ANNUAL REVIEWS

*Annual Review of Statistics and Its Application*

# Precision Medicine

Michael R. Kosorok[1] and Eric B. Laber[2]

[1]Department of Biostatistics and Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA; email: kosorok@unc.edu

[2]Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, USA; email: eblaber@ncsu.edu



Annu. Rev. Stat. Appl. 2019. 6:263–86

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-030718-105251



## Keywords

data-driven decision science, dynamic treatment regimes, machine learning, patient heterogeneity, statistical inference

## Abstract

Precision medicine seeks to maximize the quality of health care by individualizing the health-care process to the uniquely evolving health status of each patient. This endeavor spans a broad range of scientific areas including drug discovery, genetics/genomics, health communication, and causal inference, all in support of evidence-based, i.e., data-driven, decision making. Precision medicine is formalized as a treatment regime that comprises a sequence of decision rules, one per decision point, which map up-to-date patient information to a recommended action. The potential actions could be the selection of which drug to use, the selection of dose, the timing of administration, the recommendation of a specific diet or exercise, or other aspects of treatment or care. Statistics research in precision medicine is broadly focused on methodological development for estimation of and inference for treatment regimes that maximize some cumulative clinical outcome. In this review, we provide an overview of this vibrant area of research and present important and emerging challenges.

## 1. INTRODUCTION

The idea of improving health outcomes by tailoring treatment to individual patient characteristics is centuries old and remains a core component of medical practice. The scientific method began to impact medical treatment through the use of statistical inference by the late 1700s, but advances began to dramatically increase after the success of the first randomized controlled clinical trial, conducted by Austin Bradford Hill in 1946, which demonstrated the efficacy of streptomycin for treating tuberculosis (Stusser 2006). Following Hill's trial was a period of rapid methodological progress in the design and analysis of clinical trials as well as observational studies. Systematic study of the integration of data, experience, and clinical judgment into the clinical decision process led to the concept of evidence-based medicine, wherein clinical decision making is based on (to the extent possible) empirical evidence with randomized controlled trials being a gold standard for generating such evidence (Eddy 1990). However, the primary scientific aim in most clinical trials is the identification of the best treatment for a given disease area, with any heterogeneity in patient characteristics or outcomes being viewed as a nuisance to the research process.

Awareness that patient heterogeneity was important in evaluating treatments—and not just a nuisance—began to emerge late in the twentieth century among both clinicians (Sorensen 1996) and biostatisticians (Longford & Nelder 1999). That patient heterogeneity implied the need to individualize therapy in the context of evidence-based medicine was nicely articulated in Kravitz et al. (2004). These constituent concepts, combined together, yield the modern concept of precision medicine, the paradigm wherein patient heterogeneity is leveraged through data-driven approaches to improve treatment decisions so that the right treatment is given to the right patient at the right time. Precision medicine became a national priority with President Obama's announcement of the Precision Medicine Initiative in his 2015 State of the Union Address (White House Office of the Press Secretary 2015). We note that precision medicine is conceptually the same as stratified medicine (Lonergan et al. 2017) and personalized medicine (Kosorok & Moodie 2016).

The goal of this article is to provide a review of the current state of the art in statistical research for precision medicine. The chief priority of statistical research in precision medicine is to use data to inform decision making in health care; thus, precision medicine encompasses a wide range of tasks including drug discovery, biomarker identification, estimation and inference for causal treatment effects, modeling health communication and shared decision making, and study design. However, our focus in this review is on estimation and inference for treatment regimes that prescribe interventions based on individual patient characteristics. An estimated optimal treatment regime might be used as part of a decision support system within a health care organization or to generate new clinical hypotheses for future study. Thus, it is critical that statistical methodology for precision medicine be rigorous, transparent, reproducible, and generalizable.

We view precision medicine as fitting within the broader concepts of precision public health and data-driven decision science. However, the focus on data-driven, patient-centered care with its inherent challenges (e.g., patient heterogeneity, implementation cost, causal confounding) distinguishes precision medicine as its own field of study. To this point, precision medicine has led to new methodologies and insights in semiparametric modeling, causal inference, nonregular asymptotics, clinical trial design, and machine learning (see Murphy 2003, Robins 2004, Chakraborty & Moodie 2013, Laber et al. 2014c, Zhao et al. 2015a, Kosorok & Moodie 2016, and references therein). There have also been major advancements in genetics driven by the vision for precision medicine (see, e.g., Perou et al. 2000, Liang et al. 2018, Torkamani et al. 2018); however, our focus is broader, in that while the biomarkers used to inform treatment selection could be genetic or genomic factors, we also allow that treatment selections could be based on demographic and physiological measurements, comorbid conditions, individual patient preferences, lifestyle, and so on.

The remainder of the article is as follows. In Section 2, we formalize the goals of precision medicine and catalog different decision settings. In Section 3, we discuss biomarker discovery as an important supporting task in constructing an optimal treatment regime. In Section 4, we review the methodological underpinnings of precision medicine, including regression-based and direct-search estimation. In Sections 5 and 6, we discuss, respectively, managing multiple outcomes and inference for estimated optimal regimes. We conclude with a summary and discussion of pressing open problems.

## 2. GOALS OF PRECISION MEDICINE

In this section, we formalize an optimal dynamic treatment regime and contrast estimation of an optimal regime with subgroup identification and causal effect estimation. We also discuss generalizations of the dynamic treatment regime framework and how decision problems in precision medicine fit into precision public health and data-driven decision science. We view decision support—especially the estimation of optimal or near-optimal treatment regimes and those endeavors that directly support this—to be the primary goal of precision medicine. Accordingly, we view modeling the disease process per se to be of secondary importance in precision medicine, except when it is directly supportive of dynamic treatment regime discovery.

### 2.1. Discovering Dynamic Treatment Regimes

As stated previously, the goal in precision medicine is to use data to improve decision making in health care. Dynamic treatment regimes formalize decision making as a sequence of decision rules, one per decision point, that map available information to a recommended intervention. The decision points may be either fixed in calendar time or driven by patient outcomes. Thus, the timing and number of decision points may be random and can vary considerably across patients in some application domains. Furthermore, the set of allowable interventions at any given time point may vary according to a patient's health status, availability, or other factors (Bembom & van der Laan 2008, Schulte et al. 2014, Laber & Staicu 2018, Laber et al. 2018b); however, for simplicity, in describing methods for estimation and inference in Section 4, we will not do so in complete generality. We formalize the notion of an optimal treatment regime using the language of potential outcomes (Rubin 1978, 2005; Dawid 2015).

**2.1.1. The single-decision setting.** In the single-decision setting the observed data are assumed to be of the form $\{(X_i, A_i, Y_i)\}_{i=1}^n$, which comprise $n$ independent and identically distributed (i.i.d.) triples $(X, A, Y)$ where $X \in \mathcal{X}$ denotes baseline patient characteristics, $A \in \mathcal{A}$ denotes the assigned treatment, and $Y \in \mathbb{R}$ denotes the outcome coded so that higher values are better. For each $x \in \mathcal{X}$, define $\psi(x) \subseteq \mathcal{A}$ to be the set of allowable treatments for a patient presenting with $X = x$. A dynamic treatment regime in this context is a map $d : \mathcal{X} \to \mathcal{A}$ which satisfies $d(x) \in \psi(x)$ for all $x \in \mathcal{X}$; under $d$, patients presenting with $X = x$ would be assigned treatment $d(x)$. An optimal treatment regime yields the maximal mean outcome if applied to select treatments in the population of interest. Let $Y^*(a)$ denote the potential outcome under treatment $a \in \mathcal{A}$, and subsequently, for any regime $d$, define the potential outcome under $d$ to be $Y^*(d) = \sum_{a \in \mathcal{A}} Y^*(a) 1_{\{d(X)=a\}}$, where $1_B$ is the indicator of $B$. The optimal regime, say $d^{\text{opt}}$, satisfies $(a)$ $d^{\text{opt}}(x) \in \psi(x)$ for all $x \in \mathcal{X}$ and $(b)$ $\mathrm{E}Y^*(d^{\text{opt}}) \geq \mathrm{E}Y^*(d)$ for all $d$ such that $d(x) \in \psi(x)$ and all $x \in \mathcal{X}$. There are other notions of optimality, including maximizing quantiles of the outcome distribution (Linn et al. 2017, Wang et al. 2018b), maximizing efficacy subject to constraints on risk/harm (Linn et al. 2015, Wang et al. 2018c, Laber et al. 2018b), and maximizing the mean outcome subject to cost or logistical constraints (Luedtke & van der Laan 2016a, Lakkaraju & Rudin 2017). For an overview of the potential outcome framework in causal inference, see, e.g., Pearl (2009).

The optimal dynamic treatment regime, $d^{\text{opt}}$, is defined in terms of potential outcomes; however, in order to construct an estimator of $d^{\text{opt}}$, we need to identify it in terms of the data-generating model. In Section 4, we present causal conditions under which such identifiability holds. We note that the term dynamic refers to the individualization of treatment to patient characteristics, not time; hence, the term dynamic treatment regime is used even with a single decision point. One could also assume that the clinicians assigning treatment are already making nearly optimal choices, and thus a potentially good dynamic treatment is one obtained by mimicking the clinicians' decisions (Wallace et al. 2018). We consider this in Section 5 for the special case where physicians make the optimal decision at least for a small portion of the time.

**2.1.2. The multi-decision setting.** In the multi-decision setting there are two or more opportunities for treatment change for some subset of the population. Examples include the treatment of small-cell lung cancer where two or more lines of chemotherapy may be required (Zhao et al. 2011). A key concept in the multi-decision setting is that interventions can affect a patient's health status in multiple ways, including immediate effects (it may make an immediate impact on their health status), moderating effects (it may generate information that is useful for subsequent decisions, e.g., failure to respond to a drug in a given class may indicate that other drugs belonging to the same class may also perform poorly), and delayed effects (it may change the patient's health status so as to set them up for future success, e.g., providing a patient with cognitive behavioral therapy at one stage may allow them to reap greater benefits from teletherapy at later time points). Thus, applying a treatment that leads to a suboptimal proximal effect may lead to better long-term outcomes if there are strong delayed or prognostic effects (Thall et al. 2007, Kidwell 2016).

In the multi-decision setting, we assume that the observed data are of the form $\{(X_{1,i}, A_{1,i}, Y_{1,i}, \ldots, X_{T,i}, A_{T,i}, Y_{T,i})\}_{i=1}^{n}$, which comprise $n$ i.i.d. replicates of $(X_1, A_1, Y_1, \ldots, X_T, A_T, Y_T)$ where $X_1 \in \mathcal{X}_1$ denotes baseline information and $X_t \in \mathcal{X}_t$ denotes interim information collected during the course of the stage $t-1$ treatment for $t = 2, \ldots, T$; $A_t \in \mathcal{A}_t$ denotes the assigned treatment; and $Y_t$ denotes a proximal outcome measured after the treatment at stage $t$ for $t = 1, \ldots, T$. Define $H_1 = X_1$ and $H_t = (H_{t-1}, A_{t-1}, Y_{t-1}, X_t)$ so that $H_t$ is the available patient history at time $t$. Let $\mathcal{H}_t = \operatorname{dom} H_t$ and let $\psi_t(h_t) \subseteq \mathcal{A}_t$ denote the set of allowable treatments for a patient presenting with $H_t = h_t$ at time $t$. A dynamic treatment regime in this setting is a sequence of functions $d = (d_1, \ldots, d_T)$ such that $d_t : \mathcal{H}_t \to \mathcal{A}_t$ satisfies $d_t(h_t) \in \psi_t(h_t)$ for all $h_t$ for $t = 1, \ldots, T$. An optimal treatment regime maximizes the expectation of some (prespecified) cumulative outcome measure $Y = y(Y_1, \ldots, Y_T)$, e.g., $y(v_1, \ldots, v_T) = \sum_{t=1}^{T} v_t$, or $y(v_1, \ldots, v_T) = \max_t v_t$, or $y(v_1, \ldots, v_T) = v_T$.

Interventions applied at time $t$ affect both the proximal outcomes $Y_t$ and the interim measurements $X_{t+1}$; thus, to define the optimal treatment regime, we need to consider both potential proximal outcomes and potential interim measurements. We use an overline to denote history so that $\overline{a}_t = (a_1, \ldots, a_t)$. Define $X_t^*(\overline{a}_{t-1})$ to be the potential interim measurements at time $t$ under treatment sequence $\overline{a}_{t-1} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_{t-1}$, where products of sets are interpreted as Cartesian products, and define $Y_t^*(\overline{a}_t)$ to be the potential proximal outcome under $\overline{a}_t \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_t$. The potential outcome under $\overline{a}_T \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_T$ is therefore $Y^*(\overline{a}_T) = y\{Y_1^*(a_1), Y_2^*(\overline{a}_2), \ldots, Y_T^*(\overline{a}_T)\}$. For any regime $d$, the potential outcome is

$$Y^*(d) = \sum_{\overline{a}_T} Y^*(\overline{a}_T) \prod_{t=1}^{T} 1_{[d_t\{H_t^*(\overline{a}_{t-1})\}=a_t]},$$

where we have defined $H_1^*(a_0) \equiv H_1$. An optimal regime, $d^{\text{opt}}$, satisfies the following: first, $d_t^{\text{opt}}(h_t) \in \psi_t(h_t)$ for all $h_t \in \mathcal{H}_t$ and $t = 1, \ldots, T$; and second, $\mathrm{E}Y^*(d^{\text{opt}}) \geq \mathrm{E}Y^*(d)$ for all $d$ satisfying $d_t(h_t) \in \psi_t(h_t)$ for all $h_t \in \mathcal{H}_t$ and $t = 1, \ldots, T$.

As in the single-decision setting, the optimal regime is defined in terms of potential outcomes and is only identifiable in terms of the data-generating model under additional assumptions; we discuss such assumptions in Section 4. The preceding development assumes that future patients will be treated over the same time horizon as the patients in the sample; however, this need not be the case. In some settings, e.g., diabetes (Ertefaie 2015, Luckett et al. 2017) or cystic fibrosis (Tang & Kosorok 2012), interventions are applied over an indefinite time horizon. This structure would also be applicable to precision screening for cancer (Olsen & Lund 2017), among other applications. Thus, the objective in these settings is often to estimate a treatment regime that can potentially be applied beyond the time horizon over which the training data are collected. Such extrapolation typically requires additional structure to be imposed on the data-generating model; a common assumption is that the decision process is (perhaps after some suitable transformation) heterogeneous and Markov (Puterman 2005). We discuss this further in Section 4.



**2.1.3. Other decision settings.** The decision settings described above assume the following: (*a*) the observed data consist of i.i.d. replicates, (*b*) the data-generating model is fixed and indifferent to the actions of the decision maker, and (*c*) the observed data are used to construct treatment regimes for application with yet unobserved future patients. However, these assumptions may need to be relaxed in some application domains. For example, in the context of managing the spread of an emerging infectious disease over a network of individuals, spatial dependence and spillover effects preclude treating the individuals as independent replicates; furthermore, in this setting, one must manage the disease in real-time (Laber et al. 2018a). Thus, neither *a* nor *c* hold in this setting. In the context of adversarial decision making wherein one faces an intelligent (and adaptive) opponent, the data-generating model can change in response to the decision makers' actions, e.g., imagine playing poker repeatedly against the same shrewd player (Cesa-Bianchi & Lugosi 2006).

There are potentially other decision settings not covered by the above structures; however, the methods and ideas presented in this review are readily extensible to new domains. There are also additional complications associated with estimation of optimal treatment regimes that occur more generally in statistical modeling, and thus we will not discuss them here. Examples include missing data (Shortreed et al. 2011, 2014; Shortreed & Moodie 2012, Kosorok & Moodie 2016), measurement error (Shani et al. 2013), and model-building (Biernot & Moodie 2010, Rich et al. 2010, Henderson et al. 2010, Gunter et al. 2011, Lu et al. 2013, Laber et al. 2014a, Song et al. 2015a, Luedtke & van der Laan 2016b).

## 2.2. Treatment Effect Estimation and Subgroup Identification

Estimation of an optimal treatment regime is closely related to subgroup identification and estimation of the conditional average treatment effect (CATE). For simplicity, we use a single-decision problem with binary treatments, so that the data for a generic subject are $(X, A, Y)$, where $X \in \mathcal{X} \subseteq \mathbb{R}^p$ denotes baseline patient information, $A \in \mathcal{A} = \{-1, 1\}$ denotes the assigned treatment, and $Y \in \mathbb{R}$ denotes the outcome coded so that higher values are better. We assume that $\psi(x) \equiv \mathcal{A}$ for all $x \in \mathcal{X}$. In this setup, the CATE is defined as $\Delta(x) = E\{Y^*(1) - Y^*(-1)|X = x\}$. For any regime $d$, it can be seen that

$$\begin{aligned}
\mathrm{E}Y^*(d) &= \mathrm{E}[Y^*(1)1_{\{d(X)=1\}} + Y^*(-1)1_{\{d(X)=-1\}}] \\
&= \mathrm{E}\{Y^*(1) - Y^*(-1)\}1_{\{d(X)=1\}} + \mathrm{E}Y^*(-1) \\
&= \mathrm{E}\Delta(X)1_{\{d(X)=1\}} + \mathrm{E}Y^*(-1) \\
&\leq \mathrm{E}\Delta(X)1_{\{\Delta(X)>0\}} + \mathrm{E}Y^*(-1) \\
&= \mathrm{E}[Y^*(1)1_{\{\Delta(X)>0\}} + Y^*(-1)1_{\{\Delta(X)\leq 0\}}],
\end{aligned}$$

from which it can be seen that $d^{\text{opt}}(x) = \text{sign}\{\Delta(x)\}$ is optimal, where $\text{sign}(x) = \pm 1$ according to whether $x > 0$ or $< 0$. Thus, a natural approach to estimating an optimal treatment regime in this context is to first construct an estimator $\widehat{\Delta}_n(x)$ of $\Delta(x)$ and subsequently to use the plug-in estimator $\widehat{d}_n(x) = \text{sign}\{\widehat{\Delta}_n(x)\}$ of $d^{\text{opt}}(x)$; we discuss this approach in more detail in Section 4. We note this can be developed for more than two treatments, but we focus here on two treatments for simplicity.

Subgroup identification seeks to find a subgroup in the target population with an enhanced treatment effect; sometimes this is referred to as finding the right patient for the right treatment. Given a threshold $\eta \in \mathbb{R}$, one way to operationalize a subgroup is through the level set $\mathcal{T}(\eta) = \{x \in \mathcal{X} : \Delta(x) \geq \eta\}$. Thus, one could estimate this level set using the plug-in estimator $\widehat{\mathcal{T}}_n(\eta) = \{x \in \mathcal{X} : \widehat{\Delta}_n(x) \geq \eta\}$, where $\widehat{\Delta}_n(x)$ is an estimator of $\Delta(x)$. We note that while much of clinical research focuses on estimating the CATE, precision medicine seeks to go further by tailoring treatments to subgroups (as reflected in $x$) to benefit each subgroup and thereby further benefit the population on average.

To illustrate the differences between estimation of an optimal treatment regime, estimation of the CATE, and subgroup identification, consider the following generative model: $X \sim \text{Normal}(0, \tau^2)$, $A \perp X$ and $A \sim \text{Uniform}\{-1, 1\}$, $\epsilon \perp (X, A)$, $\epsilon \sim \text{Normal}(0, \sigma^2)$, and $Y = g(X, A) + \epsilon$, where $g(X, A) = \exp(\alpha_0^* + \alpha_1^* A + \alpha_2^* X + \alpha_3^* AX)$ and $\perp$ denotes independence. The optimal treatment can be seen to equal $d^{\text{opt}}(x) = \text{sign}\{\alpha_1^* + \alpha_3^* X\}$, the CATE is given by $g(x, 1) - g(x, -1) = \exp\{\alpha_0^* + \alpha_1^* + (\alpha_2^* + \alpha_3^*)x\} - \exp\{\alpha_0^* - \alpha_1^* + (\alpha_2^* - \alpha_3^*)x\}$, and the subgroup corresponding to the level set with threshold $\eta$ is given by $\mathcal{T}(\eta) = \{x \in \mathbb{R} : g(x, 1) - g(x, -1) \geq \eta\}$, which, provided it is nonempty, is an interval $[\ell, u]$ with $-\infty \leq \ell \leq u \leq \infty$. The optimal regime in this toy example is linear. Suppose that one attempted to estimate an optimal linear decision rule by postulating a linear model of the form $g(x, a; \beta) = \beta_0 + \beta_1 a + \beta_2 x + \beta_3 ax$ so that $\Delta(x; \beta) = 2\beta_1 + 2\beta_3 x$, which is to be estimated using least squares. It can be shown that the projection of $g(x, a)$ onto $g(x, a; \beta)$ is given by $g(x, a; \beta^*)$, where

$$
\begin{aligned}
\beta_0^* &= \frac{1}{2}\exp\left\{\alpha_0 + \alpha_1 + \frac{(\alpha_2 + \alpha_3)^2\tau^2}{2}\right\} + \frac{1}{2}\exp\left\{\alpha_0 - \alpha_1 + \frac{(\alpha_2 - \alpha_3)^2\tau^2}{2}\right\},\\
\beta_1^* &= \frac{1}{2}\exp\left\{\alpha_0 + \alpha_1 + \frac{(\alpha_2 + \alpha_3)^2\tau^2}{2}\right\} - \frac{1}{2}\exp\left\{\alpha_0 - \alpha_1 + \frac{(\alpha_2 - \alpha_3)^2\tau^2}{2}\right\},\\
\beta_2^* &= \frac{(\alpha_2 + \alpha_3)}{2}\exp\left\{\alpha_0 + \alpha_1 + \frac{(\alpha_2 + \alpha_3)^2\tau^2}{2}\right\} + \frac{(\alpha_2 - \alpha_3)}{2}\exp\left\{\alpha_0 - \alpha_1 + \frac{(\alpha_2 - \alpha_3)^2\tau^2}{2}\right\}, \text{ and}\\
\beta_3^* &= \frac{(\alpha_2 + \alpha_3)}{2}\exp\left\{\alpha_0 + \alpha_1 + \frac{(\alpha_2 + \alpha_3)^2\tau^2}{2}\right\} - \frac{(\alpha_2 - \alpha_3)}{2}\exp\left\{\alpha_0 - \alpha_1 + \frac{(\alpha_2 - \alpha_3)^2\tau^2}{2}\right\}.
\end{aligned}
$$

Thus, it can be seen that $\beta^*$ can be far from $\alpha^*$, and subsequently, the linear rule obtained by estimating $g(x, a)$ with a linear model need not lead to a high-quality linear decision rule, e.g., if $\tau = 1$ and $\alpha^* = (4.176, 1.720, -4.704, 0.320)^\intercal$, then the optimal rule $d^{\text{opt}}(x) = \text{sign}(\alpha_1^* + \alpha_3^* x)$ and the rule $\text{sign}\{\Delta(x; \beta^*)\} = \text{sign}(\beta_1^* + \beta_3^* x)$ disagree on more than 60% of the population.

## 2.3. Dynamic Data-Driven Decision Science

Estimation of an optimal treatment regime is an example of a reinforcement learning problem in that one must learn about optimal decision making using data on the interactions between one or more decision makers and the environment. There is an expansive literature on reinforcement learning in computer science and engineering (e.g., Sutton & Barto 1998, Si 2004, Powell 2007, Busoniu et al. 2010, Szepesvari 2010). This literature was developed with a focus on algorithmic efficiency, computational scalability, and empirical performance; state-of-the-art reinforcement learning algorithms are expected to identify complex and subtle patterns from massive data sets. In

contrast, in the precision medicine literature, methodology was developed with a focus on causal validity, generalizability, and interpretability within a domain context; state-of-the-art methodologies for precision medicine are expected to be transparent, to be rigorous, and to generate new scientific knowledge from (relatively) small data sets. However, cross-pollination between the two literatures is rapidly increasing due to technological advancements facilitating the collection and curation of massive amounts of patient-level data in real (or near-real) time and the emergence of mobile health (mHealth) (Luckett et al. 2017, Nahum-Shani et al. 2017, Tewari & Murphy 2017).

Precision medicine is also closely connected with control theory and operations research. Within these areas there is a rich history of modeling the underlying system dynamics (i.e., the generative model) and using simulation-optimization to inform decision making; standard texts include Hillier & Lieberman (1990) and Macia & Thaler (2005). Introducing stochasticity into dynamic systems to address various forms of uncertainty has led to many rich developments in stochastic differential equations (Nisio 2015) and in Markov decision processes (Puterman 2005). Simulation-based approaches that use interactions between complex agents, such as agent-based modeling, have also been developed, which allow for studying situations of greater complexity than normally achievable through systems of mathematical equations (Wilensky & Rand 2015). These approaches can be effective when there is rich scientific theory to inform the construction of the underlying models; however, such information is rarely available in the context of precision medicine, making these methods difficult to apply directly.



## 3. BIOMARKERS

In precision medicine research a common clinical goal is the identification of patient biomarkers that are important for choosing an optimal treatment. We use the term biomarker generically to represent a scalar feature constructed from current patient information; thus, a biomarker could be a single component of the available history or a composite measurement constructed from multiple components. A biomarker may provide valuable clinical information by being (*a*) prognostic, i.e., the biomarker is useful in predicting the mean outcome of a patient; (*b*) moderating, i.e., the biomarker is useful for predicting contrasts of the mean outcome across different candidate treatments; or (*c*) prescriptive, i.e., the biomarker is useful in selecting the treatment that maximizes the mean outcome (see Teran Hidalgo et al. 2016 for additional refinements on biomarker classification). We focus here on the mean outcome for simplicity, but other distributional summaries, such as the median, could also be used. **Figure 1** shows conceptual schematics for each of the three biomarker types. Below we formalize these notions and show that they are nested so that prescriptive biomarkers are moderating and prognostic, while moderating biomarkers are prognostic but need not be prescriptive. We focus on the single-decision setting; analogs for the multi-decision setting can be derived based on the approximate dynamic programming methods described in Section 4.

### 3.1. Prognostic Biomarkers

Consider the single-decision binary treatment setting with $A \in \mathcal{A} = \{-1, 1\}$. Then, it follows that $E\{Y^*(a)|X = x\} = \mu(x) + a\,\Delta(x)/2$, where $\mu(x) = E\{Y^*(1) + Y^*(-1)|X = x\}$. We note that these ideas can be generalized to more than two treatments, but we restrict ourselves here to the two-treatment situation for ease of exposition. To illustrate key concepts, we first assume models of the form $\mu(x; \beta_0^*) = x^\intercal \beta_0^*$ and $\Delta(x; \beta_1^*) = x^\intercal \beta_1^*$ and biomarkers under consideration are the components of $X = (X_1, \ldots, X_p)^\intercal$; a more general definition is given below. Under this model, the biomarker $X_j$ is a prognostic biomarker if either $\beta_{0,j}^*$ or $\beta_{1,j}^*$ is not zero. Under the causal conditions presented in Section 4, one can estimate the vectors $\beta_0^*$ and $\beta_1^*$ using ordinary least squares and

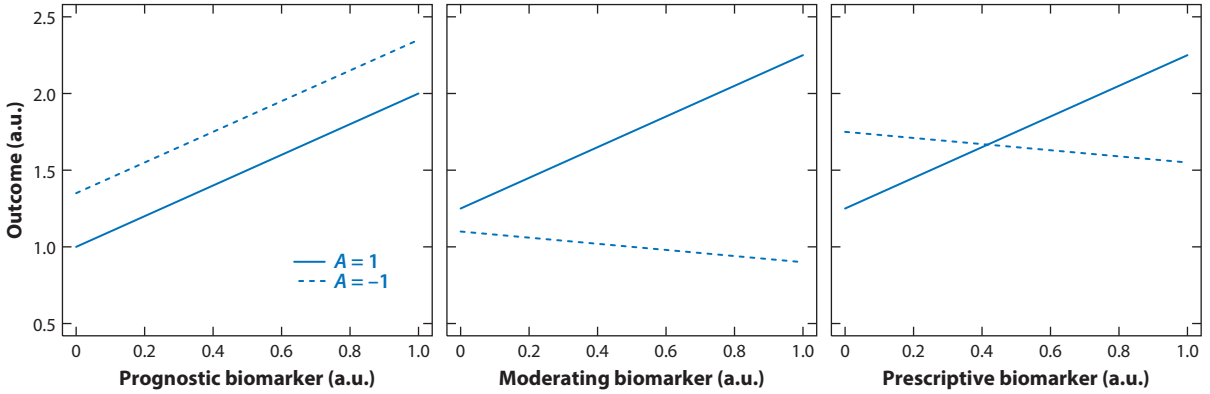


**Figure 1**

Schematics for three types of biomarker. Abbreviation: a.u., arbitrary unit.



whether a biomarker is prognostic using standard methods. The regression features found in many predictive models in medical research, such as, for example, a Cox regression model used to predict survival of non-Hodgkin's lymphoma patients (Non-Hodgkin's Lymphoma Prognostic Factors Project 1993), are examples of prognostic biomarkers.

Now consider the setting where one has a set of candidate biomarkers $\mathcal{B} = \{B_j : j = 1, \ldots, q\}$ where $B_j = B_j(X) \in \mathbb{R}$ is a possible composite summary of $X$. For any $\mathcal{J} \subseteq \{1, \ldots, q\}$, we say that $\mathcal{J}$ is sufficient for prognosis if $\sigma\{\mu(X), \Delta(X)\} \subseteq \sigma\{B_j : j \in \mathcal{J}\}$, where $\sigma(U)$ denotes the $\sigma$-algebra generated by $U$. We define a set of biomarkers, $\mathcal{J}$, to be minimal sufficient for prognosis if it is sufficient for prognosis and $\#\mathcal{J} \leq \#\mathcal{J}'$ for any other sufficient set $\mathcal{J}'$, where $\#\mathcal{B}$ denotes the number of elements in $\mathcal{B}$. We define a biomarker $B_j$ to be prognostic if $j \in \mathcal{J}$ for some minimal sufficient set $\mathcal{J}$. As in the linear setting, a general approach to identifying prognostic biomarkers is to model $\mu(x)$ and $\Delta(x)$ as functions of the candidate markers and to apply standard methods for variable selection. One could also use methods for feature construction in regression to identify the set of candidate features $\mathcal{B}$ (Cook & Ni 2005, Lee & Verleysen 2007, Li 2007).

## 3.2. Moderating Biomarkers

Moderating biomarkers are predictive of the contrast between two treatments. In the linear model example, where $\mu(x; \beta_0^*) = x^\intercal \beta_0^*$ and $\Delta(x; \beta_1^*) = x^\intercal \beta_1^*$, we say that $X_j$ is a moderating biomarker if $\beta_{1,j}^*$ is not zero. Thus, it can be seen immediately that a moderating biomarker is also prognostic. As with identifying prognostic biomarkers, under appropriate causal conditions, one can use ordinary least squares to estimate $\beta_1^*$ and test whether $\beta_{1,j}^*$ is zero to identify moderating biomarkers. Under the postulated linear model with unbounded biomarkers, every moderating biomarker is also prescriptive; however, this does not hold in general. In the more general setting with candidate biomarkers $\mathcal{B} = \{B_j : j = 1, \ldots, q\}$, we say that the subset of biomarkers $\mathcal{J} \subseteq \{1, \ldots, q\}$ is sufficient for moderation if $\sigma\{\Delta(X)\} \subseteq \sigma\{B_j : j \in \mathcal{J}\}$ and is, furthermore, minimal sufficient if $\#\mathcal{J} \leq \#\mathcal{J}'$ for any other sufficient set $\mathcal{J}'$.

## 3.3. Prescriptive Biomarkers

To illustrate the difference between moderating and prescriptive biomarkers, consider the model $\Delta(x; \beta_1^*, \beta_2^*) = \exp(x^\intercal \beta_2^*) x^\intercal \beta_1^*$ and suppose that the support of $X$ is $\mathbb{R}^p$. Then we say that $X_j$ is

a prescriptive biomarker if $\beta^*_{1,j}$ is not zero. It can be seen that if $\beta^*_{1,j}$ is zero but $\beta^*_{2,j}$ is nonzero, then $X_j$ is moderating but not prescriptive. More generally, under the setting with candidate biomarkers $\mathcal{B} = \{B_j : j = 1, \ldots, q\}$, we say that $\mathcal{J} \subseteq \{1, \ldots, q\}$ is sufficient for prescription if $\sigma[\text{sign}\{\Delta(X)\}] \subseteq \sigma(B_j : j \in \mathcal{J})$ and we say that $\mathcal{J}$ is minimally sufficient for prescription if it further satisfies $\#\mathcal{J} \leq \#\mathcal{J}'$ for all sufficient $\mathcal{J}'$. The identification of prescriptive biomarkers has been studied as a subtopic within precision medicine, with early approaches seeking to identify qualitative interactions (Gail & Simon 1985, Gunter et al. 2011, Tian et al. 2014, Shen & Cai 2016) and more recent approaches focused on identification of variables that are informative for identification of an optimal treatment regime (Fonteneau et al. 2008, Song et al. 2015a, Fan et al. 2016, Zhang & Zhang 2016).

A concrete example of both moderating and prescriptive biomarkers can be found in the work of Gail & Simon (1985), who analyze data from the National Surgical Adjuvant Breast and Bowel Project (Fisher et al. 1983). They find that a patient's age and progesterone receptor level (PR) are informative of the effect of adding tamoxifen to chemotherapy. Thus, both age and PR are moderating biomarkers. Now define the composite biomarker $D$ to be 1 if both age $< 50$ and PR $< 10$ and to be $-1$ otherwise. Gail & Simon found that if $D = 1$, tamoxifen should be added, but it should not be added otherwise. In this setting $D$ is both a moderating and a prescriptive biomarker. We note that the US Food and Drug Administration (FDA) identifies a biomarker as predictive if it is predictive of the contrast between active treatment and a control; thus, a predictive variable is prescriptive for a contrast involving a control (FDA-NIH Biomarker Working Group 2016).



## 4. ESTIMATING DYNAMIC TREATMENT REGIMES

### 4.1. The Single-Decision Setting

In the single-decision setting, as described in Section 2.1.1, the goal is to estimate a regime, $d^{\text{opt}}$, that satisfies $\text{E}Y^*(d^{\text{opt}}) \geq \text{E}Y^*(d)$ for any other regime $d$. In order to construct an estimator, we need to express $d^{\text{opt}}$ in terms of the data-generating model. To do this, we make the following assumptions: first, positivity, $\text{P}(A = a|X = x) > 0$ for all $a \in \psi(x)$ and all $x \in \mathcal{X}$; second, consistency, $Y = Y^*(A)$; and third, strong ignorability, $\{Y^*(a) : a \in \mathcal{A}\} \perp A|X$. These assumptions are standard (Kidwell 2016), though they are not as general as possible (Robins et al. 2000, Petersen et al. 2012).

Define $Q(x, a) = \text{E}(Y|X = x, A = a)$, then under the preceding assumptions, $d^{\text{opt}}(x) = \arg\max_{a \in \psi(x)} Q(x, a)$ is an optimal regime. This immediately suggests a regression-based estimator wherein one first constructs an estimator $\widehat{Q}_n(x, a)$ of $Q(x, a)$ and subsequently uses the plug-in estimator $\widehat{d}_n(x) = \arg\max_{a \in \psi(x)} \widehat{Q}_n(x, a)$. For example, one might posit a linear model of the form $Q(x, a; \beta) = \sum_{a' \in \mathcal{A}} x_{a'}^\intercal \beta_{a'} 1\{a = a'\} = x_a^\intercal \beta_a$, where $\beta = \{\beta_a : a \in \mathcal{A}\}$ and $x_a$, $a \in \mathcal{A}$ are features of $x$. Define $\mathbb{P}_n$ to be the empirical measure. Let $\widehat{\beta}_n = \arg\min_\beta \mathbb{P}_n\{Y - Q(X, A; \beta)\}^2$ and subsequently $\widehat{Q}_n(x, a) = Q(x, a; \widehat{\beta}_n)$ so that the plug-in estimator is $\widehat{d}_n(x) = \arg\max_{a \in \psi(x)} Q(x, a; \widehat{\beta}_n) = \arg\max_{a \in \psi(x)} x_a^\intercal \widehat{\beta}_{a,n}$. Linear models are commonly used because they are regarded as being easy to interpret, e.g., the coefficients of $\widehat{\beta}_{a,n}$ can be used to identify what biomarkers are likely to impact a patient's outcome under treatment $a$. However, while linearity lends itself to interpretation, this may come at the cost of misspecification; thus, as with any regression model, one should apply model diagnostics and interactive model-building techniques to ensure a high-quality model.

To mitigate misspecification, one can use a flexible class of models to estimate $Q(x, a)$, e.g., trees (Zhang et al. 2012b, Taylor et al. 2015), boosting (Kang et al. 2014), generalized additive models (Moodie et al. 2014), a change-plane (Fan et al. 2016), or nonlinear basis expansions

(Qian & Murphy 2011). However, using a flexible model for $Q(x,a)$ can render the estimated rule $x \mapsto \arg\max_{a\in\psi(x)} \widehat{Q}_n(x,a)$ unintelligible—thereby obscuring the scientific content of the estimated regime and limiting its value as a decision support tool. This issue led to the development of regression-based policy-search methods wherein the class of regimes is decoupled from the class of estimators used for $Q(x,a)$. For any regime $d$, it follows under the preceding causal conditions that $\mathrm{E}Y^*(d) = \mathrm{E}Q\{X, d(X)\}$; thus, if $\mathcal{D}$ is a prespecified class of regimes, then the optimal within this class is $d_{\mathcal{D}}^{\mathrm{opt}} = \arg\max_{d\in\mathcal{D}} \mathrm{E}Q\{X, d(X)\}$. Let $\widehat{Q}_n(x,a)$ be an estimator of $Q(x,a)$, and then the plug-in estimator of $d_{\mathcal{D}}^{\mathrm{opt}}$ is $\widehat{d}_{\mathcal{D},n} = \arg\max_{d\in\mathcal{D}} \mathbb{P}_n \widehat{Q}_n\{X, d(X)\}$. Because the class $\mathcal{D}$ is chosen independently of the class of models for $Q(x,a)$, one can use nonparametric regression estimators while maintaining control of the form of the estimated optimal regime. Furthermore, because this approach is built upon regression, it is easily extensible to settings with complex data structures, censored data, or measurement error, or to other settings for which regression models have been developed.

Regression-based estimators were derived from a regression-based representation of the optimal treatment regime, an alternative representation based on importance sampling leads to another class of estimators termed direct-search or classification-based estimators. For simplicity, we assume that treatment is binary and coded so that $A \in \mathcal{A} = \{-1, 1\}$. Under the causal conditions stated above, the marginal mean outcome under a regime $d$ is

$$V(d) \triangleq \mathrm{E}Y^*(d) = \mathrm{E}\left\{\frac{Y 1_{d(X)=A}}{\pi(A;X)}\right\},$$

where $\pi(a;x) = \mathrm{P}(A = a|X = x)$ is the propensity score (Qian & Murphy 2011). Thus, the optimal regime satisfies $d^{\mathrm{opt}} = \arg\max_d V(d)$. Given an estimator $\widehat{\pi}_n(a;x)$ of $\pi(a;x)$, which might be obtained using logistic regression, the inverse probability weighted estimator of $V(d)$ is given by $\widehat{V}_n(d) = \mathbb{P}_n\{Y\, 1_{d(X)=A}/\widehat{\pi}_n(A;X)\}$. Given a class of regimes, $\mathcal{D}$, one could construct an estimator of $d_{\mathcal{D}}^{\mathrm{opt}}$ by direct maximization, i.e., $\widehat{d}_{\mathcal{D},n} = \arg\max_{d\in\mathcal{D}} \widehat{V}_n(d)$; however, because of the discontinuous indicator function, direct optimization is not computationally feasible save for settings where $\mathcal{D}$ is small. However, it can be seen that

$$V(d) = -\mathrm{E}\left\{\frac{|Y| 1_{A\mathrm{sign}(Y)d(X)<0}}{\pi(A;X)}\right\} + \mathrm{E}\left\{\frac{(Y)_+}{\pi(A;X)}\right\},$$

where $\mathrm{sign}(u) = 1_{u>0} - 1_{u<0}$ is the sign function and $(u)_+ = \max(0,u)$ is the positive part function. Thus, given a class of regimes, $\mathcal{D}$, the optimal regime within this class satisfies

$$d_{\mathcal{D}}^{\mathrm{opt}} = \arg\min_{d\in\mathcal{D}} \mathrm{E}\left\{\frac{|Y| 1_{A\mathrm{sign}(Y)d(X)<0}}{\pi(A;X)}\right\};$$

hence, it can be seen that $d_{\mathcal{D}}^{\mathrm{opt}}$ minimizes a cost-sensitive classification problem (Elkan 2001, Zadrozny et al. 2003, Zhou & Liu 2006, Pires et al. 2013) with cost function $|Y|/\pi(A;X)$, class label $\mathrm{A}\mathrm{sign}(Y)$, and input $X \in \mathcal{X}$, over the set of classifiers $\mathcal{D}$ (see Zhang et al. 2012a,b; Zhao et al. 2012). Thus, one can estimate $d_{\mathcal{D}}^{\mathrm{opt}}$ by applying off-the-shelf classification algorithms using the estimated cost $|Y|/\widehat{\pi}_n(A;X)$ in place of $|Y|/\pi(A;X)$; for a discussion of converting cost-sensitive classification problems into standard (i.e., constant cost) classification problems, readers are directed to Zadrozny et al. (2003) and references therein. Outcome weighted learning (OWL; Zhao et al. 2012) uses this framework with support vector machines (see, e.g., chapter 12 of Hastie et al. 2009) to estimate an optimal treatment regime; convergence rates for OWL were among the first to establish the mathematical underpinnings of direct-search estimation for optimal treatment regimes and led to a series of refinements including residual OWL (Laber & Zhao 2015, Y. Wang et al. 2016, Zhou et al. 2017) and improved bounds based on efficiency theory (Athey & Wager

2017). More generally, policy-search methods based on maximizing an estimator $\widehat{V}_n(d)$ of $V(d)$ over a prespecified class of regimes have been extended to a wide range of settings, including ordinal treatments (Chen et al. 2018), right-censored outcomes (Zhao et al. 2015b, Cui et al. 2017; see also Bai et al. 2017), continuous treatments (Laber & Zhao 2015, Chen et al. 2016, Kallus 2018), high-dimensional data (Song et al. 2015a, Jeng et al. 2018), and more than two treatments in an observational setting (Kallus 2016, 2017).

## 4.2. The Multi-Decision Setting

The multi-decision setting is complicated by the need to account for delayed treatment effects and prognostic effects, e.g., information gain that improves decision making at subsequent decision points. We consider two multi-decision settings: first, a finite time horizon wherein the number of decision points is small and finite, and second, an indefinite time horizon wherein the number of decision points is large or indeterminate. There are, of course, many intermediate settings, but these two encompass many commonly encountered settings in precision medicine. The methods we discuss apply to both observational and randomized studies.

**4.2.1. Finite time horizon.** In the finite horizon setting, we can estimate the optimal dynamic treatment regimes through a variety of reinforcement learning techniques, including g-estimation (Robins 2004, Stephens 2016), Q- and A-learning (for an overview of these methods, see Schulte et al. 2014, wherein g-estimation is articulated as a special case of A-learning), modeling of the entire longitudinal process (Xu et al. 2016), and several extensions of OWL (Zhao et al. 2015a), among other approaches. Recall that in precision medicine, we are primarily interested in estimating the decision rule, and, in many settings, it is much more robust and feasible to not model the entire process if possible. Many of the learning methods listed above were motivated, at least in part, to obtain the dynamic treatment regime without needing to model the full process. Early seminal work in estimating dynamic treatment regimes includes that of Robins (1986, 1997, 2004), Murphy et al. (2001), Murphy (2003), van der Laan et al. (2001), and van der Laan & Peterson (2004). Because of its flexibility and relative ease in implementation, we present Q-learning in some detail, followed by a brief discussion of OWL and several related methods, but we do not further discuss other approaches here.

As in the single-decision setting, we derive regression-based and inverse-weighting or classification-based representations of the optimal treatment regime in terms of the data-generating model and subsequently use these representations to construct estimators of the optimal regime. Using the notation of Section 2.1.1, we make the following assumptions: first, positivity, $\mathrm{P}(A_t = a_t|H_t = h_t) > 0$ for all $a_t \in \psi_t(h_t)$ and $h_t \in \mathcal{H}_t$; second, consistency, $H_t = H_t^*(\overline{A}_{t-1})$ for $t = 2, \ldots, T$ and $Y = Y^*(\overline{A}_T)$; and third, sequential ignorability $\{Y^*(\overline{a}_T), H_T^*(\overline{a}_{T_1}, \ldots, H_2^*(a_1)), H_1 : \overline{a}_T \in \otimes_{t=1}^T \mathcal{A}_t\} \perp A_t|H_t$ for $t = 1, \ldots, T$, where $\otimes$ denotes the Cartesian product taken over the specified range of indices. Schulte et al. (2014) provides additional discussion of these assumptions.

Define $Q_T(h_T, a_T) = \mathrm{E}(Y|H_T = h_T, A_T = a_T)$, and for $t = T-1, \ldots, 1$ define $Q_t(h_t, a_t) = \mathrm{E}\{\max_{a_{t+1}} Q_{t+1}(H_{t+1}, a_{t+1})|H_t = h_t, A_t = a_t\}$; then, it follows from dynamic programming (Bellman 1957) that

$$d_t^{\mathrm{opt}}(h_t) = \arg \max_{a_t \in \psi_t(h_t)} Q_t(h_t, a_t), \qquad 1.$$

which we term the regression-based representation of the optimal regime. Q-learning is an approximate dynamic programming algorithm based on Equation 1, which proceeds as follows. Construct an estimator $\widehat{Q}_{T,n}(h_T, a_T)$ of $Q_T(h_T, a_T)$ obtained by regressing $Y$ on $H_T$ and $A_T$;

subsequently, for $t = T-1, \ldots, 1$, let $\widehat{Q}_{t,n}(h_t, a_t)$ be an estimator of $Q_t(h_t, a_t)$ obtained by regressing $\max_{a_{t+1} \in \psi_{t+1}(H_{t+1})} \widehat{Q}_{t+1,n}(H_{t+1}, a_{t+1})$ on $H_t$ and $A_t$. The Q-learning estimator of $d^{\text{opt}}$ is thus $\widehat{d}_{t,n}(h_t) = \arg\max_{a_t \in \psi_t(h_t)} \widehat{Q}_{n,t}(h_t, a_t)$ for $t = 1, \ldots, T$ (Murphy 2005a, Schulte et al. 2014).

Because Q-learning relies on a series of regression models, it is easily extensible to a variety of models and data structures (Zhao et al. 2009; Goldberg & Kosorok 2012; Moodie et al. 2012, 2014) and allows the user to interactively construct and critique the models used for the Q-functions (Rich et al. 2010, Laber et al. 2014a). However, as in the single-decision setting, in the above formulation the estimated optimal decision rule is tied to the class of models used for the Q-functions; thus, one may be forced with an unpleasant trade-off between severe model misspecification and an unintelligible black box. Instead, one can use Q-learning with policy-search wherein the class of regimes is divorced from the class of models for the Q-functions (see Laber et al. 2018b, Zhang et al. 2018).

An alternative representation of the optimal decision rule is based on inverse probability weighting. For any regime $d$, it follows that

$$V(d) = \mathrm{E}\left( Y \prod_{t=1}^{T} \frac{1_{d_t(H_t)=A_t}}{\pi_t(A_t|H_t)} \right).$$

Thus, given estimators $\widehat{\pi}_{t,n}(a_t; h_t)$ of $\pi_t(a_t; h_t)$ for $t = 1, \ldots, T$, the plug-in estimator of $V(d)$ is

$$\widehat{V}_n(d) = \mathbb{P}_n \left( Y \prod_{t=1}^{T} \frac{1_{d_t(H_t)=A_t}}{\pi_{t,n}(A_t|H_t)} \right),$$

and given a class of regimes $\mathcal{D}$, an estimator of $d^{\text{opt}}_{\mathcal{D}}$ is $\widehat{d}_{\mathcal{D},n} = \arg\max_{d \in \mathcal{D}} \widehat{V}_n(d)$. Directly computing $\widehat{d}_{\mathcal{D},n}$ is difficult except in small problems, as the indicators make this into a discontinuous optimization problem. This computational issue is ameliorated in Zhao et al. (2015a) through the use of a surrogate optimization function for $\hat{V}_n(d)$ which is smooth and has a global optimum.

However, computational issues aside, there is another difficulty with optimizing $\hat{V}_n(d)$ or one of its surrogates when the number of time points $T$ is large, as (*a*) the product of indicators can rapidly become zero for the majority of subjects—for example, with binary treatments assigned uniformly at random at each stage, the product of indicators will be zero for all but $n(1/2)^T$ of the original $n$ subjects on average—and (*b*) the product of the propensity scores can grow quite small, leading to high variance. For these reasons, direct search estimators based on $\widehat{V}_n(d)$ work best for settings where $T$ is small, e.g., $T = 2$; however, in such settings OWL and related methods can offer significant gains in terms of robustness and marginal mean outcome (see Zhang et al. 2013, Zhao et al. 2015a). As in the single-decision setting, OWL and other direct search estimators are based on converting $\widehat{V}_n(d)$ or a more efficient augmented variant of this estimator into a cost-sensitive classification problem. We note that products of indicators across time points share a similar structure to hierarchical classification problems (Gordon 1987, Wang et al. 2009), though this connection has yet to be fully explored.

**4.2.2. Infinite time horizon.** The infinite horizon setting applies when a sequence of similar decisions need to be made over an extended time, as happens, for example, when treating diabetes or other chronic diseases as mentioned previously. In this setting, decision making is typically modeled as a Markov decision process (Puterman 2005), which encompasses a tremendously broad class of decision problems (Sutton 1997). We assume that the observed data are of the form $\{(S_{1,i}, A_{1,i}, S_{2,i}, \ldots, S_{T-1,i}, A_{T-1,i}, S_{T,i})\}_{i=1}^{n}$, which comprise $n$ i.i.d. replicates of $(S_1, A_1, S_2, \ldots, S_{T-1}, A_{T-1}, S_T)$, where $S_t \in \mathcal{S}$ denotes a summary of the patient's health status at time $t$; $A_t \in \mathcal{A}$ denotes the treatment applied at time $t = 1, \ldots, T$; and $T$ is the observed time horizon. We assume that there exists a momentary reward function $y : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ so that

$y(s, a, s')$ captures the momentary goodness for a patient with health status $s$ who receives treatment $a$ and subsequently transitions to a health status $s'$. We write $Y_t = y(S_t, A_t, S_{t+1})$ to denote the observed momentary outcome at time $t$. While the observed data are collected over a horizon $T$, the goal may be to estimate a treatment regime that can be applied indefinitely to a new patient, i.e., well beyond $T$ decision points. To this end we assume that the observed data are Markov and homogeneous so that for any $\mathcal{Z} \subseteq \mathcal{S}$ and $t \geq 1$,

$$\mathrm{P}\left(S_{t+1} \in \mathcal{Z} | \overline{S}_t = \overline{s}_t, \overline{A}_t = \overline{a}_t\right) = \mathrm{P}\left(S_{t+1} \in \mathcal{Z} | S_t = s_t, A_t = a_t\right) = \mu_{s_t, a_t}(\mathcal{Z}),$$

where the measure $\mu_{s_t, a_t}$ does not depend on time. In application, the raw data may not satisfy these conditions, and one must judiciously construct the state $S_t$ as a summary of the raw data to ensure that these conditions hold (at least approximately; see Wang et al. 2018a and references therein).

For each $s_t \in \mathcal{S}$, let $\psi(s_t) \subseteq \mathcal{A}$ denote the set of allowable treatments for a patient with status $S_t = s_t$. A treatment regime in this context is a map $d : \mathcal{S} \to \mathcal{A}$ that satisfies $d(s) \in \psi(s)$ for all $s \in \mathcal{S}$ so that under $d$ a patient presenting with state $S_t = s_t$ at time $t \geq 1$ would be recommended to receive treatment $d(s_t)$; because $d$ is stationary it can be applied for all $t$, even if $t > T$.[1] Let $S_t^*(\overline{a}_{t-1})$ denote the potential patient status at time $t$ under treatment sequence $\overline{a}_{t-1} \in \otimes_{v=1}^{t-1} \mathcal{A}$ so that the potential status under a regime $d$ is

$$S_t^*(d) = \sum_{\overline{a}_{t-1}} S_t^*(\overline{a}_{t-1}) \prod_{v=1}^{t-1} 1_{\left[d\left\{S_v^*(\overline{a}_{v-1}) = a_v\right\}\right]}.$$

The potential momentary outcome for regime $d$ is $Y_t^*(d) = y[S_t^*(d), d\{S_t^*(d)\}, S_{t+1}^*(d)]$. Define the conditional discounted marginal mean outcome under $d$ to be

$$V(s; d) = \mathrm{E}\left\{ \sum_{k \geq 0} \gamma^k Y_{t+k}^*(d) \middle| S_t = s \right\},$$

where $\gamma \in [0, 1)$ is a discount factor that balances the trade-off between immediate and long-term outcomes. An optimal regime, $d^{\mathrm{opt}}$, satisfies $V(s; d^{\mathrm{opt}}) \geq V(s; d)$ for all $s \in \mathcal{S}$ and all regimes $d$. Given distribution $R$ on $\mathcal{S}$ we define the marginal mean outcome with respect to reference $R$ to be $V_R(d) = \int V(s; d) \mathrm{d}R(s)$. One can think of the reference distribution $R$ as the observed sample distribution or the distribution of a potential future population of patients, whichever is of primary interest. In the context of policy-search methods over a prespecified class of regimes, $\mathcal{D}$, we define the optimal regime within $\mathcal{D}$ with respect to reference $R$ as $d_{\mathcal{D},R}^{\mathrm{opt}} = \arg\max_{d \in \mathcal{D}} V_R(d)$.

Under causal assumptions analogous to those used in the finite horizon case, the following recursion holds:

$$V(s, d) = \mathrm{E}\left[ \frac{1_{A_t = d(S_t)}}{\pi_t(A_t; S_t)} \left\{ Y_t + \gamma V(S_{t+1}, d) \right\} \middle| S_t = s \right],$$

from which it can be seen that for any function $\phi : \mathcal{S} \to \mathbb{R}^q$ (see Precup 2000, Paduraru 2013, Luckett et al. 2017),

$$0 = \mathrm{E}\left[ \frac{1_{A_t = d(S_t)}}{\pi_t(A_t; S_t)} \left\{ Y_t + \gamma V(S_{t+1}, d) - V(S_t, d) \right\} \phi(S_t) \right].$$

[1]There is little loss in generality in restricting attention to stationary regimes; under mild regularity conditions there exists a stationary regime that leads to a discounted marginal mean outcome at least as large as any other (possibly nonstationary) regime (Bertsekas 2005).

The forgoing expression forms the basis for an estimating equation for $V(s,d)$. Let $V(s,d;\lambda)$ be a postulated class of models for $V(s,d)$ indexed by $\lambda \in \Lambda \subseteq \mathbb{R}^q$; we assume that $V(s,d;\lambda)$ is differentiable in $\lambda$ for each $s$ and $d$. Define

$$\Lambda_n(d,\lambda) = \mathbb{P}_n \sum_{t=1}^{T-1} \left[ \frac{1_{A_t=d(S_t)}}{\pi_t(A_t;S_t)} \left\{ Y_t + \gamma V(S_{t+1},d) - V(S_t,d) \right\} \nabla_\lambda V(S_t,d;\lambda) \right],$$

and let $\widehat{\lambda}_n(d)$ be a solution to $\Lambda_n(d,\lambda) = 0$. The estimated conditional marginal mean outcome is $\widehat{V}_n(s,d) = V\{s,d;\widehat{\lambda}_n(d)\}$. Furthermore, given a reference distribution, $R$, a policy-search estimator of $d^{\text{opt}}_{\mathcal{D},R}$ is $\widehat{d}_{\mathcal{D},R,n} = \arg\max_{d\in\mathcal{D}} \int \widehat{V}_n(s,d)\mathrm{d}R(s)$. Note that this approach, called V-learning, is an infinite horizon variant of OWL (see Luckett et al. 2017 for additional description and an online version that uses stochastic regimes). Q-learning in the infinite horizon setting can be derived using analogous arguments (see Ertefaie 2015) and has been applied to manage infection from *Pseudomonas aeruginosa* in patients with cystic fibrosis (Tang & Kosorok 2012) and to control diabetes (see Ertefaie 2015).

We note that inference for V-learning has been derived based on assuming that the underlying dynamic process is constant or, more precisely, stationary, at least over moderately long stretches of time (see Luckett et al. 2017). However, in some practical settings, such as in diabetes, it may be more realistic to expect the dynamics to gradually change as patients age. In this more complicated setting, inference is more challenging and remains an open research question.

**4.2.3. Mobile health.** One important motivation for infinite horizon reinforcement learning is mHealth, in which it is feasible to both collect information and provide interventions to patients in real time. This is the case for the work in both Ertefaie (2015) and Luckett et al. (2017). Such data can be collected retrospectively or by using sequential multiple assignment randomized trial (SMART) designs. A somewhat different approach to precision medicine in mHealth involves using data obtained from a microrandomized clinical trial developed by Klasnja et al. (2015). These are particularly suited for developing interventions involving prompting people to take healthy actions to improve health behavior (Bekiroglu et al. 2017). Often, these are designed to improve proximal outcomes, not necessarily longer-term outcomes, and can be framed as a contextual bandit problem (Tewari & Murphy 2017). This is an active and exciting area of precision medicine research.

## 4.3. Data Sources and Study Design

Data for estimating dynamic treatment regimes can come from a range of sources, including convenience samples, planned observational studies, randomized clinical trials, and hybrid study designs (Zatzick et al. 2016, Liu et al. 2017). In many of these sources, including randomized clinical trials, the primary study objective is not estimation of an optimal treatment regime (Lavori & Dawson 2000, 2004; Murphy 2005b; Laber et al. 2016); at best, estimation of an optimal treatment regime is a planned—but strictly exploratory—analysis. However, such data can still be a rich resource for estimation and inference for optimal treatment regimes. Electronic health records, for example, are collected for administrative or insurance purposes but have been shown to be useful for precision medicine research (see, e.g., Hripcsak et al. 2016). Planned observational studies are usually conducted in epidemiology and other fields, but they usually involve careful design, planning, and execution so that the quality of data is high (see, e.g., Thiese 2014). These designs are frequently the inspiration for causal inference research as the absence of randomized treatment assignment complicates identification of causal relationships among treatments and

risk factors. These designs can also include careful selection of subsets of convenience samples to improve quality of causal inference, as done, for example, in Lund et al. (2015).

Randomized clinical trials are a gold standard for data collection as they protect against unmeasured confounding and can be designed to ensure efficient estimation of the targeted estimand. For single-stage decision problems, a *k*-arm randomized clinical trial with equal randomization provides maximal information about average treatment effects across pairs of treatments. For multi-stage decision problems, SMARTs (Lavori & Dawson 2000, 2004; Murphy 2005b) allow for the efficient comparison of treatment sequences and fixed (i.e., not data-dependent) treatment regimes. In a SMART, a patient is randomized at each point in the treatment process where there is clinical equipoise, and thus, each patient may be randomized multiple times throughout the trial. **Figure 2** shows a schematic for a two-stage SMART for evaluating behavioral interventions for cancer pain management (Kelleher et al. 2017). In the first stage, subjects were randomized



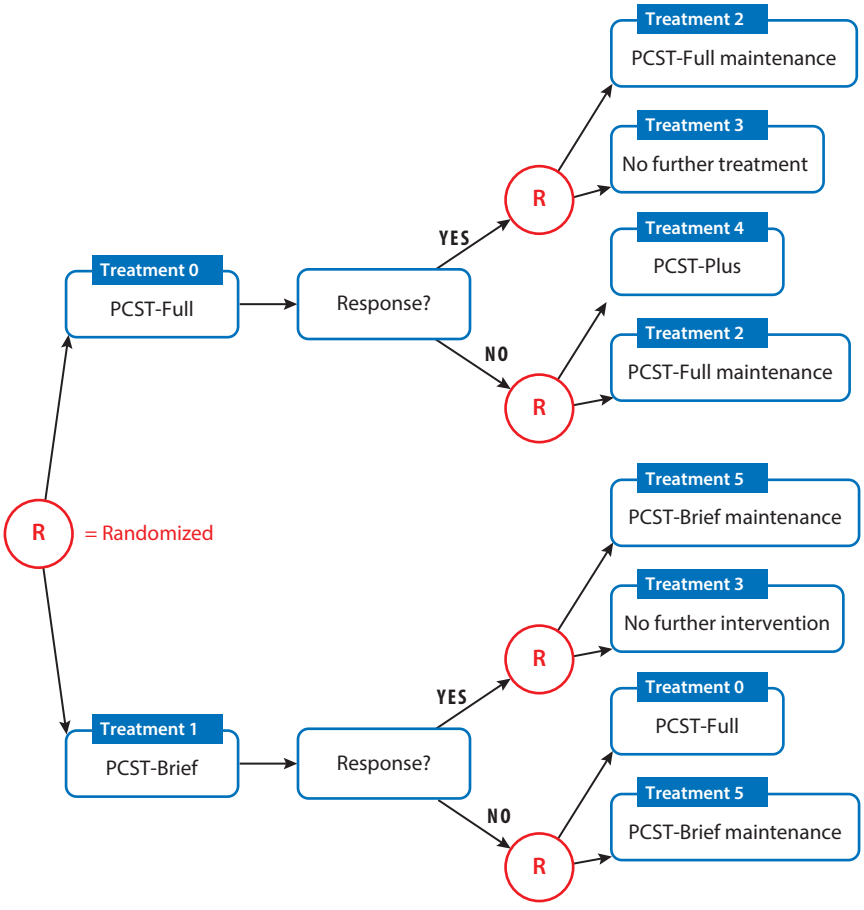


**Figure 2**

Two-stage sequential multiple assignment randomized trial (SMART) for evaluating pain coping skills training (PCST) for cancer pain management. In the first stage, subjects are randomized to receive either PCST-Full or PCST-Brief. At the second stage, responders are randomized to a maintenance therapy or no further treatments, whereas nonresponders are randomly assigned to a maintenance therapy or more intensive treatment. Kelleher et al. (2017) provide additional trial details.

with equal probability to one of two variants of pain coping skills training; in the second stage, responders were randomized to either a maintenance therapy or no further treatment, whereas nonresponders were randomized to either maintenance therapy or an intensified treatment. Two-stage SMARTs are among the most common, though they are an extremely flexible design with many different variations (Kidwell 2014, 2016, Penn State Methodology Center 2018).

For various reasons, randomization at each decision point is not always possible, ethical, feasible, or even scientifically optimal, and various hybrid designs can be considered. A pragmatic clinical trial (see, e.g., Ford & Norrie 2016) is one in which various design features are carefully incorporated to ensure similarity with treatment conditions in the real world. This can include randomizing clinics instead of patients to ensure that the treatment is the same throughout the clinic as would normally happen in practice, or recruiting a more heterogeneous patient population through liberal inclusion criteria. In a certain sense, a SMART clinical trial is more pragmatic than traditional clinical trials as the treatment decisions being evaluated are more similar to those utilized in practice. Hybrid designs have both randomization and observational components and often have a pragmatic motivation. One example is the enrichment design proposed by Liu et al. (2017), which allows the first treatment assignment to be nonrandomized but has the second treatment assignment randomized. There are many other possibilities that we will not explore further here. Two important points we want to make are, first, that heterogeneity is good for precision medicine as this is needed to estimate an optimal treatment regime which is valid for a broad range of possible patients, and, second, that design of studies used for discovering precision medicine is a crucially important aspect of precision medicine research.

## 5. MANAGING MULTIPLE OUTCOMES

So far, we have considered optimal dynamic treatment regimes in terms of a single scalar outcome that we wish to optimize. In many clinical settings, there are multiple outcomes that need to be considered when managing treatment decisions. For example, in treating schizophrenia, there can be steep trade-off between side effects and efficacy (Butler et al. 2018). As another example, consider bipolar depression, in which there is a trade-off between depressive symptoms and risk of mania (Luckett et al. 2018). Recent work on precision medicine with multiple outcomes includes set-valuated treatment regimes that recommend a set of acceptable treatments given a patient's history (Fard 2009, Lizotte et al. 2012, Laber et al. 2014b, Lizotte & Laber 2016, Wu 2016) and methods where a primary outcome is maximized while a secondary outcome is constrained to be within an acceptable region, using either regression-based or OWL methods (Linn et al. 2015, Wang et al. 2018c, Laber et al. 2018b).

In some cases, there is a trade-off between two or more endpoints that depends on patient preferences or other factors that depend on individual patient needs. Butler et al. (2018) develop a method for the single-decision setting that uses item response theory to elicit patient preferences and then combines this with Q-learning to optimize the patient-preferred composite utility, which is a convex combination of two outcomes. They show that under reasonable regularity and logic conditions, the optimal regime for any utility is equivalent to the optimal regime for a utility expressed as a convex combination of the available outcomes. They also show that the item-response model leads to the optimal patient-preferred treatment decisions under reasonable regularity conditions as the number of items grows and the sample size increases. One caveat of their approach is that the items in the patient preference instrument need to be appropriately calibrated, which can be challenging. Extensions that allow calibration to be done empirically and that can be applied to both single- and multi-decision settings are given in Butler (2016).

Luckett et al. (2018) study the situation where the trade-offs between two outcomes depend on complex individual-level factors about which clinicians have imperfect information. They consider observational data on clinicians prescribing antidepressants to patients with bipolar depression, and measures of both depression and mania outcomes are observed in the patients. They assume that the clinicians are trying to act optimally and that they succeed sometimes but not always. Based on the estimated Q-functions for each outcome (depression and mania), they estimate the weight in the combined utility of the convex combination of the two outcomes as a function of patient-level covariates, as well as the probability of correct treatment assignment, also as a function of patient-level covariates. They demonstrate that under reasonable regularity conditions, the asymptotic joint limiting distribution of the parameters is obtained at the $\sqrt{n}$ rate. The limiting distribution is non-Gaussian and requires a nonstandard bootstrap for inference. They demonstrate the validity of the inference through simulation studies and apply the method to the Systematic Treatment Enhancement Program for Bipolar Disorder study data (Sachs et al. 2007). They also demonstrate that applying the estimated dynamic treatment regime obtained from these data using the proposed method can lead to a statistically significant increased average patient-specific composite outcome for future patients. Generally speaking, addressing multiple outcomes in precision medicine is crucially important, and there is much interesting work yet to be done in this area.



## 6. STATISTICAL INFERENCE

For many of the methods described above, asymptotic consistency (the property that the estimated quantities converge to the truth as the sample size grows) has been proven. For the sake of discussion, we refer to asymptotic consistency as zero-order inference. However, typically with statistical procedures, it is valuable to be able to also provide first-order inference consisting of confidence intervals, hypothesis tests, and sample size calculations. Generally speaking, first-order inference is not yet known for many of the machine learning tools used in precision medicine, and this is an open and active area of research. Because the focus is to inform decision making, a primary emphasis is on inference for performance of a treatment regime; note that a confidence or prediction set for the marginal mean outcome of an estimated optimal treatment regime is still meaningful even if models underlying estimation of the regime are misspecified.

Although first-order inference for many of the machine learning approaches utilized has not been developed, some advances have been made in a number of settings, including for support vector machines (Laber & Murphy 2011) and random forests (Wager & Athey 2018). In addition, computation of error bounds can be a useful assessment of performance that is more precise than the presence or absence of consistency but not precise enough to obtain first-order inference. These have been developed for many machine learning tools used in precision medicine, as in Qian & Murphy (2011), Goldberg & Kosorok (2012), Zhao et al. (2012), Cui et al. (2017), and, more recently, have been improved for some settings in Athey & Wager (2017). We also note the sample size formulas for the single-decision setting have been developed based on the value function (Laber et al. 2016). Because regression-based approaches applied to the single-decision setting involve standard regression analyses, inference in this setting can sometimes be straightforward. However, in the multi-decision setting, for example, with Q-learning involving two or more decision times, the inference is nonregular even if linear regression is used at each decision time (Chakraborty et al. 2010, 2013, 2014; Moodie & Richardson 2010; Laber et al. 2014c; Song et al. 2015b).

Inference for precision medicine is an active and important area of research, and we have only touched on it briefly here. We note that many of the inferential challenges follow from the use of complex machine learning procedures. One could argue that this is a reason to avoid machine

learning methods in precision medicine. However, since the primary goal of precision medicine is to find dynamic treatment regimes that perform well on future patients, we need to use the best available tools, and this includes machine learning methods.

## 7. DISCUSSION

Precision medicine is beginning to emerge as a well-defined discipline with specific goals, areas of focus, and tailored methodology. Specifically, the primary goal is to discover treatment rules that leverage heterogeneity to improve clinical decision making in a manner that is reproducible, generalizable, and adaptable as needed. We note that patient heterogeneity is a blessing for precision medicine, although it may not be convenient for other areas of medical research. We also highlight the focus in precision medicine on discovery, as opposed to confirmatory research, and note that this makes the inferential aspects somewhat distinct from some areas of traditional medical research. Nevertheless, discovery in precision medicine should be confirmed rigorously, just as with other medical discoveries. The emphasis on both discovery and heterogeneity makes machine learning tools particularly valuable in this quest, and this means that the inferential challenges are different and in many ways more difficult.

We also note that there are many other important supporting aspects of precision medicine that we have not discussed, including implementation, national policy questions, and data storage and management, among many others. We also have not included numerous relevant research contributions to machine learning and other areas in many disciplines, both within statistics as well as outside, including many in the biomedical sciences, computer science, operations research, engineering, robotics, economics, and other areas.

Nevertheless, we hope that this review helps to clarify the goals of precision medicine and becomes a catalyst for bringing together the diverse disciplines and perspectives that are needed to make dramatic advances in precision medicine that will yield fundamental changes in human health and well-being. This is an exciting and vibrant area of research with many open questions and tremendous opportunities.

## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

The authors thank the editor and reviewers for constructive feedback that led to an improved manuscript. Both authors were funded in part by grant P01 CA142538 from the National Cancer Institute, and the second author was also funded in part by grants DMS-1555141 from the National Science Foundation and R01 DK108073 from the National Institute of Diabetes and Digestive and Kidney Diseases.

## LITERATURE CITED

Athey S, Wager S. 2017. Efficient policy learning. arXiv:1702.02896 [math.ST]

Bai X, Tsiatis AA, Lu W, Song R. 2017. Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime Data Anal.* 23:585–604

Bekiroglu K, Lagoa C, Murphy SA, Lanza ST. 2017. Control engineering methods for the design of robust behavioral treatments. *IEEE Trans. Control Syst. Technol.* 25:979–90

Bellman RE. 1957. *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press
Bembom O, van der Laan MJ. 2008. Analyzing sequentially randomized trials based on causal effect models for realistic individualized treatment rules. *Stat. Med.* 27:3689–716
Bertsekas DP. 2005. *Dynamic Programming and Optimal Control*, Vol. 1. Belmont, MA: Athena Scientific. 3rd ed.
Biernot P, Moodie EE. 2010. A comparison of variable selection approaches for dynamic treatment regimes. *Int. J. Biostat.* 6(1):6
Busoniu L, Babuska R, De Schutter B, Ernst D. 2010. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. New York: CRC
Butler EL. 2016. *Using patient preferences to estimate optimal treatment strategies for competing outcomes*. PhD Diss., Univ. N. C. Chapel Hill
Butler EL, Laber EB, Davis SM, Kosorok MR. 2018. Incorporating patient preferences into estimation of optimal individualized treatment rules. *Biometrics* 74:18–26
Cesa-Bianchi N, Lugosi G. 2006. *Prediction, Learning, and Games*. Cambridge, UK: Cambridge Univ. Press
Chakraborty B, Laber EB, Zhao YQ. 2013. Inference for dynamic treatment regimes using an *m*-out-of-*n* bootstrap scheme. *Biometrics* 69:714–23
Chakraborty B, Laber EB, Zhao YQ. 2014. Inference about the expected performance of a data-driven dynamic treatment regime. *Clin. Trials* 11(4):408–17
Chakraborty B, Moodie EE. 2013. *Statistical Methods for Dynamic Treatment Regimes*. New York: Springer
Chakraborty B, Murphy SA, Strecher V. 2010. Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.* 19:317–43
Chen G, Zeng D, Kosorok MR. 2016. Personalized dose finding using outcome weighted learning (with discussion and rejoinder). *J. Am. Stat. Assoc.* 111:1509–47
Chen J, Fu H, He X, Kosorok MR, Liu Y. 2018. Estimating individualized treatment rules for ordinal treatments. *Biometrics.* 74:924–33
Cook RD, Ni L. 2005. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J. Am. Stat. Assoc.* 100:410–28
Cui Y, Zhu R, Kosorok MR. 2017. Tree based weighted learning for estimating individualized treatment rules with censored data. *Electron. J. Stat.* 11:3927–53
Dawid AP. 2015. Statistical causality from a decision-theoretic perspective. *Annu. Rev. Stat. Appl.* 2:273–303
Eddy DM. 1990. Practice policies: guidelines for methods. *J. Am. Med. Assoc.* 263:1939–41
Elkan C. 2001. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–78. San Francisco: Morgan Kaufmann
Ertefaie A. 2015. Constructing dynamic treatment regimes in infinite-horizon settings. arXiv:1406.0764v2 [stat.ME]
Fan A, Lu W, Song R. 2016. Sequential advantage selection for optimal treatment regime. *Ann. Appl. Stat.* 10:32–53
Fard MM. 2009. *Non-deterministic policies in Markovian processes*. PhD Diss., McGill Univ.
FDA-NIH (Food. Drug. Adm. Natl. Inst. Health) Biomarker Working Group. 2016. *BEST (Biomarkers, Endpoints, and Other Tools) Resource*. Silver Spring, MD: FDA
Fisher B, Redmond C, Brown A, Wickerham DL, Wolmark N, et al. 1983. Influence of tumor estrogen and progesterone receptor levels on the response to Tamoxifen and chemotherapy in primary breast cancer. *J. Clin. Oncol.* 1:227–41
Fonteneau R, Wehenkel L, Ernst D. 2008. *Variable selection for dynamic treatment regimes: a reinforcement learning approach*. Paper presented at European Workshop on Reinforcement Learning 2008, Villeneuve d'Ascq, France, June 30–July 3
Ford I, Norrie J. 2016. Pragmatic trials. *N. Engl. J. Med.* 375:454–63
Gail M, Simon R. 1985. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 41:361–72
Goldberg Y, Kosorok MR. 2012. Q-learning with censored data. *Ann. Stat.* 40:529–60
Gordon AD. 1987. A review of hierarchical classification. *J. R. Stat. Soc. A* 150:119–37
Gunter L, Zhu J, Murphy SA. 2011. Variable selection for qualitative interactions. *Stat. Methodol.* 8:42–55

Hastie T, Tisbhsirani R, Friedman J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer
Henderson R, Ansell P, Alshibani D. 2010. Regret-regression for optimal dynamic treatment regimes. *Biometrics* 66:1192–201
Hillier FS, Lieberman GJ. 1990. *Introduction to Operations Research*. New York: McGraw-Hill. 5th ed.
Hripcsak G, Ryan PB, Duke JK, Shah NH, Park RW, et al. 2016. Characterizing treatment pathways at scale using the OHDSI network. *PNAS* 113:7329–36
Jeng XJ, Lu W, Peng H. 2018. High-dimensional inference for personalized treatment decision. *Electron. J. Stat.* 12:2074–89
Kallus N. 2016. Recursive partitioning for personalization using observational data. arXiv:1608.08925 [stat.ML]
Kallus N. 2017. Balanced policy evaluation and learning. arXiv:1705.07384 [stat.ML]
Kallus N. 2018. Policy evaluation and optimization with continuous treatment. arXiv:1802.06037 [stat.ML]
Kang C, Janes H, Huang Y. 2014. Combining biomarkers to optimize patient treatment recommendations (with discussion and rejoinder). *Biometrics* 70:695–707
Kelleher SA, Dorfman CS, Vilardaga JCP, Majestic C, Winger J, et al. 2017. Optimizing delivery of a behavioral pain intervention in cancer patients using a sequential multiple assignment randomized trial SMART. *Contemp. Clin. Trials* 57:51–7
Kidwell KM. 2014. SMART designs in cancer research: past, present, and future. *Clin. Trials* 11:445–56
Kidwell KM. 2016. DTRs and SMARTs: definitions, designs, and applications. See Kosorok & Moodie 2016, pp. 7–24
Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, et al. 2015. Micro-randomized trials: an experimental design for developing just-in-time adaptive interventions. *Health Psychol.* 34(Suppl.):1220–28
Kosorok MR, Moodie EEM. 2016. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. Philadelphia: SIAM
Kravitz RL, Duan N, Braslow J. 2004. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q.* 82:661–87
Laber EB, Linn KA, Stefanski LA. 2014a. Interactive model building for Q-learning. *Biometrika* 101:831–47
Laber EB, Lizotte DJ, Ferguson B. 2014b. Set-valued dynamic treatment regimes for competing outcomes. *Biometrics* 70:53–61
Laber EB, Lizotte DJ, Qian M, Pelham WE, Murphy SA. 2014c. Dynamic treatment regimes: technical challenges and applications. *Electron. J. Stat.* 8:1225–72
Laber EB, Meyer NJ, Reich BJ, Pacifici K, Collazon JA, Drake J. 2018a. Optimal treatment allocations in space and time for on-line control of an emerging infectious disease. *J. R. Stat. Soc. C* 67:743–89
Laber EB, Murphy SA. 2011. Adaptive confidence intervals for the test error in classification (with discussion and rejoinder). *J. Am. Stat. Assoc.* 106:904–45
Laber EB, Staicu AM. 2018. Functional feature construction for personalized dynamic treatment regimes. *J. Am. Stat. Assoc.* 113:1219–27
Laber EB, Wu F, Munera C, Lipkovich I, Colucci S, Ripa S. 2018b. Identifying optimal dosage regimes under safety constraints: an application to long term opioid treatment of chronic pain. *Stat. Med.* 37:1407–18
Laber EB, Zhao YQ. 2015. Tree-based methods for individualized treatment regimes. *Biometrika* 102:501–14
Laber EB, Zhao YQ, Regh T, Davidian M, Tsiatis A, et al. 2016. Using pilot data to size a two-arm randomized trial to find a nearly optimal personalized treatment strategy. *Stat. Med.* 35:1245–56
Lakkaraju H, Rudin C. 2017. Learning cost-effective and interpretable treatment regimes. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ed. A Singh, J Zhu, pp. 166–75. Brookline, MA: Microtome
Lavori PW, Dawson R. 2000. A design for testing clinical strategies: biased adaptive within-subject randomization. *J. R. Stat. Soc. A* 163:29–38
Lavori PW, Dawson R. 2004. Dynamic treatment regimes: practical design considerations. *Clin. Trials* 1:9–20
Lee JA, Verleysen M. 2007. *Nonlinear Dimensionality Reduction*. New York: Springer
Li L. 2007. Sparse sufficient dimension reduction. *Biometrika* 94:603–13
Liang F, Li Q, Zhou L. 2018. Bayesian neural networks for selection of drug sensitive genes. *J. Am. Stat. Assoc.* 113:955–72

Linn KA, Laber EB, Stefanski LA. 2015. Estimation of dynamic treatment regimes for complex outcomes: balancing benefits and risks. See Kosorok & Moodie 2016, pp. 249–62
Linn KA, Laber EB, Stefanski LA. 2017. Interactive Q-learning for quantiles. *J. Am. Stat. Assoc.* 112:638–49
Liu Y, Wang Y, Zeng D. 2017. Sequential multiple assignment randomized trials with enrichment design. *Biometrics* 73:378–90
Lizotte DJ, Bowling M, Murphy SA. 2012. Linear fitted-Q iteration with multiple reward functions. *J. Mach. Learn. Res.* 13:3252–95
Lizotte DJ, Laber EB. 2016. Multi-objective Markov decision processes for data-driven decision support. *J. Mach. Learn. Res.* 17:7378–405
Lonergan M, Senn SJ, McNamee C, Daly AK, Sutton R, et al. 2017. Defining drug response for stratified medicine. *Drug Discov. Today* 22:173–79
Longford NT, Nelder JA. 1999. Statistics versus statistical science in the regulatory process. *Stat. Med.* 18:2311–20
Lu W, Zhang HH, Zeng D. 2013. Variable selection for optimal treatment decision. *Stat. Methods Med. Res.* 22:493–504
Luckett DJ, Laber EB, Kahkoska AR, Maahs DM, Mayer-Davis E, Kosorok MR. 2017. Estimating dynamic treatment regimes in mobile health using V-learning. arXiv:1611.03531v2 [stat.ML]
Luckett DJ, Laber EB, Kosorok MR. 2018. Estimation and optimization of composite outcomes. arXiv:1711.10581v2 [stat.ML]
Luedtke AR, van der Laan MJ. 2016a. Optimal individualized treatments in resource-limited settings. *Int. J. Biostat.* 12:283–303
Luedtke AR, van der Laan MJ. 2016b. Super-learning of an optimal dynamic treatment rule. *Int. J. Biostat.* 12:305–32
Lund JL, Richardson DB, Stürmer T. 2015. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. *Curr. Epidemiol. Rep.* 2:221–8
Macia NF, Thaler GJ. 2005. *Modeling and Control of Dynamic Systems*. New York: Thompson Delmar Learning
Moodie EEM, Chakraborty B, Kramer MS. 2012. Q-learning for estimating optimal dynamic treatment rules from observational data. *Can. J. Stat.* 40:629–45
Moodie EEM, Dean N, Sun YR. 2014. Q-learning: flexible learning about useful utilities. *Stat. Biosci.* 6:223–43
Moodie EEM, Richardson TS. 2010. Estimating optimal dynamic regimes: correcting bias under the null. *Scand. J. Stat.* 37:126–46
Murphy SA. 2003. Optimal dynamic treatment regimes (with discussion and rejoinder). *J. R. Stat. Soc. B* 65:331–66
Murphy SA. 2005a. A generalization error for Q-learning. *J. Mach. Learn. Res.* 6:1073–97
Murphy SA. 2005b. An experimental design for the development of adaptive treatment strategies. *Stat. Med.* 24:1455–81
Murphy SA, van der Laan MJ, Robins JM, Conduct Problems Prevention Research Group. 2001. Marginal mean models for dynamic regimes. *J. Am. Stat. Assoc.* 96:1410–23
Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, et al. 2017. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Ann. Behav. Med.* 52:446–62
Nisio M. 2015. *Stochastic Control Theory: Dynamic Programming Principle*. New York: Springer. 2nd ed.
Non-Hodgkin's Lymphoma Prognostic Factors Project. 1993. A prediction model for aggressive non-Hodgkin's lymphoma: The international non-Hodgkin's lymphoma prognostic factors project. *N. Engl. J. Med.* 329:987–94
Olsen KS, Lund E. 2017. Population-based precision cancer screening—Letter. *Cancer Epidemiol. Biomark. Prev.* 26:975
Paduraru C. 2013. *Off-policy evaluation in Markov decision processes*. PhD Thesis, McGill Univ.
Pearl J. 2009. Causal inference in statistics: an overview. *Stat. Surv.* 3:96–146
Penn State Methodology Center. 2018. Projects using SMARTs. *Penn State Methodology Center*. **https://methodology.psu.edu/ra/adap-inter/projects**
Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. 2000. Molecular portraits of human breast tumors. *Nature* 406:747–52

Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. 2012. Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* 21:31–54

Pires BA, Szepesvari C, Ghavamzadeh M. 2013. Cost-sensitive multiclass classification risk bounds. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, pp. 1391–9. Brookline, MA: Microtome

Powell WB. 2007. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. New York: Wiley

Precup D. 2000. *Temporal abstraction in reinforcement learning*. PhD Thesis, Univ. Mass. Amherst

Puterman ML. 2005. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: Wiley. 2nd ed.

Qian M, Murphy SA. 2011. Performance guarantees for individualized treatment rules. *Ann. Stat.* 39:1180–2011

Rich B, Moodie EE, Stephens DA, Platt RW. 2010. Model checking with residuals for g-estimation of optimal dynamic treatment regimes. *Int. J. Biostat.* 6(2): Article 12

Robins JM. 1986. A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the health worker survivor effect. *Comput. Math. Appl.* 7:1393–512

Robins JM. 1997. Causal inference from complex longitudinal data. *Lect. Notes Stat.* 120:69–117

Robins JM. 2004. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, ed. DY Lin, PJ Heagerty, pp. 189–326. New York: Springer

Robins JM, Hernan MA, Brumback B. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11:550–60

Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6:34–58

Rubin DB. 2005. Causal inference using potential outcomes: designs, modeling, decisions. *J. Am. Stat. Assoc.* 100:322–31

Sachs GS, Nierenberg AA, Calabrese JR, Marangell LB, Wisniewski SR, et al. 2007. Effectiveness of adjunctive antidepressant treatment for bipolar depression. *N. Engl. J. Med.* 356:1711–22

Schulte PJ, Tsiatis AA, Laber EB, Davidian M. 2014. Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Stat. Sci.* 29:640–61

Shani G, Pineau J, Kaplow R. 2013. A survey of point-based POMDP solvers. *Auton. Agents Multi-Agent Syst.* 27:1–51

Shen Y, Cai T. 2016. Identifying predictive markers for personalized treatment selection. *Biometrics* 72:1017–25

Shortreed SM, Laber EB, Lizotte DJ, Stroup TS, Pineau J, Murphy SA. 2011. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Mach. Learn.* 84:109–36

Shortreed SM, Laber EB, Scott ST, Pineau J, Murphy SA. 2014. A multiple imputation strategy for sequential multiple assignment randomized trials. *Stat. Med.* 33:4202–14

Shortreed SM, Moodie EE. 2012. Estimating the optimal dynamic antipsychotic treatment regime: evidence from the sequential multiple assignment randomized Clinical Antipsychotic Trials of Intervention and Effectiveness schizophrenia study. *J. R. Stat. Soc. C* 61:577–99

Si J, ed. 2004. *Handbook of Learning and Approximate Dynamic Programming*, Vol. 2. New York: Wiley

Song R, Kosorok MR, Zeng D, Zhao YQ, Laber EB, Yuan M. 2015a. On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat* 4:59–68

Song R, Wang W, Zeng D, Kosorok MR. 2015b. Penalized Q-learning for dynamic treatment regimens. *Stat. Sin.* 25:901–20

Sorensen TIA. 1996. Which patients may be harmed by good treatments? *Lancet* 348:351–2

Stephens DA. 2016. G-estimation for dynamic treatment regimes in the longitudinal setting. See Kosorok & Moodie 2016, pp. 89–117

Stusser RJ. 2006. Reflections on the scientific method in medicine. In *Medical Sciences*, Vol. 1, ed. YLG Verhasselt, AV Jablensky, A Pellegrini, C Sayegh, A Kazanjian, et al., pp. 23–47. Paris: UNESCO-EOLSS Joint Comm. Secr.

Sutton RS. 1997. On the significance of Markov decision processes. In *Proceedings of the 7th International Conference on Artificial Neural Networks*, pp. 273–82. New York: Springer

Sutton RS, Barto AG. 1998. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press

Szepesvari C. 2010. *Algorithms for Reinforcement Learning*. Williston, VT: Morgan & Claypool
Tang Y, Kosorok MR. 2012. *Developing adaptive personalized therapy for cystic fibrosis using reinforcement learning*. Work. Pap. 30, Dep. Biostat., Univ. N.C., Chapel Hill. **http://biostats.bepress.com/uncbiostat/art30**
Taylor JM, Cheng W, Foster JC. 2015. Reader reaction to "A robust method for estimating optimal treatment regimes" by Zhang et al. (2012). *Biometrics* 71:267–73
Teran Hidalgo SJ, Lawson MT, Luckett DJ, Chaudhari M, Chen J, et al. 2016. A master pipeline for discovery and validation of biomarkers. In *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*, ed. A Holzinger, pp. 259–88. New York: Springer
Tewari A, Murphy SA. 2017. From ads to interventions: contextual bandits in mobile health. In *Mobile Health—Sensors, Analytic Methods, and Applications*, ed. J Regh, S Murphy, S Kumar, pp. 495–517. New York: Springer
Thall PF, Leiko HW, Logothetis CJ, Millikan RE, Tannir NM. 2007. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Stat. Med.* 26:4687–702
Thiese MS. 2014. Observational and interventional study design types: an overview. *Biochem. Medica* 24:199–210
Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. 2014. A simple method for estimating interactions between a treatment and a large number of covariates. *J. Am. Stat. Assoc.* 109:1517–32
Torkamani A, Wineinger NE, Topol EJ. 2018. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19:581–90
van der Laan MJ, Murphy SA, Robins JM. 2001. *Analyzing dynamic regimes using structural nested mean models*. Unpublished manuscript
van der Laan MJ, Peterson ML. 2004. *History-adjusted marginal structural models and statistically-optimal dynamic treatment regimes*. Work. Pap. 158, Div. Biostat., Univ. Calif., Berkeley. **https://biostats.bepress.com/ucbbiostat/paper158**
Wager A, Athey S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113:1228–42
Wallace MP, Moodie EEM, Stephens DA. 2018. Reward ignorant modeling of dynamic treatment regimes. *Biom. J.* 60:991–1002
Wang J, Shen X, Pan W. 2009. On large margin hierarchical classification with multiple paths. *J. Am. Stat. Assoc.* 104:1213–23
Wang L, Laber EB, Witkiewitz K. 2018a. Sufficient Markov decision processes. arXiv:1704.07531v2 [stat.ME]
Wang L, Zhou Y, Song R, Sherwood B. 2018b. Quantile-optimal treatment regimes. *J. Am. Stat. Assoc.* 113:1243–54
Wang Y, Fu H, Zeng D. 2018c. 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.* 113:1–13
Wang Y, Wu P, Liu Y, Weng C, Zeng D. 2016. Learning optimal individualized treatment rules from electronic health record data. *IEEE Int. Conf. Healthc. Inform.* 2016:65–71
White House Office of the Press Secretary. 2015. *Fact sheet: President Obama's Precision Medicine Initiative*. Press release, Jan. 30
Wilensky U, Rand W. 2015. *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo*. Cambridge, MA: MIT Press
Wu T. 2016. *Set valued dynamic treatment regimes*. PhD Diss., Univ. Mich.
Xu Y, Müller P, Wahed AS, Thall PF. 2016. Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times (with discussion and rejoinder). *J. Am. Stat. Assoc.* 111:921–50
Zadrozny B, Langford J, Abe N. 2003. Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE International Conference on Data Mining*, pp. 435–42. New York: IEEE
Zatzick DF, Russo J, Darnell D, Chambers DA, Palinkas L, et al. 2016. An effectiveness-implementation hybrid trial study protocol targeting posttraumatic stress disorder and comorbidity. *Implement. Sci.* 11:58
Zhang B, Tsiatis BB, Davidian M, Zhang M, Laber EB. 2012a. Estimating optimal treatment regimes from a classification perspective. *Statistics* 1:103–14
Zhang B, Tsiatis AA, Laber EB, Davidian M. 2012b. A robust method for estimating optimal treatment regimes. *Biometrics* 68:1010–18

Zhang B, Tsiatis AA, Laber EB, Davidian M. 2013. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* 100:681–94

Zhang B, Zhang M. 2016. *Variable selection for estimating the optimal treatment regimes in the presence of a large number of covariate*. Tech. Rep. 120, Sch. Public Health, Univ. Mich.

Zhang Y, Laber EB, Davidian M, Tsiatis AA. 2018. Estimation of optimal treatment regimes using lists. *J. Am. Stat. Assoc.* In press

Zhao YF, Kosorok MR, Zeng D. 2009. Reinforcement learning design for cancer clinical trials. *Stat. Med.* 28:3294–315

Zhao YQ, Zeng D, Laber EB, Kosorok MR. 2015a. New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Am. Stat. Assoc.* 110:583–98

Zhao YQ, Zeng D, Laber EB, Song R, Yuan M, Kosorok MR. 2015b. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika* 102:151–68

Zhao YQ, Zeng D, Rush AJ, Kosorok MR. 2012. Estimating individualized treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.* 107:1106–18

Zhao YF, Zeng D, Socinski MA, Kosorok MR. 2011. Reinforcement learning strategies for clinical trials in non-small cell lung cancer. *Biometrics* 67:1422–33

Zhou ZH, Liu XY. 2006. On multi-class cost-sensitive learning. In *Proceedings of the 21st National Conference on Artificial Intelligence*, Vol. 1, pp. 567–72. Palo Alto, CA: AAAI

Zhou X, Mayer-Hamblett N, Khan U, Kosorok MR. 2017. Residual weighted learning for estimating individualized treatment rules. *J. Am. Stat. Assoc.* 112:169–87

Annual Review of Statistics and its Application

Volume 6, 2019

# Contents

**Errata**

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics