QBS
Vol. 36, No. 1, pp. 7-13 (2017)
https://doi.org/10.22283/qbs.2017.36.1.7
pISSN 2288-1344 eISSN 2508-7185 **Review Article**

Quantitative Bio-Science
Official Journal of the Institute of Natural Science at Keimyung University
Available Online at https://qbs.kmu.ac.kr:442/

Individualized Treatment Regime for Personalized Medicine: A Review

Young-Geun Choi[1,*]

[1]Postdoctoral Fellow, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
(Received April 14, 2017; Revised May 15, 2017; Accepted May 17, 2017)

ABSTRACT

The heterogeneity of patients' responses in many treatment programs stems from the characteristics of their diverse background. Accordingly, paradigms of clinical decisions are shifting from "one-size-fits-all" rules to individualized treatment rules (ITRs). In statistical frameworks, an ITR is defined as a mapping of patient information to treatment recommendations, aiming to optimize the average future outcome of the program. In recent years, many researchers have developed methods for estimating optimal ITRs. This paper reviews some of the recent approaches, focusing on "indirect" ("*Q*-learning" and "*A*-learning") and "direct" ("*O*-learning") methods. We also briefly discuss further extensions of ITRs, which include multi-staged treatment rules also known as a dynamic treatment regime.

Key words : *A*-learning, *O*-learning, *Q*-learning, Individualized treatment rules, Personalized medicine

1. Introduction

It is widely recognized that even if the same medical decision applies, response patterns may be heterogeneous among the patients treated. For example, patients suffering from mental disorders (e.g. depression, drug abuse) reportedly showed different responses to given treatment [1,2]. To improve personalized healthcare, one should diversify the type of treatment or the dosage according to patients' information. One of the statistical efforts to operationalize personalized treatment is an individualized treatment rule (ITR) approach. An ITR is a decision rule that assigns each patient to treatment methods based on their information. The term "individual treatment regime" is also known as "personalized treatment regimes" [3] or "personalized treatment rules" [4]. An ITR is said to be optimal when it optimizes the mean of a desired clinical outcome (e.g. maximizing efficacy or minimizing side-effects) if applied.

There is growing interest in estimating optimal ITR and relevant rules in the statistical literature [5-27]. The aim of this paper is to provide a limited survey of those recent approaches to stimulate the reader's interest. In a broad sense, there are two approaches to estimate the optimal ITR, which depends on estimating the expected future outcome. In particular, we review so-called "*Q*-learning", "*A*-learning", and "*O*-learning" for the examples of indirect and direct methods. Although we will deal with single-staged decisions, the framework can extend to a situation involving multi-staged decision making over time also known as "dynamic treatment regimes" (DTRs) [17-27]. DTRs will also be discussed briefly.

Prior to developing the review, we clarify how the individual treatment rules are different from supervised learning (classification/regression). Typical interest in supervised learning is classifying symptoms or predicting risk from a patient's information, which involves datasets that can described by (X, Y), where X denotes covariates and Y is a response. In this

* Correspondence should be addressed to Dr. Young-Geun Choi, Postdoctoral Fellow, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Mail stop M2-B500, Seattle, WA 98109, USA. Tel: +1-206-667-6202, Fax: +1-206-667-7004, E-mail: ychoi2@fredhutch.org

case, a user aims to select such a model that minimizes the misclassification rate or mean-squared error. Prognosis and risk prediction may be suitably relevant medical problems. On the other hand, ITR may be relevant to decision support systems. The ITRs do not aim to predict future outcomes, but aim to find optimal decisions, which leads to the best future outcome, under the assumption that the treatment has a causal effect on the future outcome. Thus, datasets that are typically used in ITR can be described by (X, A, Y), where X and Y can be treated similarly as in the supervised learning and A denotes the treatment selected.

Because the notational setting in ITR is the same as that used in the causal inference literature, the difference between the causal inference and ITR should be noted. In causal inference, the interest is typically in inferring the difference of the potential outcomes between two choices of treatments from the treatment effect. The interested reader can refer to [28,29] for an overview of causal inference. In that field, recently there has been growing interest in the estimation of personalized treatment effect, the treatment effect conditioned on covariates [30-32]. Once the personalized treatment effect is estimated, it can be used to determine which treatment should be applied to an individual with a given covariate, i.e. an ITR. The approaches provided in this paper can be seen as shortcuts to estimate the optimal ITR without estimating the (personalized) treatment effects.

The rest of the paper is organized as follows. In Section 2, basic notation and assumptions are introduced as well as the properties of optimal ITR. In Section 3, the Q-, A-, and O-learning-based methods are introduced to estimate the optimal ITR. In Section 4, we discuss pros and cons when using indirect/direct methods and extension to DTR. In Section 5, we present concluding remarks and discuss some possible research directions.

2. Preliminaries

2.1 Notation and assumptions

The notational setup follows the Neyman's potential outcome framework. Let $\{(X_i, A_i, Y_i)\}_{i=1}^n$ be a random sample of (X, A, Y). Here, X, a p-dimensional random vector, denotes a subject's covariates. Note that it may include the intercept without loss of generality. Here $A = \pm 1$ is the treatment administered to the subject. Here, we assume that we can choose binary treatments for simplicity, but extensions to multiple treatments are straightforward. The clinical outcome is denoted by Y and assumed to be encoded as a higher value is preferred. Propensity of treatment assignments given covariates is defined by $\pi(a|x) = P(A=a|X=x)$. The expected clinical outcome given a covariate and treatment decision will be said to be the *quality* function and denoted by $Q(x, a) = E(Y|X=x, A=a)$. Our decision based on pre-treatment covariates is described by $d(x)$, a mapping of covariates to decisions that takes values ± 1. The description of Y needs additional care. We further introduce the notation $Y(-1)$ and $Y(1)$, the potential outcomes of a subject under treatment -1 and 1, respectively. Since each patient receives one treatment only, either one is unobservable. The observed outcome will be denoted by Y, i.e. $Y = Y(1)I(A=1) + Y(-1)I(A=-1)$, where $I(\cdot)$ denotes the indicator function. Like a standard causal inference problems, the inference of optimal ITR requires the following three assumptions to accommodate the systemically unobserved outcomes:

(A1) (strong ignorability) $\{Y(-1), Y(1)\} \perp A|X$;
(A2) (consistency) $Y = Y(A)$;
(A3) (positivity) there exist $c > 0$
such that $P(A = a|X = x) \geq c$ for almost sure x and a.

The first assumption says that observed data behave like a randomized clinical trial in each small region wherein the covariates are almost the same. The condition is also typically understood as that there are no unmeasured confounders. Randomized experiments satisfy (A1) by design. In an observational study, however, it can never be proved and should be assumed. Assumption (A2) ensures the outcomes observed in the study are what would be potentially seen under the received treatments. Assumption (A3) implies that we have both observations for the entire domain of the covariates so that the bias-correction information is available.

2.2 Optimal individualized treatment rule

We now assume that (X, A, Y) follows a distribution P. Let P^d be the distribution of (X, A, Y) when treatments are assigned according to d. What will P^d look like? Let us assume X and Y are either discrete or continuous. We can write P as $p(x, a, y) = f_X(x)\pi(a|x)f_{Y|X,A}(y|x, a)$. Then the distribution of P^d becomes $p^d(x, a, y) = f_X(x)I(a = d(x))f_{Y|X,A}(y|x, a)$. We assess the performance of d by its *value*, which we will define by $V(d) = E^d(Y)$, the expectation of outcome if patients

were treated by the decision d. An *optimal* decision d^* is defined by a decision rule satisfying $V(d^*) \geq V(d)$ for all $d \in D$ (a given class of decision functions).

We introduce several properties and equivalent representations for $V(d)$. First, from the positivity assumption, $I(a = d(x)) = \frac{I(a=d(x))}{\pi(a|x)}\pi(a|x)$ for the support of P and so $p^d(x,a,y) = \frac{I(a=d(x))}{\pi(a|x)}p(x,a,y)$. This gives

$$V(d) = E^d(Y) = E\left[\frac{Y}{\pi(A|X)} I(A = d(X))\right]. \quad (1)$$

This equation will be used in "direct" methods later. The value is also related to the quality; by the double expectation theorem,

$$\begin{aligned} V(d) &= E\left[E\left(\frac{Y}{\pi(A|X)} I(A = d(X))\middle| X, A\right)\right] \\ &= E\left[\frac{Q(X,A)}{\pi(A|X)} I(A = d(X))\right] \end{aligned}$$

and, with slight abuse of notation on the integral with discrete variable,

$$\begin{aligned} & E\left[\frac{Q(X,A)}{\pi(A|X)} I(A = d(X))\right] \quad (2) \\ &= \int \frac{Q(x,a)}{\pi(a|x)} I(a = d(x)) \cdot f_X(x)\pi(a|x)dadx \\ &= \int Q(x, d(x)) f_X(x)dx = E[Q(X, d(X))]. \end{aligned}$$

Equation (2) gives a valuable insight on how one can find an optimal treatment d^*. First, for any decision d, it holds that

$$E[Q(X, d(X))] \leq E\left[\max_a Q(X,a)\right].$$

On the other hand, considering a decision $d(x) = \arg\max_a Q(x,a)$ (the maximizer is not necessarily unique and one can pick any one maximizer), the definition of d^* tells us that

$$E[Q(X, d^*(X))] \geq E\left[\max_a Q(X,a)\right].$$

From those two inequalities, we observe that $d^*(X) \in \arg\max_a Q(X,a)$ almost surely.

How can we estimate the optimal decision d^*? First, note that $d^*(x) = \arg\min_a Q(x,a)$. This relation invoked researchers to estimate some part of or the whole $Q(x,a)$ and select such a decision maximizing the (estimated) outcome function. We refer to these series of methods as "indirect" methods.

On the other hand, one may focus on equation (1), which is equivalent to a weighted version of the 0-1 loss. Thus, maximizing (1) could lead to the optimal decision. Methods motivated by this idea will be referred to as "direct" methods.

3. Estimation of the Optimal Individualized Treatment Regime

3.1 Indirect methods

3.1.1 *Q*-learning

One natural approach to estimate d^* is to model and estimate the quality function $Q(X,A)$. These methods are also called "Q-learning" methods. Typically, the Q-learning methods were developed in estimating the optimal DTR. Qian and Murphy [5] inspected analytical bounds for the performance of estimated treatment rules when decision making is single stage (i.e. ITR). Furthermore, they proposed a l_1-penalized least square to approximate $Q(X,A)$ to Y.

The analytical details of Q-learning for estimating the optimal ITR are as follows. We explain the version given in [5]. First, one should model the space of quality functions, say $\mathcal{Q} = \{Q(x,a;\theta) = \Phi(x,a)^T\theta : \theta \in \mathbb{R}^p\}$, where $\Phi(x,a)$ is a vector-valued function and its components form a basis for $\mathcal{Q}$. Denote by $E_n f = \sum_{i=1}^n f(X_i, A_i, Y_i)/n$ an empirical expectation where f is a real-valued function. The authors assume that p can be possibly high and propose a least square with the l_1-penalty:

$$\hat{\theta} = \arg\min_\theta [E_n\{(Y - \Phi(X,A)^T\theta)^2\} + P_\lambda(\theta)], \quad (3)$$

where $P_\lambda(\theta)$ is a penalty function, $\Phi(X,A)_j$ and θ_j are the jth components of $\Phi(X,A)$ and θ, and λ is a tuning parameter. For the penalty function, Qian and Murphy [5] used an l_1-penalty that is weighted to balance the scale of the basis functions. After the fitting, one may estimate the optimal ITR using

$$\hat{d}(x) = \arg\max_a \Phi(X,a).$$

For readability, let us revisit an example without the penalty term as discussed in [5]. Let X be a scalar random variable following the uniform distribution on $[-1,1]$. Here A is generated evenly regardless of X, i.e. $\pi(a|x) = \frac{1}{2}$ for all a and x. Set the true quality function as $Q(x,a) = \left(x - \frac{1}{4}\right)^2 \cdot a$ and let $Y = Q(X,A) + \epsilon$, where ϵ follows the standard normal distribution. Because $d^*(X) \in \arg\max_a Q(X,a)$, clearly $d^*(X) = 1$ almost surely. The corresponding optimal value is $V(d^*) = 19/48$. What if we set a model space as $\mathcal{Q} = \{Q(x,a;\theta) = (1, x, a, xa)^T\theta : \theta \in \mathbb{R}^4\}$, which does not contain the true d^*? First, calculating (3) in the population-version mean without penalty yields $\hat{\theta} =$

$\left(0, 0, \frac{19}{48}, -\frac{1}{2}\right)$, which leads to $\hat{Q}(x,a)=\left(\frac{19}{48}-\frac{1}{2}x\right)a$. The estimated optimal decision becomes $\hat{d}(x)=sign\left(\frac{19}{48}-\frac{1}{2}x\right)$. Note that this is not equal to the true optimal rule $d^*=1$. However, one can see d^* is, in fact, included in the decision space: for a fixed θ, the estimated decision can be expressed as $\hat{d}(x)=\arg\max_a\{\theta_1+\theta_2 x+a(\theta_3+\theta_4 x)\}=sign(\theta_3+\theta_4 x)$. Then d^* can be identified as the case when $\theta_3=1$ and $\theta_4=0$. Thus, although the associated decision space contains the true decision, the misspecification of Q could lead to an inaccurate estimation of the decision. Generally speaking, when method (3) is used, it is safer to set a large class of model space. The l_1 penalty in (3) can accommodate the high-dimensionality coming from the model size (p), encouraging ease and parsimony of interpretation.

3.1.2 *A*-learning

"*A*-learning" (advantage learning, named in [6]) refers to methods that exploit the fact that one need not specify the entire Q-function to estimate the optimal treatment rule. To see this fact, we first note that the Q-function $Q(x,a)=E(Y|X=x, A=a)$ can be characterized by $m(x)+a\times c(x)$ for real-valued functions $m(x)$ and $c(x)$. Indeed, $m(x)=\frac{Q(x,1)+Q(x,-1)}{2}$ and $c(x)=\frac{Q(x,1)-Q(x,-1)}{2}$. Then the optimal decision $d^*(x)=\arg\min_a Q(x,a)$ can be restated as $d^*(x)=sign\{c(x)\}$. Thus, it suffices to estimate $c(x)$, not the whole $Q(x,a)$, for decision making.

For the estimation of $c(x)$, it is common to use an estimating equation as proposed in [7]. See [6,8-10] for details. We present Lu et al.'s [10] reformulation that recasts the estimating equation as a squared loss minimization problem that can be easily tailored to a certain regularization when the number of covariates are large. Specifically, let us consider the linear case, $c(x)=x^T\beta$. Then the estimated β is estimated from

$$\hat{\beta}=\arg\min_{\beta}\left[E_n\{\left(Y-h(X)-\beta^T X(A-\pi(1|X)\right)^2\}+P_\lambda(\beta)\right], \quad (4)$$

where $h(x)$ is an arbitrary function of covariates and we recall $\pi(1|x)=P(A=1|X=x)$ and $P_\lambda(\beta)$ is a penalty function on β. For the choice of $h(x)$, Lu et al. [10] refer to a constant model $h(x)=\gamma$ or a linear model $h(X)=\gamma^T X$.

3.2 Direct methods (*O*-learning)

As seen above, the indirect methods model a part or the whole body of the quality function Q to estimate the optimal treatment rule and potential error can occur from the misspecification of the model. Instead, one may consider directly estimating the rule maximizing the response as formulated in (1). Maximizing (the empirical version of) (1) is discontinuous and nonconvex with respect to d, which can be computationally intensive.

The "*O*-learning" (outcome weighted learning) first proposed by Zhao et al. [11] maximizes a convex surrogate loss of (1) so that the problem becomes tractable and can be solved efficiently. First, since $E\left(\frac{Y}{\pi(A|X)}\right)$ is constant, one notes that (1) is equivalent to minimizing

$$E\left[\frac{Y}{\pi(A|X)}I\left(A\neq d(X)\right)\right]$$

over d. Note that any d can be represented by $d(x)=sign\left(f(x)\right)$. Here, we define $sign(0)=1$, although $sign(0)=-1$ also gives valid arguments. Then, one sees $I\left(A\neq d(X)\right)=I(Af(X)<0)$. In the machine learning literature, the computational difficulty of the 0-1 loss function $I(Af(X)<0)$ was alleviated by a convex surrogate loss function $\phi(t)$ replaced with the 0-1 loss $I(t<0)$. The most popular choice is the hinge loss function, $\phi(t)=(1-t)_+=\max(1-t,0)$. Furthermore, one might want to penalize f to avoid overfitting. As a result, Zhao et al. [11] proposed to minimize

$$E_n\left[\frac{Y}{\pi(A|X)}\left(1-Af(X)\right)_+\right]+\lambda\|f\|^2, \quad (5)$$

where $\|\cdot\|$ is some norm on a function space and λ is a tuning parameter. Equation (5) can be a support vector machine (SVM) problem with weights and one can apply the theory of SVMs. For the class of f, a typical example is a space of linear classifiers; letting $f(X)=x^T\beta$ and the norm as in l_2-sense, one solves

$$E_n\left[\frac{Y}{\pi(A|X)}(1-AX^T\beta)_+\right]+\lambda\|\beta\|_2^2.$$

In addition, one can consider a nonlinear decision rule, for example, Zhao et al. [11] considered a so-called kernel trick. This stream of methods is called "outcome weighted learning" or simply "*O*-learning."

One notes that (5) becomes convex only when Y is positive. What if the outcome has a negative value? A natural strategy is a constant shift, i.e. coding the outcome as $Y+c$ instead of Y for some constant c. It is easy to see that the constant shift does not change the optimal decision defined from (4). However, it changes in a finite sample, i.e. the constant shift will lead to a differently estimated rule. Zhou et al. [12] proposed the use of a residual $Y-g(X)$ as the outcome weight in (4)

instead of Y, for some function g. Here, the authors proposed to pick such g reducing the variance for estimating (4). Their choice of g made the problem being invariant under constant shift of the outcome even in a finite sample.

Several extensions of the O-learning exist in the literature. Song et al. [13] considered sparse linear decision rules for high-dimensional covariates, by replacing the l_1 or smoothly clipped absolute deviation (SCAD) penalty with the squared penalty in (5). Zhao et al. [14] extended the O-learning to censored data. Laber and Zhao [15] and Zhu et al. [16] minimized (5) by considering tree-based rules instead of convex surrogate losses.

4. Discussion

4.1 Tuning parameter selection

One notes that all the presented methods, (3), (4), and (5), involve a tuning parameter λ. A natural question is how to select λ from given data in practical implementation. If one considers cross-validation (CV), it is common to minimize the predictive loss function (averaged over folds) in standard regression literature. In ITR, λ can be selected in a different way: recall that the goal of ITR is maximizing the value rather than minimizing predictive loss. For example, Qian and Murphy [5] choose λ to maximize the predictive value, which can be estimated in an unbiased manner by

$$\frac{E_n\left[\frac{Y}{\pi(A|X)} I\left(A=d(X)\right)\right]}{E_n\left[\frac{1}{\pi(A|X)} I\left(A=d(X)\right)\right]}$$

where the ITR d is fitted from a training dataset and a fixed tuning parameter and the empirical average is taken over a testing dataset.

4.2 Indirect or direct?

A possible advantage of indirect methods is that we can use standard statistical tools and packages, such as linear models, to model the outcome. There are well-developed disseminated methods such as the goodness-of-fit checking. If one needs interpretability that scientific theory or expert opinion usually requires, indirect methods would be useful. Instead, as shown in the example in Section 3.1, the outcome models should be correctly specified for valid inference, otherwise there would be a bias on the estimation. Direct methods are superior to the indirect methods in that they do not require models for the quality function or other traits on the outcome. Their potential drawback is that the coefficients in estimated treatment rules generally show larger variance than those from the indirect methods [18]. Indeed, in the computer science literature (reinforcement learning), many studies use both direct and indirect methods to construct optimal policies. The interested reader may refer to the textbook on reinforcement learning [33] for examples.

4.3 Dynamic (multi-staged) treatment regimes

The statistical framework for ITR can be extended to DTR, a sequence of personalized decision rules in response to a subject's changing needs. We provide a brief review where there are two stages of decisions with binary treatment, which is sufficient to provide a comprehensive intuition. Let $(X_1, A_1, X_2, A_2, X_3)$ be a time-ordered trajectory of each patient's information on covariates (X_1, X_2, and X_3), decisions (A_1 and A_2). The available information for the decision maker is $H_1=X_1$ at the first time and $H_2=(X_1, A_1, X_2)$ at the second time. As for the outcome, let us say that there is one outcome at the end of the study, which we will denote by Y, although a more general framework allows the presence of an outcome (like the quality function in ITR) after each of the two decisions. A DTR is defined by a pair of decisions, $d=(d_1, d_2)$, where d_t is a mapping from the domain of H_t to $\{-1, +1\}$, $t=1, 2$. The value of a DTR d is then defined by the expected outcome when d is applied to the given population. An optimal DTR is similarly defined in a recursive way in which each formulation resembles (4), which is omitted for simplicity. In the two-staged setting, the quality function is defined recursively by

$$Q_2(h_2, a_2)=E(Y|H_2=h_2, A_2=a_2),$$
$$Q_1(h_1, a_1)=E(\max_{a_2} Q_2(h_2, a_2)|H_1=h_1, A_1=a_1),$$

and it holds that the optimal DTR satisfies $d_t^*(h_t)=\arg\max_{a_t} Q_t(h_t, a_t)$, $t=1, 2$. There is a vast literature to infer the optimal DTR. Estimating the optimal DTR from Q-learning and A-learning was proposed in [9,18-20,23,25] and [6,8,9], respectively. The O-learning methods inferring DTR were developed in [21,25,27]. Here, we provide a brief introduction to Q-learning methods. Consider linearly parameterized models on both quality functions, say $Q_1(h_1, a_1; \theta_1)$ and $Q_2(h_2, a_2; \theta_2)$ as in the single-staged Q-learning. One of the Q-learning methods solves the following dynamic programming problem in a

backward sense.

Step 1. Second stage regression: let $\hat{\theta}_2$ be the minimizer of $E_n\left(Y - Q_2(H_2, A_2; \theta_2)\right)^2$.

Step 2. Second stage outcome is predicted by $\tilde{Y} = \max_{a_2} Q_2\left(H_2, a_2; \hat{\theta}_2\right)$.

Step 3. First stage regression: let $\hat{\theta}_1$ be the minimizer of $E_n\left(\tilde{Y} - Q_1(H_1, A_1; \theta_1)\right)^2$.

Step 4. Estimate DTR: $\hat{d}_t(h_t) = \arg\max_{a_t} Q_t(h_t, a_t; \hat{\theta}_t)$, t = 1, 2.

5. Concluding Remarks

We have provided a short survey on recent methodologies estimating the optimal ITR. It turned out that the optimal ITR can be expressed not only as the maximizer of the quality function with respect to possible decisions, but also as the minimizer of outcome-weighted 0-1 loss. The former formulation led to indirect methods including *A*-learning and *Q*-learning, which model the quality function and can leverage standard statistical theories and packages. On the other hand, the direct methods including *O*-learning aim to minimize an estimator of the weighted 0-1 loss and are advantageous in that they require fewer model assumptions on the quality function than the direct methods.

We conclude the paper with some open problems and possible future directions for further research. The first direction is hypothesis testing on estimated ITRs. Even though an ITR is estimated, one might be interested in the significance of the model or some of the variables. When the number of covariates is small, traditional statistics theory may be applied for the desired testing. For high-dimensional covariates, hypothesis testing frameworks have been developed in very recent years; their application to ITRs appears to be a promising future approach. The second possible direction is to leverage more machine learning approaches to the *O*-learning methods. Currently, SVMs or regression trees have been imported to accommodate the weighted 0-1 loss. It remains open to the application of random forests or deep neural networks, which have had great success in machine learning. Finally, for real-world applications, studies dealing with missing data and measurement errors should also be developed.

References

1. McLellan AT. Have we evaluated addiction treatment correctly? Implications from a chronic care perspective. Addiction 2002;97:249-252.
2. Fava M, Rush AJ, Trivedi MH, Nierenberg AA, Thase ME, Sackeim HA, et al. Background and rationale for the sequenced treatment alternatives to relieve depression (STAR*D) study. Psychiatr Clin North Am 2003;26:457-494.
3. Wang Y, Powell W. An optimal learning method for developing personalized treatment regimes. arXiv Prepr 2016;arxiv:1607.01462.
4. Wang Y, Fu H, Zeng D. Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies. J Am Stat Assoc 2017;to appear.
5. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Ann Stat 2011;39:1180-1210.
6. Blatt D, Murphy SA, Zhu J. *A*-learning for approximate planning. Technical Report 04-63, The Methodology Center, Pennsylvania State University, State College, PA; 2012.
7. Robins JM. Optimal structural nested models for optimal sequential decisions. In: Lin D, Heagerty, PJ, editors. Proceedings of the second Seattle symposium in biostatistics. New York: Springer; 2004.
8. Moodie EEM, Richardson TS. Estimating optimal dynamic regimes: correcting bias under the null. Scand J Stat 2010;37:126-146.
9. Schulte PJ, Tsiatis AA, Laber E, Davidian M. *Q*- and *A*-learning methods for estimating optimal dynamic treatment regimes. Stat Sci 2014;29:640-661.
10. Lu W, Zhang HH, Zeng D. Variable selection for optimal treatment decision. Stat Methods Med Res 2013;22:493-504.
11. Zhao Y, Zeng D, Rush AJ, Kosorok, MR. Estimating individualized treatment rules using outcome weighted learning. J Am Stat Assoc 2012;107:1106-1118.
12. Zhou X, Mayer-Hamblett N, Khan U, Kosorok MR. Residual weighted learning for estimating individualized treatment rules. J Am Stat Assoc 2017;to appear.
13. Song R, Kosorok MR, Zeng D, Zhao Y, Laber E, Yuan M. On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. Statistics 2015;4:59-68.
14. Zhao Y, Zeng D, Laber E, Song R, Yuan M, Kosorok MR. Doubly robust learning for estimating individualized treatment with censored data. Biometrika 2015;102:151-168.
15. Laber E, Zhao Y. Tree-based methods for individualized treatment regimes. Biometrika 2015;102:501-514.
16. Zhu R, Zhao Y, Chen G, Ma S, Zhao H. Greedy outcome weighted tree learning of optimal personalized treatment rules.

Biometrics 2017;to appear.
17. Zhang B, Tsiatis AA, Laber E, Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika 2013;100:681-694.
18. Laber E, Lizotte, DJ, Qian M, Pelham WE, Murphy SA. Dynamic treatment regimes: technical challenges and applications. Electron J Stat 2014;8:1225-1272.
19. Chakraborty B, Murphy SA. Dynamic treatment regimes. Ann Rev Stat Appl 2014;1:447-464.
20. Huang X, Choi S, Wang L, Thall PF. Optimization of multistage dynamic treatment regimes utilizing accumulated data. Stat Med 2015;34:3424-3443.
21. Zhao Y, Zeng D, Laber E, Kosorok MR. New statistical learning methods for estimating optimal dynamic treatment regimes. J Am Stat Assoc 2015;110:583-598.
22. Chakraborty B, Laber E, Zhao Y. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. Biometrics 2013;69:714-723.
23. Zhang Y, Laber E, Tsiatis AA, Davidian M, Carolina, N. Interpretable dynamic treatment regimes. arXiv Prepr 2016; arXiv:1606.01472.
24. Shi C, Fan A, Song R, Lu W. High-dimensional *A*-learning for dynamic treatment regimes. Ann Stat 2017;to appear.
25. Moodie EEM, Chakraborty B, Kramer MS. *Q*-learning for estimating optimal dynamic treatment rules from observational data. Can J Stat 2012;40:629-645.
26. Zhao YQ, Laber EB. Estimation of optimal dynamic treatment regimes. Clin Trials 2014;11:400-407.
27. Liu Y, Wang Y, Kosorok MR, Zhao Y, Zeng D. Robust hybrid learning for estimating personalized dynamic treatment regimens, arXiv Prepr 2016;arxiv:1611.02314.
28. Rubin DB. Causal inference using potential outcomes. J Am Stat Assoc 2005;100:322-331.
29. Pearl, J. Causal inference in statistics: an overview. Statist Surv 2009;3:96-146.
30. Belloni A, Chernozhukov V, Fernandez-Val I, Hansen C. Program evaluation and causal inference with high-dimensional data. Econometrica 2017;85:233-298.
31. Farrell MH. Robust inference on average treatment effects with possibly more covariates than observations. J Econometrics 2015;189:1-23.
32. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. arXiv Prepr 2016;arXiv: 1510.04342.
33. Wiering M, van Otterlo M. Reinforcement learning: state-of-the-art, volume 12. New York: Springer; 2012.