# Inference for Dynamic Treatment Regimes: Non-regular Asymptotics under Different Settings

Yating Zou

Gillings School of Public Health
University of North Carolina at Chapel Hill

Nov 11, 2022

# Overview

1. Intro to Paper

   Inference of Value Function for RL in Infinite-Horizon Settings[1]

2. Regularity Conditions
   - What are they?
   - When is there a problem?
   - How to solve?

3. Asymptotic Inference for DTR when Using:
   - Outcome Weighted Learning (OWL)
   - Reinforcement learning (RL)
     - Finite Horizon
     - Infinite Horizon *

[1] Shi et al., (2021)

## Motivating Example

Mobile Health - infinite timepoints, needs to find the best policy when there's no pre-determined stopping point.

**Question**:
How to Quantify Uncertainty using asymptotic Confidence Intervals (CI) for the Value function associated with the estimated optimal DTR?

**Other Major Contributions:**

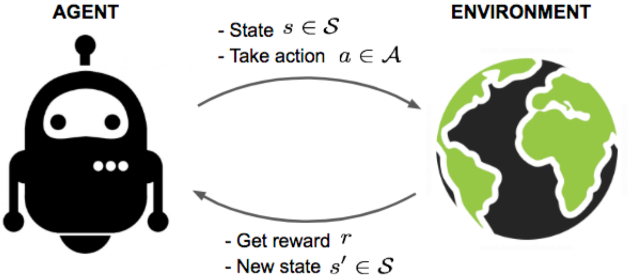
- Non-asyptotic error bound
- Characterize the approximation error for the Value Function.
- Valid in non-regular cases where opt DTR is not unique
- Converge as long as either n or t $\to \infty$.

**Other Major Contributions:**

- Converge as long as either n or t $\to \infty$.
- Non-asyptotic *error bound*
- Characterize the *approximation error* for the Value Function.
- Valid in *non-regular* cases where opt DTR is not unique

## Quantifying Uncertainty

Recall Basic Ingredients in Reinforcement Learning:



Recall Standard Notations:

- Markov Decision Process (MDP): $(\mathbb{X}, \mathcal{A}, \mathcal{P}, \gamma, R)$, where
  - $\mathbb{X}$ a subspace of $\mathbb{R}^d$
  - $\mathcal{A} = \{0, 1, \ldots, m-1\}$
  - $\mathcal{P}(S|x, a)$ transition probability given $x$ and $a$
  - $\gamma$ the discount factor
  - R: $\mathbb{X} \times \mathcal{A} \to \mathbb{R}$, $R(x, a) := \mathbb{E}(Y|X = x, A = a)$
- Policy $\pi(\cdot|x) : \mathcal{X} \to \mathcal{P}(\mathcal{A})$, a probability distribution over $\mathcal{A}$
- Value Function associated with a Policy:

$$V(\pi; x) = \sum_{t \geq 0} \gamma^t \mathbb{E}^\pi (Y_t | X_{t=0} = x)$$

$$Q(\pi; x, a) = \sum_{t \geq 0} \gamma^t \mathbb{E}^\pi (Y_t | X_{t=0} = x, A_{t=0} = a)$$

- Goal: Use data $\{(X, A, Y)_{i,t}\}_{i\in\{1,2,\ldots,n\},t\geq 0}$,

  Find $\pi^* \in \arg\max_{\pi\in\Pi} V^\pi$, and quantify its uncertainty

# Section 1: Intro to Paper

Big Picture:

- $\pi^* \in \arg\max_{\pi \in \Pi} V^\pi$ (Why exist?)
- Value Function can be formulated using either Q or V (Why? Which?)
- Don't know the Value Function $\rightarrow$ Estimate (How?)

Big Picture:

- $\pi^* \in \arg\max_{\pi \in \Pi} ValueFunction^*$ (Why exist?)

## Add Assumptions

Denote history up to and not including t as $H_t = \{(Y_j, X_j, A_j)\}_{0 \leq j < t}$

1. Markov Assumption (MA):

$$Pr(X_{t+1} \in S | X_t = x, A_t = a, H_t) = \mathcal{P}(S|x, a) \text{ , for } S \text{ any subset of } \mathbb{X}$$

2. Conditional Mean Independence Assumption (CMIA):

$$\mathbb{E}(Y_t | X_t = x, A_t = a, H_t) = \mathbb{E}(Y_t | X_t = x, A_t = a) = r(x, a)$$

3. There exists at least one optimal policy $\pi^*$ such that $V(\pi^*; x) \geq V(\pi; x)$, $\forall \pi, x$ (Puterman, 1994)

- Value Function can be formulated using either Q or V (Why? Which?)
- Don't know the Value Function $\rightarrow$ Estimate (How?)

Big Picture:

- $\pi^* \in \arg\max_{\pi \in \Pi} ValueFunction^*$ (Why $\pi^*$ exist?)

### Fixed Point Theorem

1. Bellman Operator $\mathcal{B} : \mathcal{F} \to \mathcal{F}$, where $\mathcal{F}$ is a space of functions on $\mathcal{S}$

$$\|\mathcal{B}V_1 - \mathcal{B}V_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$$

2. Fixed Point Theorem: The sequence $V, \mathcal{B}V, \mathcal{B}^2V, \ldots$ converges for every $V$, and the limit $V^*$ is a unique fixed point, 'fixed' in the sense $\mathcal{B}V^* = V^*$
3. Rmk: $V^*$ unique, but $\pi$ might not be unique*.

- Value Function can be formulated using either Q or V (Why? Which?)
- Don't know the Value Function $\to$ Estimate (How?)

Big Picture:

- $\pi^* = \arg\max ValueFunction^*$ (Why exist?)
- Value Function can be formulated using either Q or V (Why? Which?)

**Infinite-Horizon**

1. $V(\pi; x) = \sum_{a \in \mathcal{A}} Q(\pi; x, a)\pi(a|x)$
2. $V^*(s) = Q^*(s, \pi^*(s))$

- Don't know the Value Function $\rightarrow$ Estimate (How?)

Big Picture:

- $\pi^{opt} = \arg\max ValueFunction^{opt}$ (Why exist?)
- Value Function can be formulated using either Q or V (Why? Which?)

### Take Into Account our goal of Inference

When would there be sufficient smoothness?

1. When $\pi$ is not continuous in $a$ for any given $x$, $V(\pi; \cdot)$ would not be continuous $\rightarrow$ Raising a problem for non-constant deterministic policy.
2. When $r(\cdot, a)$ is smooth, $Q(\pi, \cdot, a)$ is p-smooth $\rightarrow$ Can deal with both deterministic and random policies.

- Don't know the Value Function $\rightarrow$ Estimate (How?)

Big Picture:

- $\pi^{opt} = \arg\max ValueFunction^{opt}$ (Why exist?)
- Value Function can be formulated using either Q or V (Why? Which?)

## Another Advantage

1. If use $V$, need to estimate the data generating behavior policy, say $b(a|x)$, and adjust the value function $V(\pi; X)$ by a weight $\frac{\pi(A,X)}{b(A,X)}$

$$0 = \mathbb{E}\Big[\frac{\pi(A_t; X_t)}{b(A_t; X_t)}(Y_t + \gamma V(\pi, X_{t+1}) - V(\pi, X_t))|X_t = x_t\Big]$$

- Don't know the Value Function $\rightarrow$ Estimate (How?)

Big Picture:

- $\pi^{opt} = \arg\max ValueFunction^{opt}$ (Why exist?)
- Value Function can be formulated using either Q or V (Why? Which?)
- Don't know the Value Function $\rightarrow$ Estimate (How?)

## A Common Solution – Linear Parametrization

1. $Q(\pi; x, a) = \Phi_L(x)^\intercal \beta_{\pi,a}$, $\forall x \in \mathbb{X}, a \in \mathcal{A}$, where $\Phi_L(\cdot) = \{\phi_{L,1}(\cdot), \phi_{L,2}(\cdot), \cdots, \phi_{L,L}(\cdot)\}^\intercal$ a vector of $L$ basis functions.

$$
\begin{aligned}
V(\pi; x) &= \sum_{a \in \mathcal{A}} Q(\pi; x, a)\pi(a|x) \\
&= \sum_{a \in \mathcal{A}} \Phi_L(x)^\intercal \beta_{\pi,a}\pi(a|x) = U_\pi(x)^\intercal \beta_\pi
\end{aligned}
$$

All Together - Inference under a fixed policy (thus unique):

1. parametrize $Q(\pi; x, a) = \Phi_L^T(x)\beta_{\pi,a}$
2. Estimate $\beta$ from the Bellman Equation

$$\mathbb{E}\Big[\Big\{ \underbrace{Y_t + \gamma \sum_{a \in \mathcal{A}} Q(\pi; X_{t+1}, a)\pi(a|X_{t+1}) - Q(\pi; X_t, A_t)}_{\text{temporal difference error}} \Big| X_t, A_t \Big]\Big\} = 0$$

3. obtain CI for $\widehat{V}(\pi; \mathbb{G})$, $\mathbb{G}$ a reference distribution for $X$, using

$$\frac{V(\pi; \mathbb{G}) - \widehat{V}(\pi; \mathbb{G})}{(nT)^{-1/2}\widehat{\sigma}(\pi; \mathbb{G})}$$

$$= \frac{(nT)^{-1/2}}{\sigma(\pi; \mathbb{G})} \sum_{i,t} \left\{ \int U_\pi(x)\mathbb{G}(x) \right\}^T \Sigma_\pi^{-1} \xi_{i,t}\epsilon_{\pi,i,t} + o_p(1)$$

All Together - Inference under an estimated policy (possibly not unique):

1. parametrize $Q(\pi; x, a) = \Phi_L^T(x)\beta_{\pi,a}$
2. Estimate $\beta$ from the Bellman Equation
3. obtain CI for $\widehat{V}(\widehat{\pi}; \mathbb{G})$, $\mathbb{G}$ a reference distribution of $X$, using

$$\frac{V(\widehat{\pi}; \mathbb{G}) - \widehat{V}(\widehat{\pi}; \mathbb{G})}{(nT)^{-1/2}\widehat{\sigma}(\widehat{\pi}; \mathbb{G})}$$
$$= \frac{(nT)^{-1/2}}{\sigma(\widehat{\pi}; \mathbb{G})} \sum_{i,t} \left\{ \int U_{\widehat{\pi}}(x)\mathbb{G}(x) \right\}^T \Sigma_{\widehat{\pi}}^{-1} \xi_{i,t} \epsilon_{\widehat{\pi},i,t} + o_p(1)$$

smoothness... uniqueness...

But what is ‘Regularity’, specifically?

# Section 2: Regularity

Usually, the steps to quantifying uncertainty:
point approximation $\rightarrow$ local approximation

asymptotically unbiased (consistency)
$\rightarrow$ asymptotic normality
$\rightarrow$ smallest possible variance (efficiency)
$\rightarrow$ finite inference

We usually assume regularity conditions to begin with proving these results. The specific conditions differ case-by-case.

A example using M-estimators
For $\{m(X, \theta) : \theta \in \Theta\}, m_\theta : \mathbb{X} \to \mathbb{R}$, $\{X_i\}_{i=1,\ldots,n}$ i.i.d.

- Assume the true parameter $\theta_0 = \arg\min_{\theta\in\Theta} Pm(X, \theta)$
- However, only have $\widehat{P}_n$, an empirical measure. So $\widehat{\theta}_n = \arg\min_{\theta\in\Theta} \widehat{P}_n m(X, \theta) = \arg\min_{\theta\in\Theta} \frac{1}{n} \sum_{i\leq n} m(X_i, \theta)$

How good does $\widehat{\theta}$ approximate $\theta_0$?
Assume consistency: $\widehat{\theta}_n = \theta_0 + o_p(1)$

Consistency:

- $\theta_0$ is the unique minimizer of $Pm(X, \theta)$
  - Assume a true $\theta_0$ exist (a philosophical argument...)
  - Assume $\theta_0$ can be identified

- $\widehat{\theta}_n$ is the unique minimizer of $\widehat{P}_n m(X, \theta)$
  - If $m(X, \theta)$ continuous in $\theta$, a compact parameter space
  - exist by Extreme Value Theorem

- If $\widehat{P}_n m(X, \theta) \to Pm(X, \theta)$ uniformly over $\Theta$

- Then $\widehat{\theta}_n \to \theta_0$ in probability

**Local quadratic approximation using Taylor Expansion:**

$$f(x) = f(a) + f^{(1)}(a)(x-a) + \frac{1}{2}f^{(2)}(a)(x-a)^2 + \cdots$$

Replace x with $\theta_0$, $a$ with $\widehat{\theta}_n$, integrate over $\widehat{P}_n$ of a random variable $X$:

$$\widehat{P}_n m(X, \widehat{\theta}_n) =$$
$$\frac{1}{n}\sum_{i \le n}\left\{m(X_i, \theta_0) + m^{(1)}(X_i, \theta_0)d + \frac{1}{2}m^{(2)}(X_i, \theta_0)d^2 + R_n(|d|^3)\right\}$$

where $d = (\widehat{\theta}_n - \theta_0)$.

Equivalently, in a cleaner, vector form, with $Z_n = \frac{1}{\sqrt{n}} \sum_{i \leq n} m^{(1)}(X_i, \theta_0)$, $J_n = \sum_{i \leq n} m^{(2)}(X_i, \theta_0)$

**Local approximation**

$$\widehat{P}_n m(X, \widehat{\theta}_n) - \widehat{P}_n m(X, \theta_0) = \frac{1}{\sqrt{n}} d^\intercal Z_n + \frac{1}{2} d^\intercal J_n d + R_n(|d|^3)$$

- If we want to minimize this expression
- Consider the existence of such expansion and the validity of desired operations under a distributional argument

# Section 2: Regularity - M-estimators

## Local approximation

$$\widehat{P}_n m(X, \widehat{\theta}_n) - \widehat{P}_n m(X, \theta_0) = \frac{1}{\sqrt{n}} d^\intercal Z_n + \frac{1}{2} d^\intercal J_n d + R_n(|d|^3)$$

**Classical Regularity Conditions**:
If at $\theta_0$, there is a neighborhood $\mathcal{N}(\theta_0)$ of $\theta_0 = \arg\min_\theta Pm(X, \theta)$, with $\theta \in \Theta$, satisfying

- Interior Point:
  - $\theta_0$ an interior point of $\Theta$ (otherwise zero derivative might not be equivalent to being an extreme point)
- Smoothness within $\mathcal{N}(\theta_0)$:
  - For almost all x under $P$, derivatives up to the third order exists and derivatives up to the second order can go under the integral sign.
  - $J = Pm^{(2)}(X, \theta_0)$, the Fisher Information, is positive definite.
  - $R_n(|d|^3)$ can be bounded,
    that is, $sup_{\theta \in \Theta} |m^{(3)}(x, \theta)| \leq M(x)$, with $\mathbb{E}M(X) < \infty$.

**Local approximation**

$$\widehat{P}_n m(X, \widehat{\theta}_n) - \widehat{P}_n m(X, \theta_0) = \frac{1}{\sqrt{n}} d^\intercal Z_n + \frac{1}{2} d^\intercal J_n d + R_n(|d|^3)$$

Under regularity conditions, with $\widehat{\theta}_n \xrightarrow{p} \theta_0$, show

- $R_n(|d|^3)/|d|^3 \xrightarrow{p} 0$
- $J_n \xrightarrow{p} J$ (if J exists, by WLLN)
- $\widehat{\theta}_n$ within $o_p(1/n)$ in minimizing $P_n m(X, \theta)$

Then

- $\widehat{\theta}_n = \theta_0 - J_n^{-1} Z_n / \sqrt{n} + o_p(1/\sqrt{n})$
- If $Z_n \xrightarrow{d} N(0, \Sigma)$, then $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(J^{-1} \Sigma J)$

## Asymptotic Inference under Semiparametric Setting

The von Mises Expansion:

$$\sqrt{n}(\widehat{\Psi} - \Psi) = \sqrt{n} \int \phi(O_i, \widehat{P}_n) d(\widehat{P}_n - P)(O) + R_2(\widehat{P}_n, P)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\phi(O_i, P)\} - \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\phi(O_i, \widehat{P}_n)\}}_{\text{Plug-in bias}}$$

$$+ \underbrace{\sqrt{n}(P_n - P)\{\phi(O, \widehat{P}_n) - \phi(O, P)\}}_{\text{Empirical Process Term}} + \underbrace{R_2(\widehat{P}_n, P)}_{\text{Remainder}}.$$

- First term: $\xrightarrow{d} \mathcal{N}(0, Var(\phi(O, P)))$
- Empirical Process term: assume Donsker condition / use cross-validation
- Remainder term: by controlling convergence rate of nuisance parameter

Rmk: Here, a similar problem of non-smoothness would be pathwise non-differentiable.

When is there a problem of Regularity in DTR research?

Recall Models related to DTR:

- DTR
  - RL
    - Finite-Horizon
      - ...
      - Non-unique DTR (Luedtke & van der Laan, 2016)
    - Infinite-Horizon
      - GGQ-Learning (Ertefaie & Strawderman, 2018)
      - V-Learning (Luckett et al., 2020)
      - pT-Learning (Zhou et al., 2021)
      - A debiased approach (Shi et al., 2021)
      - Double Reinforcement Learning (Kallus & Uehara, 2021)
      - Non-unique DTR (Shi et al., 2021)
  - OWL
    - One-Stage
      - Outcome Weighted Learning (Zhao et al., 2012)
      - Residual Weighted Learning (Zhou et al., 2017)
      - Multiple treatment options with risk bounds (Zhou, 2018)
    - Multi-Stage
      - Backward/Simultaneous OWL with finite sample bounds (Zhao et al., 2015)
      - ...

## Reinforcement Learning

A possible issue:

- Greedy Gradient Q-Learning (Ertefaie & Strawderman, 2018)

$$0 = \mathbb{E}\Big[R^t + \gamma \max_{a \in \mathcal{A}} Q^*(S^{t+1}, a) - Q^*(s^t, a^t))|S^t = s^t, A^t = a^t\Big]$$

## Reinforcement Learning

$$Q_2(H_2, A_2) = \mathbb{E}[Y_2 | H_2, A_2] \tag{1}$$
$$Q_1(H_1, A_1) = \mathbb{E}[Y_1 + \max_{a_j} Q_2(H_2, a_2) | H_1, A_1] \tag{2}$$

Consider a linear model with $\psi$ our parameter of interest:

$$Q(H, A; \beta, \phi) = \beta^T H_1 + (\psi^T H_2)A,$$
$$\text{where } A \in \{-1, 1\}, H = (H_1, H_2)^T$$

Then in step (2),

$$\widehat{Y}_1 = Y_1 + \widehat{\beta}_2^T H_{2,0} + |\widehat{\psi}_2^T H_{2,1}|, \text{ non-smooth in } \psi$$

$$\widehat{Y}_1 = Y_1 + \widehat{\beta}_2^T H_{2,0} + |\widehat{\psi}_2^T H_{2,1}|, \text{ non-smooth in } \psi$$

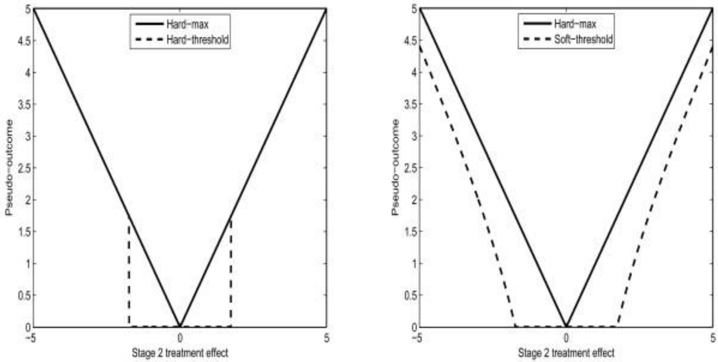
Hard Threshold:

$$|\widehat{\psi}_2^T H_{2,1}| I\left\{ \frac{\sqrt{n}|\widehat{\psi}_2^T H_{2,1}|}{\sqrt{H_{2,1}^\intercal \widehat{\Sigma}_2 H_{2,1}}} \geq z_{\alpha/2} \right\}$$

Soft Threshold:

$$|\widehat{\psi}_2^T H_{2,1}| I\left(1 - \frac{\lambda}{|\widehat{\psi}_2^\intercal H_{2,1}|^2}\right)^+$$

mitigate the problem



## Outcome Weighted Learning

DTR as a weighted classification problem:

- Single stage OWL (Zhao et al., 2012)

  Minimize classification risk

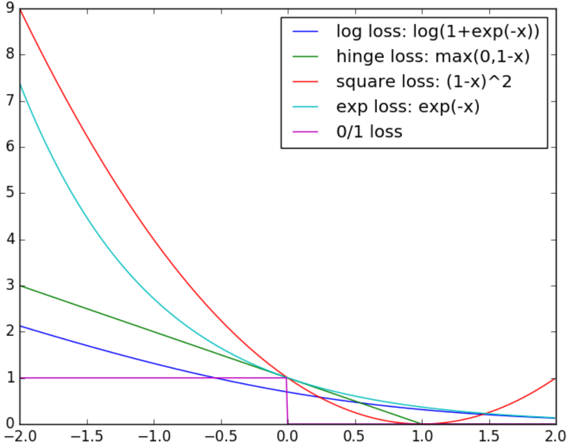
$$\mathbb{E}\left[\frac{Y}{b(A,X)} I(A \neq \pi(X))\right]$$

- Finite Multi-stage OWL (Zhao et al., 2015)

  Minimize weighted cumulative risk

$$\mathbb{E}\left[\frac{(\sum_{j=t}^{T} Y_j) \prod_{j=t+1}^{T} I(A_j = \pi_j^*(H_j))}{\prod_{j=t}^{T} b_j(A_j, H_j)} I(A_t \neq \pi_t(H_t))\right]$$

mitigate the problem using convex surrogates



What about Bootstrap?

When the estimator non-smooth,

- "n out of n bootstrap" would be inconsistent.
- "m out of n bootstrap" would be consistent with valid asymptotics as both $m$ and $n$ goes large. (Bickel, 2008)

  However, it will sacrifice convergence rate and would introduce a data-adaptive tunning parameter $m$ which might not be obvious. Its use for small sample is limited partly because performance is sensitive to $m$.

# References


1. Andrews, D. W. K. Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space. *Econometrica* **68,** 399–405. ISSN: 00129682, 14680262. http://www.jstor.org/stable/2999432 (2022) (2000).
2. Bickel, P. J. & Sakov, A. On the choice of m in the m out of n Bootstrap and Confidence Bounds for Extrema. *Statistica Sinica* **18,** 967–985. ISSN: 10170405, 19968507. http://www.jstor.org/stable/24308525 (2022) (2008).
3. Ertefaie, A. & Strawderman, R. L. Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika* **105,** 963–977 (Dec. 2018).
4. Goldberg, Y. & Kosorok, M. R. Q-Learning with Censored Data. *Annals of statistics* **40 1,** 529–560 (2012).
5. In, Ian, Ing, Uen & Heung. *Generalization Error Bounds of Dynamic Treatment Regimes in Penalized Regression-Based Learning.* in (2021).
6. Luckett, D. J. *et al.* Estimating Dynamic Treatment Regimes in Mobile Health Using V-Learning. *Journal of the American Statistical Association* **115,** 692–706 (2020).
7. Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* 1st. ISBN: 0471619779 (John Wiley & Sons, Inc., USA, 1994).
8. Shao, J. Bootstrap Sample Size in Nonregular Cases. *Proceedings of the IEEE* (1994).
9. Shi, C., Zhang, S., Lu, W. & Song, R. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **84,** 765–793 (2021).
10. Song, R., Wang, W., Zeng, D. & Kosorok, M. R. Penalized Q-Learning for Dynamic Treatment Regimens. *Statistica Sinica* **25,** 901–920 (July 2015).
11. Zhao, Y. Q. *et al.* Doubly Robust Learning for Estimating Individualized Treatment with Censored Data. *Biometrika* **102 1,** 151–168 (2015).
12. Zhao, Y.-Q., Zeng, D., Laber, E. B. & Kosorok, M. R. New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Journal of the American Statistical Association* **110,** 583–598 (2015).
13. Zhao, Y.-Q., Zeng, D., Rush, A. J. & Kosorok, M. R. Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association* **107,** 1106–1118 (2012).
14. Zhou, X., Mayer-Hamblett, N., Khan, U. & Kosorok, M. R. Residual Weighted Learning for Estimating Individualized Treatment Rules. *Journal of the American Statistical Association* **112,** 169–187 (2017).