left(A_1-\frac{1}{2}\right)H_{2,1}^T v 1_{\widehat{T}(H_{2,1})>\lambda_n}\right]\\ &\quad+\sup_{\gamma\in\mathbb{R}^8} c_1^T\widehat{\Sigma}_{1,1}^{-1}\mathbb{P}_n\left[H_{1,1}\left(A_1-\frac{1}{2}\right)\left\{\left[H_{2,1}^T\left(v+\gamma\right)\right]_+-\left[H_{2,1}^T\gamma\right]_+\right\}1_{\widehat{T}(H_{2,1})<\lambda_n}\right]\end{aligned}$$

and

$$\begin{aligned}&l_n^{(sb)}\left(c_1^*;s,v\right)\\ &\equiv s+c_1^T\widehat{\Sigma}_{1,1}^{-1}\mathbb{P}_n\left[H_{1,1}\left(A_1-\frac{1}{2}\right)H_{2,1}^T v 1_{\widehat{T}(H_{2,1})>\lambda_n}\right]\\ &\quad+\inf_{\gamma\in\mathbb{R}^8} c_1^T\widehat{\Sigma}_{1,1}^{-1}\mathbb{P}_n\left[H_{1,1}\left(A_1-\frac{1}{2}\right)\left\{\left[H_{2,1}^T\left(v+\gamma\right)\right]_+-\left[H_{2,1}^T\gamma\right]_+\right\}1_{\widehat{T}(H_{2,1})<\lambda_n}\right]\end{aligned}$$

are regarded as fixed, i.e. non-random functions of s and v. We conjecture that under assumptions (A1), (A2) and (A4) of LLQPM, the interval $(c_1^T\widehat{\beta}_{1,1}-\widetilde{u}_1/\sqrt{n}, c_1^T\widehat{\beta}_{1,1}-\widetilde{l}_1/\sqrt{n})$ satisfies

$$P\left(c_1^T\widehat{\beta}_{1,1}-\widetilde{u}_1/\sqrt{n}\leq c_1^T\beta_{1,1}^{\pi_2,*}|_{\pi_2=\widehat{\pi}_2^{dp}}\leq c_1^T\widehat{\beta}_{1,1}-\widetilde{l}_1/\sqrt{n}\right)\geq 1-\alpha+o_P\left(1\right)$$

and

$$P\left(c_1^T\widehat{\beta}_{1,1}-\widetilde{u}_1/\sqrt{n}\leq c_1^T\beta_{1,1}^*\leq c_1^T\widehat{\beta}_{1,1}-\widetilde{l}_1/\sqrt{n}\right)\geq 1-\alpha+o_P\left(1\right)$$

where the inequalities in the two preceding displays are equalities if $P(H_{2,1}^T\beta_{2,1}^*=0)=0$. One advantage of the semiparametric bootstrap approach is that the Monte-Carlo approximation to $\widetilde{l}_1$ and $\widetilde{u}_1$, unlike the MC approximation to $\widehat{l}_1$ and $\widehat{u}_1$, does not require recalculating the estimators of $\beta_{2,1}$ and $\beta_{1,1}$ in each Monte-Carlo draw.

Incoherent decision making based on examining univariate confidence intervals In the discussion of the ADHD study, the authors display in table 9 univariate confidence intervals for $h_{t,1}^T\beta_{t,1}^*$ for each possible value $h_{t,1}$ of $H_{t,1}, t=1,2$. The authors argue that if a confidence interval (say a 90% interval) for the effect parameter $h_{t,1}^T\beta_{t,1}^*$ for a patient with features $h_{t,1}$ at stage t includes zero, then no treatment recommendation should be made by the trialists; rather it is preferable that the choice of which treatment to administer should be left to the discretion of the treating physician. We will now show, by example, that this use of univariate confidence intervals in decision making can result in implementation on a sizeable subset of the patient population of a treatment regime that is apriori known, i.e. even before the data are collected, to be nonoptimal. We will also argue that this phenomena cannot occur if joint, rather than univariate confidence intervals, are used in an analogous fashion. The issue we raise here is of concern even in one stage studies, so to avoid distracting complications we will discuss it in this simple setting.

Consider a one stage randomized study with dichotomous treatment $A \in \{-1, 1\}$ and a single trivariate covariate $X \in \{-1, 0, 1\}$. Suppose that $P(X = x) = 1/3$ for $x = -1, 0, 1$, that $P(A = 1|X) = 1/2$, and that it is known, based on prior knowledge, that Y follows the linear homoscedastic model

$$E\left(Y|A,X\right) = \psi_0 + \psi_1 X + A\left(\psi_3 + \psi_4 X\right) \tag{0.2}$$

with $\text{var}(Y|A, X) = \sigma^2$. Note the model implies linear treatment effect modification by X, that is, the treatment effect function $\Delta(x) = E(Y|A = 1, X = x) - E(Y|A = 1, X = x)$ is linear in x.

With a single trivariate covariate and a dichotomous treatment, there exist only eight possible treatment regimes. For ease of reference, we write them as tuples $(a_{-1}, a_0, a_1), a_j \in \{-1, 1\}$ with the first coordinate indicating treatment assignment to covariate $x = -1$, the second to covariate $x = 0$, and the third to covariate $x = 1$. For instance, the tuple $(1, -1, 1)$ corresponds to the regime that assigns $A = 1$ to covariates $x = -1$ and 1 and treatment $A = -1$ to covariate $x = 0$. For a given value of ψ, the optimal regime is $\pi_\psi(X) = 2 \times 1_{\psi_3+\psi_4 X>0} - 1$. Figure 1 plots with the same color the regions where the many to one map $\psi \to \pi_\psi$ takes the same value. For instance, on the region $\{(\psi_3, \psi_4) : \psi_3 > 0, \psi_4 > 0, \psi_4 > \psi_3\}$, the optimal regime assigns $A = -1$ to $x = -1$ and $A = 1$ to $x = 0$ and $x = 1$. Likewise, on the half-line $\{(\psi_3, \psi_4) : \psi_3 = 0, \psi_4 > 0\}$, the optimal regime assigns $A = -1$ to $x = -1$, $A = 1$ to $x = 1$ and is indifferent about treatment assignment to $x = 0$ (indifferent treatment assignment for a covariate x is denoted with U in the corresponding entry of the tuple). Note that the regime $(1, -1, 1)$ does not appear in any region, as it is incompatible with the assumption of linear treatment effect modification. This implies that a data analyst that correctly postulates the linear model (0.2) a-priori excludes regime $(1, -1, 1)$ from the set of possible optimal regimes.

Now, let $\widehat{\psi}$ be the ordinary least squares estimator of ψ and let ψ^* be the true value of ψ. Under the assumed data generating process, $\sqrt{n}(\widehat{\psi} - \psi^*) \to N(0, \Sigma)$ where Σ is the diagonal matrix with diagonal $\sigma^2 \times (1, 3/2, 1, 3/2)$. Suppose that $\widehat{\psi}_{tr} \equiv (\widehat{\psi}_3, \widehat{\psi}_4) = (0.4, 1)$ and that $\widehat{\sigma}^2/n = 0.1$ where $\widehat{\sigma}^2$ estimates consistently σ^2. Let $\widehat{\Sigma}_{tr}$ be the 2×2 diagonal matrix with diagonal $(\widehat{\sigma}^2/n) \times (1, 3/2)$. The ellipse

$$\begin{aligned} C\left(.9\right) &= \left\{\psi_{tr} = \left(\psi_3, \psi_4\right); \left(\psi_{tr} - \widehat{\psi}_{tr}\right)^T \widehat{\Sigma}_{tr}^{-1} \left(\psi_{tr} - \widehat{\psi}_{tr}\right) < 4.61\right\} \\ &= \left\{\psi_{tr} = \left(\psi_3, \psi_4\right); 10 \times \left(\psi_3 - 0.4\right)^2 + 20 \times \left(\psi_4 - 1\right)^2/3 < 4.61\right\} \end{aligned}$$

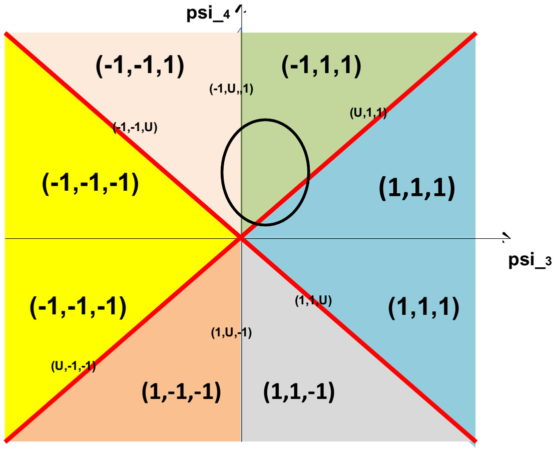

Fig 1. *Ellipsoid is a 90% Confidence region for (psi_3, psi_4). Values of (psi_3, psi_4) in regions with the same colour correspomd to the same optimal treatment regime, indicated as a tuple over the region.*

Table 1

	90% confidence interval for $\psi_3^* + \psi_4^* x$
$x = -1$	-0.6 ± 0.82
$x = 0$	$0.4 \pm 0.518\,61$
$x = 1$	1.4 ± 0.82

shown in the Figure 1 is a 90% joint confidence region for $\psi_{tr}^* \equiv (\psi_3^*, \psi_4^*)$. Ninety percent univariate Wald confidence intervals $(\widehat{\psi}_3 + \widehat{\psi}_4 x) \pm 1.64\,\sigma\sqrt{(1 + x^2 3/2)/n}$ for each $\psi_3^* + \psi_4^* x$ are given in table 1.

For $x = 0$ and $x = -1$, the confidence intervals for $\psi_3 + \psi_4 x$ do not exclude 0, so the authors would recommend to leave it to the doctor to decide whether or not to treat patients with $x = 0$ or -1. In contrast, for $x = 1$, the recommendation would be to treat with $A = 1$. Now, suppose that such a recommendation is implemented in a given population. Suppose that if given no indication on how to treat patients with covariates $x = 0$ or $x = -1$, one quarter of the doctors in the target population would choose to treat both kind of patients with $A = 1$, a second quarter would choose to treat both kind of patients with $A = -1$, a third quarter would treat $x = 0$ patients with $A = 1$ and $x = -1$ patients with $A = -1$, and a fourth quarter would treat $x = 0$ patients with $A = -1$ and $x = -1$ patients with $A = 1$ .Then, effectively, each of the four regimes $(1, 1, 1), (-1, 1, 1), (1, -1, 1)$ and $(-1, -1, 1)$ will have been implemented in one quarter of the population. However, this would lead to implementing in one quarter of the population the regime $(1, -1, 1)$ which a priori is known not to be optimal under any state of nature.

Now, given the 90% joint confidence region $C(.9)$ for ψ_{tr}, the set $\Pi_{op}(0.9) = \{\pi^{\psi_{tr}}; \psi_{tr} \in C(0.9)\}$ with $\pi^{\psi_{tr}}(x) = 1_{\psi_3 + \psi_4 x > 0}$ contains the optimal strategy $\pi^{dp} = \pi^{\psi_{tr}^*}$ with probability at least 0.9 in large samples and cannot contain a regime that was known apriori to be non optimal. Thus, rather than leaving the physicians the option of deciding what to administer to *specific patients*, we could offer them the opportunity to choose from the *specific regimes* $(-1,-1,1),(-1,1,1),(1,1,1)$ that comprise the set $\Pi_{op}(0.9)$. With this strategy we guarantee that no physician will end up implementing a regime known a-priori to be non-optimal. Of course, this strategy may confuse a physician who, unversed in the magic of linear models, may have a hard time grasping that once the specific treatment option $A=1$ is chosen (and consequently, regarded as optimal) for a patient with $x=-1$, the linear model implies that only the option $A=1$ can be optimal for a patient with $x=0$. Univariate confidence intervals do not encode the a priori restrictions imposed by models on the Q-functions. In contrast, joint confidence regions for ψ_{tr} do. We note that, as pointed out by Robins (2004, pp. 222–224), even in the multistage setting, it is possible to construct valid uniform large sample confidence regions for the vector of all treatment effect parameters by inverting tests, even if some of the parameters are non-regular in the sense that they do not have a pathwise derivative at certain generative laws.

One might wonder what regime the conventional decision theoretic approaches would choose in our simple one-stage decision problem. In large samples the Bayesian strategy that maximizes the posterior expected value would choose the regime $(0,1,1)$ associated with the MLE $\widehat{\psi}_{tr}=(\widehat{\psi}_3,\widehat{\psi}_4)=(0.4,1)$, since the posterior distribution for ψ_{tr} in large samples is normal with mean $\widehat{\psi}_{tr}$. The minimax regret regime that minimizes the maximum regret value and the maximin regime that maximizes the minimum value over all possible $\psi_{tr}\in C(0.9)$ for any alpha is also $(0,1,1)$ because the set $C(0.9)$ is an ellipse.

Sensitivity to model misspecification Q-learning and A-learning (also known as g-estimation of optimal regime structural nested mean models) methods model the dependence of the quality function on information accrued. In general this dependence is modelled parametrically in Q-learning and semiparametrically in A learning, although nonparametric fitting is also possible (Qian and Murphy, 2011; van der Laan, 2013).

A second class of methods search for the optimal regime in a specific class which often, but not always (see Zhao et al., 2012), is parametric (Zhang et al., 2012). These methods first estimate the value function for each regime in the class, and then estimate the optimal regime as the one that maximizes the estimated value functions. When the dimension of the class is large, some authors have proposed further modeling the dependence of the value functions on the regimes in the class (van der Laan and Petersen, 2007; Robins et al., 2008; Orellana et al., 2010). LLQPM refer to Q and A learning as indirect methods and to methods that rely on estimates of the value function as direct methods. Throughout the following discussion we assume that data are from a randomized experiment, so that the stage-specific probabilities of treatment are known to the analyst.

In the introduction, the authors mention the well known fact that indirect methods based on parametric or semiparametric models are not robust

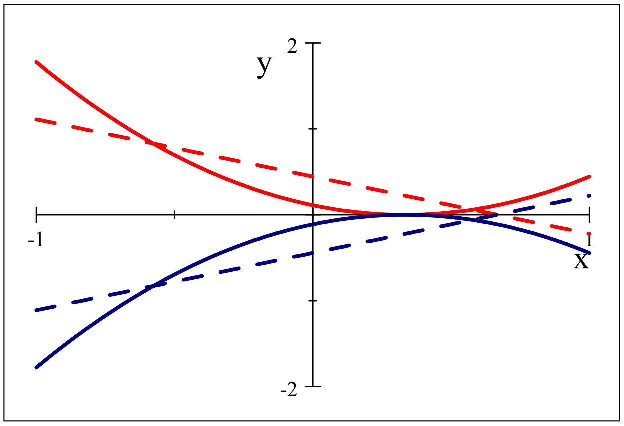

Fig 2.

to misspecification of the models for the quality function. This point is vividly portrayed in the following example borrowed from Qian and Murphy (2011). Consider a one-stage randomized study with binary treatment $A \in \{-1, 1\}$ and baseline covariate X uniformly distributed on $(-1, 1)$ and suppose that $e_j(x) \equiv E(Y|A = j, X = x)$, for $j = -1$ or 1 satisfies $e_j(x) = j \times (x - 1/3)^2$. In Figure 2, the parabolas are the curves $e_j(x)$, for $j = -1$ and 1. If larger outcomes are preferable, then for no x is treatment $A = -1$ preferable to $A = 1$ since for all values of x, $e_1(x) \geq e_{-1}(x)$.

The best linear approximation to $e_j(x)$ (in the sense of minimizing mean squared error) is the line $l_j(x) \equiv j4/9 - j2/3x$, i.e. for $j = 1$ and -1, $(j4/9, -j2/3) = \arg\min_{(\beta_{0,j},\beta_{1,j})} \int_{-1}^{1} (e_j(x) - \beta_{0,j} - \beta_{1,j}x)^2 \; dx$. The dashed lines in Figure 2 are the lines $l_j(x), j = -1, 1$. Consider Q-learning under the incorrect assumption that $E(Y|A = j, X) = \beta_{0,j} + \beta_{1,j}X$. Suppose the method proceeds by computing $\widehat{l}_j(x) = \widehat{\beta}_{0,j} + \widehat{\beta}_{1,j}x$, the ordinary least squares fit in the regression of Y on X, separately in each treatment arm. As the sample size increases, $\widehat{l}_j(x)$ converges to $l_j(x)$ and thus with probability going to 1, Q-learning leads to incorrectly conclude that treatment $A = -1$ is preferable for any x greater than 2/3.

Irregularity of direct method estimators Recognizing the sensitivity to model misspecification of indirect methods, a number of authors (Zhao et al., 2012; Zhang et al., 2012) have recommended that direct methods be used. In the Appendix, LLQPM argue that even direct methods suffer from the problem of nonregularity. While, as discussed further below, we are in agreement with this remark we wish to point out that the argument given by the authors to support this claim is incorrect because it misrepresents the estimators that the direct methods use. In particular, the authors' proof in Appendix A considers estimators of optimal regimes whose consistency is only ensured when the models for the Q-functions are correctly specified, thus defeating the purpose of direct methods. For this reason we question whether the method they analyze should be referred to as a 'direct method'. Specifically, as in Appendix A of LLQPM, let $A_t \in \{0,1\}, t = 1,2$, with $P(A_t|H_t) = 1/2$. Let $\pi_\psi = (\pi_{1,\psi_{1,1}}, \pi_{2,\psi_{2,1}})$ denote the regime $\pi_{1,\psi_{1,1}}(H_{1,1}) = 1_{H_{1,1}^T\psi_{1,1}>0}, \pi_{2,\psi_{2,1}}(H_{2,1}) = 1_{H_{2,1}^T\psi_{2,1}>0}$ and let $V_\psi \equiv \mathbb{E}^{\pi_\psi}(Y) = 4P[Y1_{(2A_1-1)H_{1,1}^T\psi_{1,1}>0}1_{(2A_2-1)H_{2,1}^T\psi_{2,1}>0}]$ be its value function when the randomization probabilities at the first and second stage are 1/2. Direct methods are designed to provide estimators of the optimal regime in the class $\mathcal{N} = \{\pi_\psi : \psi \text{ is arbitrary}\}$, i.e. of the regime $\pi_{\psi_{opt}}$ where $\psi_{opt} \equiv \arg\max_\psi V_\psi$, which are consistent regardless of whether or not the models for the Q-functions are correctly specified. Thus, if the models for the Q-functions are correct the direct method consistently estimates the global optimal regime; under misspecification of the Q-model the method still consistently estimates the optimal regime in the parametric class $\mathcal{N}$.

Setting aside computational difficulties, an instance of a direct method (see Zhang et al., 2012) would be to separately estimate V_ψ for each fixed ψ with its empirical analog $\widehat{V}_\psi = 4\mathbb{P}_n[Y1_{(2A_1-1)H_{1,1}^T\psi_{1,1}>0}1_{(2A_1-1)H_{2,1}^T\psi_{2,1}>0}]$ and next to estimate ψ_{opt} with $\arg\max_\psi \widehat{V}_\psi$. In contrast, the method studied in Appendix A estimates ψ_{opt} with $\widehat{\psi} = (\widehat{\psi}_{1,1}, \widehat{\psi}_{2,1})$, the estimator returned by the two-stage algorithm in that Appendix. In particular, $\widehat{\psi}_{2,1}$ converges in probability to $\psi_{2,1}^* = \arg\max_{\psi_{2,1}} P[YI_{(2A_2-1)H_{2,1}^T\psi_{2,1}>0}]$. However, $\psi_{2,1}^*$ is not equal to $\psi_{2,opt}$ if the model $H_{2,1}^T\psi_{2,1}$ is misspecified in the sense that there is no vector $\psi_{2,1}$ for which $E[Y|H_2, A_2 = 1] - E[Y|H_2, A_2 = 0] = H_{2,1}^T\psi_{2,1}$.This happens essentially because $P[YI_{(2A_2-1)H_{2,1}^T\psi_{2,1}>0}]$ is the mean over a subpopulation that includes subjects that did not follow the regime $1_{H_{1,1}^T\psi_{1,1}>0}$ at stage one and for these subjects the mean outcome under the stage two regime $I_{H_{2,1}^T\psi_{2,1}>0}$ might differ from the mean outcome of subjects that did follow the stage one regime $1_{H_{1,1}^T\psi_{1,1}>0}$ if the model $H_{2,1}^T\psi_{2,1}$ is incorrect.

To illustrate this point consider the simple setting in which no covariates X_1 and X_2 are measured and A_1, A_2 are sequentially randomized with probabilities $P(A_1 = 1) = 1/2$ and $P(A_2 = 1|A_1) = 1/2$. Suppose that, unknown to the data analyst,

$$\begin{aligned} &E\left(Y|A_1 = 1, A_2 = 1\right) = 100,\ \ E\left(Y|A_1 = 1, A_2 = 0\right) = 60, \\ &E\left(Y|A_1 = 0, A_2 = 1\right) = 20,\ \ \ E\left(Y|A_1 = 0, A_2 = 0\right) = 80 \end{aligned} \tag{0.3}$$

With these values, the optimal treatment regime is $A_1 = 1, A_2 = 1$. Also, $E(Y|A_1, A_2 = 1) - E(Y|A_1, A_2 = 0)$ depends on A_1. Suppose however that we mistakenly assume that $E(Y|A_1, A_2 = 1) - E(Y|A_1, A_2 = 0) = \psi_{2,1}$ is constant. Unaware of our error and before we see the data, we would reason as follows. The optimal stage two treatment is $A_2 = 1$ if $\psi_{2,1} > 0$ and $A_2 = 0$ if $\psi_{2,1} < 0$, that is, $\pi_{2,\psi_{2,1}}(A_1) = 1_{\psi_{2,1}>0}$. Letting $E(Y|A_1 = 1) - E(Y|A_1 = 0) \equiv \psi_{1,1}$ we likewise reason that the optimal stage one treatment is $\pi_{1,\psi_{1,1}} = 1_{\psi_{1,1}>0}$.Thus we would conclude that each of the four possible regimes (defined by the four possible assignments to the pair (a_1, a_2)) are implied by some parameter vector $(\psi_{1,1}, \psi_{2,1})$ as $(\psi_{1,1}, \psi_{2,1})$ varies over $\mathbb{R}^2$.

Suppose now the data with which to estimate the optimal regime are made available. Consider first the direct method based on maximizing the four empirical means. For a pair $(\psi_{1,1}, \psi_{2,1})$ such that $\psi_{1,1} > 0$ and $\psi_{2,1} > 0$, the empirical mean direct method estimator $\widehat{V}_\psi = 4\mathbb{P}_n[Y1_{(2A_1-1)\psi_{1,1}>0}1_{(2A_1-1)\psi_{2,1}>0}]$ is equal to $4\mathbb{P}_n[YA_1A_2]$ which converges in probability to $E(Y|A_1 = 1, A_2 = 1) = 100$. Likewise $\widehat{V}_\psi$ converges to $E(Y|A_1 = 1, A_2 = 0) = 70$ if $\psi_{1,1} > 0$ and $\psi_{2,1} < 0$ and so on. Thus, the direct method, consistently estimates the values of the four possible regimes and consequently with probability converging to 1, picks the optimal regime $A_1 = 1, A_2 = 1$. Consider instead the estimator studied in Appendix A. The sample mean $\mathbb{P}_n[Y1_{(2A_2-1)\psi_{2,1}>0}] = \mathbb{P}_n[YA_2]1_{\psi_{2,1}>0} + \mathbb{P}_n[Y(1-A_2)]1_{\psi_{2,1}<0}$ converges to $P[YA_2] = P[YA_1A_2] + P[Y(1-A_1)A_2] = (100+20)/4 = 30$ if $\psi_{2,1} > 0$ and to $P[Y(1-A_2)] = (60+80)/4 = 35$ if $\psi_{2,1} < 0$. Thus, the algorithm in Appendix A returns $\widehat{\psi}_{2,1} < 0$ with probability going to 1 and consequently erroneously estimates that the optimal stage two treatment is $A_2 = 0$.

Incidentally, in this example we can use a higher level argument to recognize that the authors estimators in Appendix A cannot be consistent under all data generating laws. Specifically, it is well known that any estimator of V_ψ that is regular and asymptotically linear (RAL) *under any data law*, must be asymptotically equivalent to an estimator of the form

$$\begin{aligned}\widehat{V}_\psi(r_1, r_2) = 4\mathbb{P}_n&\left[Y1_{(2A_1-1)\psi_{1,1}>0}1_{(2A_1-1)\psi_{2,1}>0}\right]\\ &+ \mathbb{P}_n\left[(A_2 - 1/2)r_2 + (A_1 - 1/2)r_1\right]\end{aligned} \tag{0.4}$$

for some constants r_1 and r_2. At the law (0.3), any RAL estimator of the value of the optimal regime must therefore be asymptotically equivalent to $\max_\psi\{\widehat{V}_\psi(r_1, r_2)\}$. The authors' estimator of the value of the optimal regime is RAL at laws with $E(Y|A_1, A_2 = 1) - E(Y|A_1, A_2 = 0)$ truly independent of A_1. If it were consistent for the optimal value function *under any data law*, then it would have to be asymptotically equivalent to $\max_\psi\{\widehat{V}_\psi(r_1, r_2)\}$ for some (r_1, r_2). However, this is not the case, so it cannot be consistent under all data laws.

Our remark here is meant as an observation rather than a critique because (a) when a concave relaxation is required to make the computations feasible, we do not know how to implement such a relaxation in two or multi stage studies without introducing model dependence, (b) estimators of the value function whose consistency relies solely on knowledge of the randomization probabilities are less efficient than their model based counterparts, and (c) even when a concave relaxation is not required as in our simple example with no covariates, and the unbiased estimator $4\mathbb{P}_n[Y1_{A_1=\pi_1}1_{A_2=\pi_2}]$ is used to estimate $\mathbb{E}^\pi[Y]$, we believe, with the authors, that the estimation of the optimal value and the associated optimal regime is not a regular problem; for instance in the preceding example the estimator of the optimal value is $4\max(\mathbb{P}_n[YA_1A_2], \mathbb{P}_n[Y(1-A_1)A_2], \mathbb{P}_n[YA_1(1-A_2)], \mathbb{P}_n[Y(1-A_1)(1-A_2)])$ which is not a regular estimator of $\max\{E(Y|A_1 = a_1, A_2 = a_2) : a_1, a_2 = 0, 1\}$ if the maximum is achieved at two or more of the four conditional means in the set.

Finally, we note that in the comprehensive monograph on optimal dynamic treatment regimes of van der Laan (2013), Theorem 2 claims that the optimal value is a regular pathwise differentiable parameter at all laws of the model, which includes those laws at which the optimal value is achieved at two or more conditional means. In contrast, Theorem 3.3 of the present article by Laber et al. implies that estimation of the optimal value function is an irregular problem at such exceptional laws, even when estimated with direct methods insensitive to model misspecification. We believe Laber et al's. result to be the correct one; Van der Laan's proof of Theorem 2 is incorrect owing to inappropriately applying the chain rule for differentiation to a composition of functions, one of which is non-differentiable at any such exceptional law.

Appendix

Proof of the Lemma. Sketch of the proof of part 1. Let $\beta^*_{2,1,n} = \beta^*_{2,1} + s/\sqrt{n}$.

$$
\begin{aligned}
&\sqrt{n}\left\{\widehat{\beta}_{1,1} - \beta^{\pi_2,*}_{1,1,n}|_{\pi_2=\widehat{\pi}_2^{dp}}\right\} \\
&= \mathbb{S}_{1,n} + \widehat{\Sigma}^{-1}_{1,1}\mathbb{P}_n\left[H_{1,1}\left\{A_1 - \mathbb{P}_n\left(A_1\right)\right\}\mathbb{U}_n\right] \\
&\quad - \Sigma^{-1}_{1,1\infty}\sqrt{n}P_n\left[H_{1,1}\left(A_1 - \frac{1}{2}\right)\left(1_{H^T_{2,1}\beta_{2,1}>0} - 1_{H^T_{2,1}\beta^*_{2,1,n}>0}\right)\right. \\
&\quad \left.\left.\times\, H^T_{2,1}\beta^*_{2,1,n}1_{H^T_{2,1}\beta^*_{2,1}>0}\right]\right|_{\beta_{2,1}=\widehat{\beta}_{2,1}} \\
&\quad - \Sigma^{-1}_{1,1\infty}\sqrt{n}P_n\left[H_{1,1}\left(A_1 - \frac{1}{2}\right)\left(1_{H^T_{2,1}\beta_{2,1}>0} - 1_{H^T_{2,1}\beta^*_{2,1,n}>0}\right)\right. \\
&\quad \left.\left.\times\, H^T_{2,1}\beta^*_{2,1,n}1_{H^T_{2,1}\beta^*_{2,1}=0}\right]\right|_{\beta_{2,1}=\widehat{\beta}_{2,1}}
\end{aligned}
$$

Under the underlying generative distribution P_n satisfying $(A3)$ of LLQPM, by their Theorem 4.1, part 2. we have

$$
\begin{aligned}
&\mathbb{S}_{1,n} + \widehat{\Sigma}^{-1}_{1,1}\mathbb{P}_n\left[H_{1,1}\left\{A_1 - \mathbb{P}_n\left(A_1\right)\right\}\mathbb{U}_n\right] \\
&\rightsquigarrow \mathbb{S}_{1,\infty} + \Sigma^{-1}_{1,1\infty}P\left[H_{1,1}\left(A_1 - \frac{1}{2}\right)H^T_{2,1}\mathbb{V}_\infty 1_{H^T_{2,1}\beta^*_{2,1}>0}\right] \\
&\quad + \Sigma^{-1}_{1,1\infty}P\left[H_{1,1}\left(A_1 - \frac{1}{2}\right)\left\{\left[H^T_{2,1}\left(\mathbb{V}_\infty + s\right)\right]_+ - \left[H^T_{2,1}s\right]_+\right\}1_{H^T_{2,1}\beta^*_{2,1}=0}\right]
\end{aligned}
$$

Now, $P_n[1_{H^T_{2,1}\widehat{\beta}_{2,1}>0} = 1|H^T_{2,1}\beta^*_{2,1} > 0] = P_n[1_{\sqrt{n}H^T_{2,1}(\widehat{\beta}_{2,1}-\beta^*_{2,1,n})+\sqrt{n}H^T_{2,1}\beta^*_{2,1,n}>0} = 1|H^T_{2,1}\beta^*_{2,1} > 0] \to 1$. Also, for $h^T_{2,1}\beta^*_{2,1} > 0$ it holds that $1_{h^T_{2,1}\beta^*_{2,1,n}>0} = 1$ for n sufficiently large. Then,

$$
\begin{aligned}
&P_n\left[P_n\left[H_{1,1}\left(A_1 - \frac{1}{2}\right)\left(1_{H^T_{2,1}\beta_{2,1}>0} - 1_{H^T_{2,1}\beta^*_{2,1,n}>0}\right)\right.\right. \\
&\quad \left.\left.\left.\times\, H^T_{2,1}\beta^*_{2,1,n}1_{H^T_{2,1}\beta^*_{2,1}>0}\right]\right|_{\beta_{2,1}=\widehat{\beta}_{2,1}} = 0\right] \to 1
\end{aligned}
$$

Consequently

$$
\begin{aligned}
&\Sigma_{1,1\infty}^{-1}\sqrt{n}P_n\left[H_{1,1}\left(A_1-\frac{1}{2}\right)\left(1_{H_{2,1}^T\beta_{2,1}>0}-1_{H_{2,1}^T\beta_{2,1,n}^*>0}\right)\right.\\
&\quad\left.\left.\times H_{2,1}^T\beta_{2,1,n}^*1_{H_{2,1}^T\beta_{2,1}^*>0}\right]\right|_{\beta_{2,1}=\widehat{\beta}_{2,1}}\\
&\rightsquigarrow 0
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&\Sigma_{1,1\infty}^{-1}\sqrt{n}P_n\left[H_{1,1}\left(A_1-\frac{1}{2}\right)\left(1_{H_{2,1}^T\beta_{2,1}>0}-1_{H_{2,1}^T\beta_{2,1,n}^*>0}\right)\right.\\
&\quad\left.\left.\times H_{2,1}^T\beta_{2,1,n}^*1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\right|_{\beta_{2,1}=\widehat{\beta}_{2,1}}\\
&=\Sigma_{1,1\infty}^{-1}P_n\left[H_{1,1}\left(A_1-\frac{1}{2}\right)\left(1_{\sqrt{n}H_{2,1}^T\left(\beta_{2,1}-\beta_{2,1,n}^*\right)+H_{2,1}^Ts>0}-1_{H_{2,1}^Ts>0}\right)\right.\\
&\quad\left.\left.\times H_{2,1}^Ts1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\right|_{\beta_{2,1}=\widehat{\beta}_{2,1}}\\
&\rightsquigarrow\Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)\left(1_{H_{2,1}^T(\mathbb{V}_\infty+s)>0}-1_{H_{2,1}^Ts>0}\right)H_{2,1}^Ts1_{H_{2,1}^T\beta_{2,1}^*=0}\right]
\end{aligned}
$$

So,

$$
\begin{aligned}
&\sqrt{n}\left\{\widehat{\beta}_{1,1}-\beta_{1,1,n}^{\pi_2,*}|_{\pi_2=\widehat{\pi}_2^{dp}}\right\}\\
&\rightsquigarrow\mathbb{S}_{1,\infty}+\Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)H_{2,1}^T\mathbb{V}_\infty1_{H_{2,1}^T\beta_{2,1}^*>0}\right]\\
&\quad+\Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)\left\{\left[H_{2,1}^T\left(\mathbb{V}_\infty+s\right)\right]_+-\left[H_{2,1}^Ts\right]_+\right\}1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\\
&\quad-\Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)\left(1_{H_{2,1}^T(\mathbb{V}_\infty+s)>0}-1_{H_{2,1}^Ts>0}\right)H_{2,1}^Ts1_{H_{2,1}^T\beta_{2,1}^*=0}\right]
\end{aligned}
$$

The result follows from

$$
\begin{aligned}
&\Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)\left(1_{H_{2,1}^T(\mathbb{V}_\infty+s)>0}-1_{H_{2,1}^Ts>0}\right)H_{2,1}^Ts1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\\
&=\Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)\left(1_{H_{2,1}^T(\mathbb{V}_\infty+s)>0}\right)H_{2,1}^T\left(\mathbb{V}_\infty+s\right)1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\\
&\quad-\Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)1_{H_{2,1}^T(\mathbb{V}_\infty+s)>0}H_{2,1}^T\mathbb{V}_\infty1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\\
&\quad-\Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)1_{H_{2,1}^Ts>0}H_{2,1}^Ts1_{H_{2,1}^T\beta_{2,1}^*=0}\right]\\
&=\Sigma_{1,1\infty}^{-1}P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)\left\{\left[H_{2,1}^T\left(\mathbb{V}_\infty+s\right)\right]_+-\left[H_{2,1}^Ts\right]_+\right\}1_{H_{2,1}^T\beta_{2,1}^*=0}\right]
\end{aligned}
$$

$$- \Sigma_{1,1\infty}^{-1} P\left[H_{1,1}\left(A_1-\frac{1}{2}\right)1_{H_{2,1}^T(\mathbb{V}_\infty+s)>0}H_{2,1}^T\mathbb{V}_\infty 1_{H_{2,1}^T\beta_{2,1}^*=0}\right]$$

Proof of part (2).

Let $p_{a_1,x_1,x_2} \equiv \Pr[A_1=a_1, X_1=x_1, X_2=x_2]$, $g(X_1,A_1) \equiv c_1^T\Sigma_{1,1\infty}^{-1}H_{1,1}(A_1-\frac{1}{2})$, $\mathbb{Z}(\gamma) \equiv P[g(X_1,A_1)1_{H_{2,1}^T(\mathbb{V}_\infty+\gamma)>0}H_{2,1}^T\mathbb{V}_\infty 1_{H_{2,1}^T\beta_{2,1}^*=0}]$ and $\mathbb{W}(\gamma) \equiv P[g(X_1,A_1)\{[H_{2,1}^T(\mathbb{V}_\infty+\gamma)]_+ - [H_{2,1}^T\gamma]_+\}1_{H_{2,1}^T\beta_{2,1}^*=0}]$. For any $r \in \mathbb{R}^8$ let r_{a_1,x_1,x_2} denote the realization of $H_{2,1}^T r$ when $A_1=a_1, X_1=x_1$ and $X_2=x_2$. When the realized value of $\mathbb{V}_\infty$ is $v \in \mathbb{R}^8$ we let $z(\gamma)$ and $w(\gamma)$ denote the realized values of $\mathbb{Z}(\gamma)$ and $\mathbb{W}(\gamma)$ respectively and in a slight abuse of notation we let

$$z\left(a_1,x_1,x_2;\gamma_{a_1,x_1,x_2}\right) \equiv g\left(a_1,x_1\right)1_{v_{a_1,x_1,x_2}+\gamma_{a_1,x_1,x_2}>0}v_{a_1,x_1,x_2}$$

be the realized value of $g(X_1,A_1)1_{H_{2,1}^T(\mathbb{V}_\infty+\gamma)>0}H_{2,1}^T\mathbb{V}_\infty$ when $A_1=a_1, X_1=x_1$ and $X_2=x_2$. Likewise let

$$w\left(a_1,x_1,x_2;\gamma_{a_1,x_1,x_2}\right) \equiv g\left(a_1,x_1\right)\left\{1_{v_{a_1,x_1,x_2}+\gamma_{a_1,x_1,x_2}>0}\left(v_{a_1,x_1,x_2}+\gamma_{a_1,x_1,x_2}\right)\right. \\ \left. - 1_{\gamma_{a_1,x_1,x_2}>0}\gamma_{a_1,x_1,x_2}\right\}$$

be the realized value of $g(X_1,A_1)\{[H_{2,1}^T(\mathbb{V}_\infty+\gamma)]_+ - [H_{2,1}^T\gamma]_+\}$.

We then have that the realized values of $\sup_{\gamma\in\mathbb{R}^8}\mathbb{Z}(\gamma)$ and of $\sup_{\gamma\in\mathbb{R}^8}\mathbb{W}(\gamma)$ are respectively,

$$\sup_{\gamma\in\mathbb{R}^8} z\left(\gamma\right) \\ = \sum_{a_1=0}^{1}\sum_{x_1=0}^{1}\sum_{x_2=0}^{1} 1_{\left(\beta_{2,1}^*\right)_{a_1,x_1,x_2}=0}p_{a_1,x_1,x_2}\sup_{\gamma_{a_1,x_1,x_2}\in\mathbb{R}}\left[z\left(a_1,x_1,x_2;\gamma_{a_1,x_1,x_2}\right)\right]$$

$$\sup_{\gamma\in\mathbb{R}^8} w\left(\gamma\right) \\ = \sum_{a_1=0}^{1}\sum_{x_1=0}^{1}\sum_{x_2=0}^{1} 1_{\left(\beta_{2,1}^*\right)_{a_1,x_1,x_2}=0}p_{a_1,x_1,x_2}\sup_{\gamma_{a_1,x_1,x_2}\in\mathbb{R}}\left[w\left(a_1,x_1,x_2;\gamma_{a_1,x_1,x_2}\right)\right]$$

Now,

$$z\left(a_1,x_1,x_2;\gamma_{a_1,x_1,x_2}\right) = \begin{cases} 0 & \text{if } \gamma_{a_1,x_1,x_2} \leq -v_{a_1,x_1,x_2} \\ g\left(a_1,x_1\right)v_{a_1,x_1,x_2} & \text{if } \gamma_{a_1,x_1,x_2} > -v_{a_1,x_1,x_2} \end{cases}$$

On the other hand, if $v_{a_1,x_1,x_2} > 0$ then

$$w\left(a_1,x_1,x_2;\gamma_{a_1,x_1,x_2}\right) = \begin{cases} 0 & \text{if } \gamma_{a_1,x_1,x_2} \leq -v_{a_1,x_1,x_2} \\ g\left(a_1,x_1\right)\left(v_{a_1,x_1,x_2}+\gamma_{a_1,x_1,x_2}\right) & \\ \quad \text{if } -v_{a_1,x_1,x_2} < \gamma_{a_1,x_1,x_2} \leq 0 & \\ g\left(a_1,x_1\right)v_{a_1,x_1,x_2} & \text{if } \ 0 \leq \gamma_{a_1,x_1,x_2} \end{cases}$$

So, if $v_{a_1,x_1,x_2} > 0$ we conclude that if $g(a_1,x_1) > 0$, then $\sup_{\gamma_{a_1,x_1,x_2}\in\mathbb{R}}[z(a_1,x_1,x_2;\gamma_{a_1,x_1,x_2})] = \sup_{\gamma_{a_1,x_1,x_2}\in\mathbb{R}}[w(a_1,x_1,x_2;\gamma_{a_1,x_1,x_2})] = g(a_1,x_1)v_{a_1,x_1,x_2}$ and if

$g(a_1, x_1) \leq 0$, then $\sup_{\gamma_{a_1,x_1,x_2} \in \mathbb{R}}[z(a_1, x_1, x_2; \gamma_{a_1,x_1,x_2})] = \sup_{\gamma_{a_1,x_1,x_2} \in \mathbb{R}}[w(a_1, x_1, x_2; \gamma_{a_1,x_1,x_2})] = 0$.

If $v_{a_1,x_1,x_2} \leq 0$ then

$$w\left(a_1, x_1, x_2; \gamma_{a_1,x_1,x_2}\right) = \begin{cases} 0 & \text{if } \gamma_{a_1,x_1,x_2} \leq 0 \\ -g\left(a_1, x_1\right) \gamma_{a_1,x_1,x_2} & \text{if } 0 < \gamma_{a_1,x_1,x_2} \leq -v_{a_1,x_1,x_2} \\ g\left(a_1, x_1\right) v_{a_1,x_1,x_2} & \text{if } \gamma_{a_1,x_1,x_2} > -v_{a_1,x_1,x_2} \end{cases}$$

So, if $v_{a_1,x_1,x_2} \leq 0$ we conclude that if $g(a_1, x_1) > 0$, then $\sup_{\gamma_{a_1,x_1,x_2} \in \mathbb{R}}[z(a_1, x_1, x_2; \gamma_{a_1,x_1,x_2})] = \sup_{\gamma_{a_1,x_1,x_2} \in \mathbb{R}}[w(a_1, x_1, x_2; \gamma_{a_1,x_1,x_2})] = 0$ and if $g(a_1, x_1) \leq 0$, then $\sup_{\gamma_{a_1,x_1,x_2} \in \mathbb{R}}[z(a_1, x_1, x_2; \gamma_{a_1,x_1,x_2})] = \sup_{\gamma_{a_1,x_1,x_2} \in \mathbb{R}}[w(a_1, x_1, x_2; \gamma_{a_1,x_1,x_2})] = g(a_1, x_1) v_{a_1,x_1,x_2}$.

We thus conclude that regardless of the values of v_{a_1,x_1,x_2} and $g(a_1, x_1)$, it holds that $\sup_{\gamma_{a_1,x_1,x_2} \in \mathbb{R}}[z(a_1, x_1, x_2; \gamma_{a_1,x_1,x_2})] = \sup_{\gamma_{a_1,x_1,x_2} \in \mathbb{R}}[w(a_1, x_1, x_2; \gamma_{a_1,x_1,x_2})]$ and consequently, that $\sup_{\gamma \in \mathbb{R}^8} z(\gamma) = \sup_{\gamma \in \mathbb{R}^8} w(\gamma)$.

The claim for the equality of the infimums follows by an analogous analysis.

Proof of Part (3). This follows from Parts (1) and (2) of the Lemma and part (3) of Theorem 4.1 of LLQPM. □

References

Laber, E.B. et al. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electron. J. Statist.*, 8, 1225–1272.

Orellana, L., Rotnitzky, A. and Robins, J. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: Main content. *Int. Journal of Biostatistics*, 6(2). MR2602551

Qian, M. and Murphy, S. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2), 1180–1210. MR2816351

Robins, J.M., Orellana, L. and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, pages 4678–4721. MR2528576

Robins, J.M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*. MR2129402

van der Laan, M.J. and Petersen, M.L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1), Article 3. MR2306841

van der Laan, M. (October 2013). Targetted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. *Berkeley Division of Biostatistics Working Paper Series. Working Paper 317.* `http://biostats.bepress.com/ucbbiostat/paper317`.

Zhang, B., Tsiatis, A., Laber, E. and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, 68, 1010–1018. MR3040007

Zhao, Y., Zeng, D., Rush, J. and Kosorok, M. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499), 1106–1118. MR3010898