

# Dynamic Treatment Regimes

Bibhas Chakraborty<sup>1</sup> and Susan A. Murphy<sup>2</sup>

<sup>1</sup>Duke-NUS Graduate Medical School, National University of Singapore, Singapore 169857; email: bibhas.chakraborty@duke-nus.edu.sg

<sup>2</sup>Department of Statistics and Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48109; email: samurphy@umich.edu

Annu. Rev. Stat. Appl. 2014. 1:447–64

First published online as a Review in Advance on August 26, 2013

The *Annual Review of Statistics and Its Application* is online at [statistics.annualreviews.org](http://statistics.annualreviews.org)

This article's doi:  
10.1146/annurev-statistics-022513-115553

Copyright © 2014 by Annual Reviews.  
All rights reserved

## Keywords

dynamic treatment regime, reinforcement learning, sequential randomization, nonregularity, Q-learning

## Abstract

A dynamic treatment regime consists of a sequence of decision rules, one per stage of intervention, that dictate how to individualize treatments to patients, based on evolving treatment and covariate history. These regimes are particularly useful for managing chronic disorders and fit well into the larger paradigm of personalized medicine. They provide one way to operationalize a clinical decision support system. Statistics plays a key role in the construction of evidence-based dynamic treatment regimes—informing the best study design as well as efficient estimation and valid inference. Owing to the many novel methodological challenges this area offers, it has been growing in popularity among statisticians in recent years. In this article, we review the key developments in this exciting field of research. In particular, we discuss the sequential multiple assignment randomized trial designs, estimation techniques like Q-learning and marginal structural models, and several inference techniques designed to address the associated nonstandard asymptotics. We reference software whenever available. We also outline some important future directions.

# 1. INTRODUCTION

Personalized medicine is an increasingly popular theme in today's health care. Operationally, personalized treatments are decision rules that dictate what treatment to provide given a patient state (consisting of, e.g., demographics, results of diagnostic tests, and genetic information). Dynamic treatment regimes (DTRs) (Murphy 2003; Robins 1986, 1989, 1993, 1997, 2004) generalize personalized medicine to time-varying treatment settings in which treatment is repeatedly tailored to a patient's time-varying—or dynamic—state. DTRs are alternatively known as adaptive treatment strategies (Lavori & Dawson 2000, 2008; Murphy 2005b; Thall et al. 2000, 2002) or treatment policies (Lunceford et al. 2002; Wahed & Tsiatis 2004, 2006). These decision rules offer an effective vehicle for personalized management of chronic conditions (e.g., alcohol and drug abuse, cancer, diabetes, HIV infection, and mental illnesses), for which a patient typically has to be treated at multiple stages, and help clinicians adapt the treatment (type, dosage, and timing) at each stage to the evolving treatment and covariate history. DTRs underpin clinical decision support systems, which represent a key element of the chronic care model (Wagner et al. 2001).

A simple example of a DTR arising in the treatment of alcohol dependence is this: After the patient completes an intensive outpatient program, provide the medication naltrexone (NTX) along with face-to-face medical management. If within the following two months the patient experiences two or more heavy-drinking days, then immediately augment the NTX with cognitive behavioral therapy (CBT). Otherwise, at the end of the two months, provide telephone disease management in addition to the NTX. Rosthøj et al. (2006) give an example of a DTR used in guiding warfarin dosing to control the risk of both clotting and excessive bleeding. Here, the decision rules input summaries of the trajectory of the international normalized ratio (a measure of clotting tendency of blood) over the recent past and output recommendations concerning how much to change the dose of warfarin (if at all). Robins et al. (2008) provide a DTR example also, this one concerning decision rules that input summaries of the trajectories of plasma HIV RNA and CD4 counts over the recent past and output when to start an asymptomatic HIV-infected subject on highly active antiretroviral therapy. In Section 3, we review different statistical methods for constructing the decision rules in a DTR.

Traditionally, personalized medicine concerns single-stage decision making. In a single-stage (nondynamic) decision problem, one observes a random vector, the first observation,  $\mathbf{O}_1$ ; then one selects an action (here a treatment action),  $a_1$ , from a set  $\mathcal{A}_1$  of actions. Then, depending on which action was selected, one makes a second observation,  $\mathbf{O}_2(a_1)$ . To avoid technical details and for simplicity, we assume sufficient regularity for all statements here and below. A decision rule, e.g.,  $d_1$ , is a mapping from the range of  $\mathbf{O}_1$  into  $\mathcal{A}_1$ . The quality of a treatment for a particular value of  $\mathbf{O}_1$  is evaluated in terms of its utility, e.g.,  $r(\mathbf{O}_1, a_1, \mathbf{O}_2(a_1))$ , for a known function  $r$ . The utility may be a summary of one outcome, such as percent days abstinent in an alcohol dependence study, or a composite outcome; for example, in Wang et al. (2012), the utility is a compound score numerically combining information on treatment efficacy, toxicity, and the risk of disease progression. The optimal decision rule outputs the treatment (action) that maximizes the expected utility,  $\mathcal{U}(o_1; a_1) = E[r(\mathbf{O}_1, a_1, \mathbf{O}_2(a_1)) \mid \mathbf{O}_1 = o_1]$ ; because the expected utility depends on  $o_1$ , the treatment that maximizes the expected utility may depend on  $o_1$  and thus provide a personalized treatment decision. Equivalently, the optimal decision rule is given by  $\arg \max_{d_1} E[\mathcal{U}(\mathbf{O}_1; d_1(\mathbf{O}_1))]$ , where the maximum is taken over all functions on the range of  $\mathbf{O}_1$ .  $E[\mathcal{U}(\mathbf{O}_1; d_1(\mathbf{O}_1))]$  is called the value of the decision rule,  $d_1$ .

Constructing DTRs involves solving, or estimating quantities relevant in, a multistage decision problem. In multistage decision problems, observations are interwoven with action selection; we can denote such a sequence by  $\mathbf{O}_1, a_1, \mathbf{O}_2(a_1), a_2, \mathbf{O}_3(\bar{a}_2), \dots, \mathbf{O}_K(\bar{a}_{K-1}), a_K, \mathbf{O}_{K+1}(\bar{a}_K)$ , where

$\bar{a}_j = \{a_1, \dots, a_j\}$ , and  $\mathbf{O}_{j+1}(\bar{a}_j)$  denotes the observation made at stage  $j + 1$  subsequent to the selection of the action sequence  $\bar{a}_j$ . A DTR is a sequence of decision rules,  $\bar{d}_K = (d_1, \dots, d_K)$ ; the decision rule  $d_j$  is a mapping from the range of  $(\mathbf{O}_1, a_1, \dots, \mathbf{O}_j(\bar{a}_{j-1}))$  into the  $j$ th action space,  $\mathcal{A}_j$ . When  $K = 2$  and the treatment actions are discrete, the value of the DTR  $(d_1, d_2)$  can be written on one line as

$$E \left[ \sum_{a_1 \in \mathcal{A}_1} 1_{a_1=d_1(\mathbf{O}_1)} \sum_{a_2 \in \mathcal{A}_2} 1_{a_2=d_2(\mathbf{O}_1, a_1, \mathbf{O}_2(a_1))} r(\mathbf{O}_1, a_1, \mathbf{O}_2(a_1), a_2, \mathbf{O}_3(a_1, a_2)) \right]. \quad 1.$$

(The generalization to more than two stages is straightforward.) Using this formula, we might compare two or more DTRs in terms of their value or, equivalently, their expected utility. The optimal DTR is the set of decision rules,  $\bar{d}_K$ , that maximizes the value.

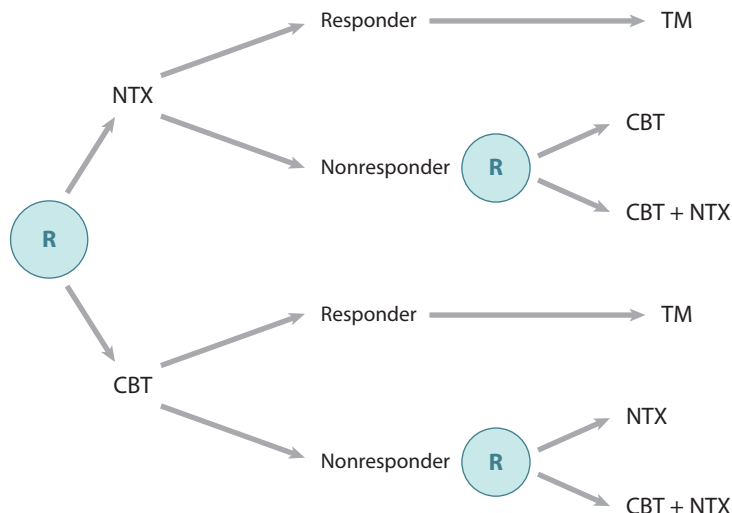
Constructing the optimal decision rules in multistage decision problems is challenging because of the time-varying or dynamic nature of this problem. Historically, an early method for solving (e.g., constructing the optimal decision rules) multistage decision problems is dynamic programming (DP), which dates at least as far back as Bellman (1957). The primary reason why classical DP algorithms have seen little use in DTR research is that these algorithms require complete knowledge of, or a full model for, the multivariate distribution of the data for any set of actions; this requirement is impractical in many application areas (the curse of modeling) (Kulkarni et al. 2011). Moreover, DP methods are computationally very expensive, and they become hard to manage in moderately high-dimensional problems; in other words, they suffer from the curse of dimensionality (Sutton & Barto 1998). But DP provides an important theoretical and conceptual foundation for research in multistage decision problems. In fact, as we illustrate below, many present-day estimation methods build on classical DP algorithms but relax their stringent requirements.

## 2. DATA SOURCES FOR CONSTRUCTING DTRs

Most statistical research in the arena of DTRs concerns (a) the comparison of two or more preconceived DTRs in terms of their value and (b) the estimation of the optimal DTR, i.e., estimating the sequence of decision rules, one per stage, that results in the highest value, within a class of DTRs. In each case, the data used in comparing or constructing DTRs are usually from (a) sequentially randomized studies, (b) longitudinal observational studies, or (c) dynamical system models. Research based on the first source of data, i.e., sequentially randomized studies, is experiencing a period of rapid growth as a result of the increasing number of clinical trials in which many of the patients are randomized multiple times, in a sequential manner. However, by far the majority of statistical research, led by Robins's (1986, 1989, 1993, 1997) pioneering work, concerns the use of data from longitudinal, observational studies. The third data source, based on simulating from or otherwise using existing dynamical system models, has received much less attention in DTR development. In this section, we briefly review the first two types of data sources, their advantages and drawbacks, and the assumptions required to perform valid analyses in each, along with some examples. Dynamical system models are discussed in Section 3.

### 2.1. Sequential Multiple Assignment Randomized Trials

Beginning with Robins's (1986, 1989, 1993, 1997) work, sequentially randomized trials were used as a conceptual tool to precisely state the inferential goals in DTR research. More recently, trial designs, known as sequential multiple assignment randomized trial (SMART) designs (Lavori & Dawson 2000, 2004; Murphy 2005b), have been implemented in practice. SMART designs involve



**Figure 1**

Hypothetical sequential multiple assignment randomized trial (SMART) design schematic for the addiction management example (an R within a circle denotes randomization at a critical decision point). Abbreviations: CBT, cognitive behavioral therapy; NTX, naltrexone; TM, telephone monitoring.

an initial randomization of patients to available treatment actions, followed by rerandomizations (of some or all of the patients) at each subsequent stage to treatment actions available at that stage. The rerandomizations and set of treatment actions at each subsequent stage may depend on information collected in prior stages, such as how well the patient responded to the previous treatment.

Recent SMARTs include a smoking cessation study (Chakraborty et al. 2010), a study involving treatment of autism among children (Kasari 2009, Lei et al. 2011), a study involving interventions for children with attention deficit hyperactivity disorder (Nahum-Shani et al. 2012a,b), a study involving treatment for pregnant drug abusers (Jones 2010, Lei et al. 2011), and a study involving alcohol-dependent individuals (Lei et al. 2011). For a list of some other SMARTs, we refer the reader to the following website: <http://methodology.psu.edu/ra/adap-treat-strat/projects>.

To make the discussion more concrete, **Figure 1** shows a hypothetical SMART design based on the addiction management example introduced earlier. In this trial, each participant is randomly assigned to one of two possible initial treatments: CBT or NTX. Participants are classified as nonresponders or responders to the initial treatment according to whether they do or do not experience more than two heavy-drinking days during the next two months. A nonresponder to NTX is rerandomized to one of the two subsequent treatment options: either a switch to CBT or an augmentation of NTX with CBT (CBT + NTX). Similarly, a nonresponder to CBT is rerandomized to either a switch to NTX or an augmentation (CBT + NTX). Responders to the initial treatment receive telephone monitoring (TM) for an additional six months. One goal of the study might be to construct a DTR leading to a maximal mean number of non-heavy-drinking days over 12 months.

We denote the observable data trajectory for a participant in a two-stage SMART by  $(\mathbf{O}_1, A_1, \mathbf{O}_2, A_2, \mathbf{O}_3)$ , where  $\mathbf{O}_1$ ,  $\mathbf{O}_2$ , and  $\mathbf{O}_3$  are the pretreatment information, intermediate outcomes, and final outcomes, respectively. The randomized treatment actions are  $A_1$  and  $A_2$ , and

the primary outcome is  $Y = r(\mathbf{O}_1, A_1, \mathbf{O}_2, A_2, \mathbf{O}_3)$  for a known function  $r$ . For example, in the addiction management study above,  $\mathbf{O}_1$  may include addiction severity and comorbid conditions;  $\mathbf{O}_2$  may include the participant's binary response status, side effects, and adherence to the initial treatment; and  $Y$  may be the number of non-heavy-drinking days over the 12-month study period.

To connect the distribution of the data collected in the above SMART to the distributions considered in the multistage decision problem in Section 1, we make a short digression into the field of causal inference. Recall that, in the case of two stages (Section 1), we denoted the sequence of random observations by  $(\mathbf{O}_1, a_1, \mathbf{O}_2(a_1), a_2, \mathbf{O}_3(a_1, a_2))$  for the selected actions  $(a_1, a_2)$ . These observations are potential outcomes (Robins 1986, Rubin 1974). Potential outcomes or counterfactual outcomes are defined as a participant's outcome had he or she followed a particular treatment (sequence), which is possibly different from the treatment (sequence) he or she was actually observed to follow. Consider, for example, a single-stage randomized trial in which participants can receive either  $a$  or  $a'$ . Accordingly, any participant in this study is conceptualized as having two potential second observations,  $\mathbf{O}_2(a)$  and  $\mathbf{O}_2(a')$ . However, only one of these—the one corresponding to the treatment to which a participant is randomized—will be observed. Clearly, observing the  $\mathbf{O}_2$  under both treatments  $a$  and  $a'$  is impossible without further data and assumptions (e.g., in a crossover trial with no carryover effect). Now suppose that participants are treated over two stages and can receive at each stage either  $a$  or  $a'$  ( $\mathcal{A}_1 = \mathcal{A}_2 = \{a, a'\}$ ). In this case, four sequences of potential observations exist,  $(\mathbf{O}_2(a), \mathbf{O}_3(a, a))$ ,  $(\mathbf{O}_2(a), \mathbf{O}_3(a, a'))$ ,  $(\mathbf{O}_2(a'), \mathbf{O}_3(a', a))$ , and  $(\mathbf{O}_2(a'), \mathbf{O}_3(a', a'))$ ; only one of these sequences will be observed for any given participant.

To connect the potential observations to the observations made during a SMART, we make two assumptions (Robins 1997):

1. Consistency: The potential outcome under the observed treatment and the observed outcome agree.
2. No unmeasured confounders: For any treatment sequence  $\bar{a}_K$ , and conditional on the history  $H_j = (\bar{\mathbf{O}}_j, \bar{A}_{j-1})$ , treatment  $A_j$  is independent of future (potential) outcomes,  $\mathbf{O}_{j+1}(\bar{a}_j), \dots, \mathbf{O}_K(\bar{a}_{K-1}), \mathbf{O}_{K+1}(\bar{a}_K)$ . That is, for any possible treatment sequence  $\bar{a}_K$ ,

$$A_j \perp (\mathbf{O}_{j+1}(\bar{a}_j), \dots, \mathbf{O}_K(\bar{a}_{K-1}), \mathbf{O}_{K+1}(\bar{a}_K)) \mid H_j \quad \forall j = 1, \dots, K.$$

The consistency assumption subsumes Rubin's (1980) more explanatory stable unit treatment value assumption (SUTVA), which is this: Each participant's potential outcome is not influenced by the treatment applied to the other participants. In clinical trials, SUTVA is most often violated when the treatment is not well defined. For example, the treatment as defined may not specify that some aspects of the treatment are provided in a group setting containing multiple participants from the trial. In this case, the response of one participant to treatment may influence the response of another participant if they are in the same group.

Under the consistency assumption, the potential outcomes in a two-stage SMART are connected to the observable data by  $\mathbf{O}_2 = \mathbf{O}_2(A_1)$  and  $\mathbf{O}_3 = \mathbf{O}_3(A_1, A_2)$ . The no-unmeasured-confounders assumption holds in a SMART design if the randomization probabilities depend at most on the past observations; more precisely, the randomization probabilities for  $A_1$  and  $A_2$  may depend on  $\mathbf{O}_1$  and  $(\mathbf{O}_1, A_1, \mathbf{O}_2)$ , respectively. Under this assumption,

$$P(\mathbf{O}_2(a_1) \leq o_2 \mid \mathbf{O}_1 = o_1) = P(\mathbf{O}_2 \leq o_2 \mid \mathbf{O}_1 = o_1, A_1 = a_1), \text{ and}$$

$$P(\mathbf{O}_3(a_1, a_2) \leq y \mid \mathbf{O}_1 = o_1, \mathbf{O}_2(a_1) = o_2) = P(\mathbf{O}_3 \leq y \mid \mathbf{O}_1 = o_1, A_1 = a_1, \mathbf{O}_2 = o_2, A_2 = a_2).$$

The above probability statements imply that the value for a DTR can be written as a function of the multivariate distribution of the observable data obtained from a SMART. In the case of two

stages, Equation 1 can be written as

$$E \left[ E \left[ \sum_{a_1 \in \mathcal{A}_1} 1_{a_1=d_1(H_1)} E \left[ \sum_{a_2 \in \mathcal{A}_2} 1_{a_2=d_2(H_2)} E[r(H_2, A_2, \mathbf{O}_3) | H_2, A_2 = a_2] \right] | H_1, A_1 = a_1 \right] \right]$$

[recall that  $H_1 = \mathbf{O}_1$  and  $H_2 = (\mathbf{O}_1, A_1, \mathbf{O}_2)$ ]. A similar result holds for settings with more than two stages. Thus, the validity of the two assumptions ensures that data from SMARTs can be effectively used to evaluate prespecified DTRs or to estimate the optimal DTR within a certain class.

**2.1.1. Some practical considerations in designing a SMART.** Many authors recommend that the design of a SMART be no more complicated than necessary. Indeed, the class of treatment options at each stage should not be unnecessarily restricted (Lavori & Dawson 2004, Murphy 2005b). For example, using a low-dimensional summary criterion (e.g., responder/nonresponder status, as used in the addiction management SMART example) to restrict the class of possible treatments is preferable to using all intermediate outcomes (e.g., improvement of symptom severity, side effects, adherence). Furthermore, a SMART is best viewed as one trial among a series of randomized trials intended to develop and/or refine a DTR. It should eventually be followed by a confirmatory randomized trial that compares the developed regime and an appropriate control (Murphy 2005b). That is, the construction of DTRs is developmental as opposed to confirmatory. In this sense, a scientist employing a SMART design has a similar goal to Box et al.'s (1978) goal of developing multicomponent treatments. Indeed, the SMART can be viewed as an extension of the factorial design to a setting in which the time and sequencing of treatments play crucial roles (Murphy & Bingham 2009). As a result, the primary hypothesis, i.e., the hypothesis used to determine the sample size for the trial, often concerns a main effect. However, because of the multiple randomizations, we can consider many interesting secondary research questions with randomized data, though the SMART may or may not have enough power to address these secondary hypothesis questions.

Most often, the primary hypothesis concerns the main effect of the first-stage treatment. For example, in the addiction management study, an interesting primary research question might ask what the best initial treatment would be on average if we marginalized over secondary treatments. In other words, here the researcher wants to compare the mean primary outcome of the patients receiving NTX as the initial treatment with the mean primary outcome of those receiving CBT. Another interesting primary question could concern the main effect of a second-stage treatment: On average, what is the best secondary treatment, a switch or an augmentation, for nonresponders to initial treatment? Here, the researcher might compare the mean primary outcome of nonresponders assigned to switch with the mean primary outcome of nonresponders assigned to augmentation. In all of these cases, sample size formulae are standard or easily derived.

Alternatively, the primary research question may concern the comparison of two of the embedded DTRs. In the example of the addiction management SMART, four embedded DTRs exist, corresponding to two options for the first-stage treatment and two options for the second-stage treatment for nonresponders (only one option exists for the responders). For example, one embedded regime in this SMART is to treat the patient with NTX at stage 1, to give TM at stage 2 if the patient is a responder, and to give CBT at stage 2 if the patient is a nonresponder; other embedded regimes can be described similarly. Dawson & Lavori (2010, 2012), Murphy (2005b), and Oetting et al. (2011) consider how to determine appropriate sample sizes to compare two embedded DTRs in the context of a continuous outcome. A web application that calculates the required sample size for a SMART design for a continuous

endpoint can be found at <http://methodologymedia.psu.edu/smart/samplesize>. Much work has concerned survival endpoints (Feng & Wahed 2008; Lunceford et al. 2002; Wahed & Tsiatis 2004, 2006). Relevant sample size formulae can be found in Feng & Wahed (2009) and Li & Murphy (2011). A web application for sample size calculation in this case can be found at <http://methodologymedia.psu.edu/logranktest/samplesize>.

**2.1.2. SMART versus other designs.** The SMART design discussed above involves stages of treatment and/or experimentation. In this regard, it bears superficial similarity with adaptive designs (Coffey et al. 2012). Adaptive design is an umbrella term used to denote different trial designs that allow certain trial features to change based on accumulating data while maintaining the statistical, scientific, and ethical integrity of the trial (Coffey et al. 2012). In a SMART design, each participant moves through multiple stages of treatment, whereas in adaptive designs, each stage involves different participants. The goal of a SMART is to develop a good DTR that could benefit future patients. Many adaptive designs try to provide the most efficacious treatment to each patient in the trial based on the current knowledge available at the time that a participant is randomized. In a SMART, unlike in an adaptive design, design elements such as the final sample size, randomization probabilities, and treatment options are prespecified. SMART designs involve within-participant adaptation of treatment, whereas adaptive designs involve between-participant adaptation. Although in some settings the incorporation of adaptive elements into a SMART design is possible (Thall et al. 2002, Thall & Wathen 2005), how to achieve optimal incorporation is an open question that warrants further research.

SMART designs have some operational similarity with classical crossover trial designs; however, they differ greatly in the scientific goal. In particular, a crossover design is typically used to contrast the effects of stand-alone treatments, whereas a SMART is used to develop a DTR, i.e., a sequence of treatments. Treatment allocation at any stage after the initial stage of a SMART typically depends on a participant's intermediate outcome (response/nonresponse). However, in a crossover trial, participants receive all the candidate treatments irrespective of their intermediate outcomes. And most importantly, an attempt to wash out the carryover effect is crucial in a crossover trial, whereas the process of constructing a DTR involves harnessing carryover effects so as to improve outcomes. That is, carryover effects such as synergistic interactions between treatments at different stages may lead to a better DTR, as compared with a DTR in which no carryover effects exist.

## 2.2. Observational Studies

In observational studies, the treatments are not randomized. In particular, we do not know with certainty the reasons why different individuals receive differing treatments or the reasons why one individual receives different treatments at different times. Indeed, data in which the treatments are (sequentially) randomized, when available, are preferable for making inferences concerning DTRs. However, observational studies are the most common source of data for constructing DTRs, and most research in statistics has concentrated on how best to use observational data.

In observational data, associations observed in the data (e.g., between treatment and outcome) may partially stem from the unobserved or unknown reasons why individuals receive differing treatments, as opposed to stemming from the effects of the treatments. Thus, to conduct inference, one requires assumptions. Assumptions such as the consistency and the no-unmeasured-confounders assumptions discussed earlier can be used to justify estimation and inference based on observational data; the plausibility of these assumptions is generally best justified by scientific,



expert knowledge. Researchers have undertaken many studies aimed at constructing DTRs from observational data. Data sources include hospital databases (Cotton & Heagerty 2011, Orellana et al. 2010a, Robins et al. 2008, Rosthøj et al. 2006), randomized encouragement trials (Moodie et al. 2009), and cohort studies (van der Laan & Petersen 2007b).

The assumption of no unmeasured confounders deserves careful consideration and thought in the observational data setting. The assumption states that, conditional on the past history, treatment received at stage  $j$  is independent of future potential observations and outcome:  $P(A_j = a \mid H_j, \mathbf{O}_{j+1}(\bar{a}_j), \dots, \mathbf{O}_K(\bar{a}_{K-1}), \mathbf{O}_{K+1}(\bar{a}_K)) = P(A_j = a \mid H_j)$ . This assumption allows us to effectively view the observational data as coming from a sequentially randomized trial, albeit with unknown as opposed to known randomization probabilities at stage  $j$ . The assumption may be (approximately) true in observational settings where all relevant common causes of outcomes and treatment have been observed.

In addition to careful consideration of causal inference issues, using observational data to construct DTRs requires careful thought concerning how the data may restrict the set of DTRs that can be assessed, absent further assumptions. This set is called the feasible (Robins 1994) or viable (Wang et al. 2012) DTRs. Feasibility of a DTR  $\bar{d}_K$  requires a positive probability that some participants in the study will have followed  $\bar{d}_K$ .

### 3. DATA ANALYSIS

As mentioned in Section 2, two common goals are (a) the estimation/comparison of a few DTRs in terms of their value and (b) the estimation of the optimal DTR within a certain class. In this section, we review the analysis strategies for both. Throughout, we assume that both the no-unmeasured-confounders and consistency assumptions hold and that all DTRs considered are feasible.

Weighting is often used to address both goals. Weights, or inverse probability of treatment weights (IPTWs), were originally developed to estimate the value of nondynamic regimes (Robins 1999, Robins et al. 2000) but later were adapted to the problem of estimating the value of DTRs. Murphy et al. (2001) and Wang et al. (2012) use IPTWs to estimate the values of a few DTRs. To see why weights might be used, consider a SMART like that in **Figure 1**, which gives only one option for responding participants (e.g., TM). Suppose that the treatment assignment probabilities at stage 1 and the treatment assignment probabilities for the nonresponders are uniform (randomization probability is 0.5). Now, suppose we want to estimate the value of the embedded DTR: Treat the patient with NTX at stage 1, give TM at stage 2 if the patient is a responder, and give CBT at stage 2 if the patient is a nonresponder. To estimate the value, we utilize the outcome of all participants with treatment patterns consistent with this DTR. However, within this group of participants, responders are overrepresented compared with nonresponders because the nonresponders were subdivided in the trial, whereas the responders were not. The IPTWs are used to adjust for overrepresentation of participants across the treatment patterns consistent with a given DTR. In this example, data from responders would have a weight of  $1/0.5$ , as responders were randomized only in stage 1 (with a probability of 0.5), whereas data from nonresponders would have a weight of  $1/(0.5)(0.5)$ , as they have been randomized twice (each with a probability of 0.5). We refer the reader to Nahum-Shani et al. (2012a) and Wang et al. (2012) for detailed explanations of how IPTWs can account for this over- and underrepresentation in SMARTs. Lunceford et al. (2002), Miyahara & Wahed (2010), and Wahed & Tsiatis (2004, 2006) use IPTWs in estimating the value of DTRs in the survival analysis setting. Improved versions of the IPTW estimator are available in papers by Robins and colleagues (Murphy et al. 2001; Orellana et al. 2010a,b; Robins et al. 2008) and Zhang et al. (2012b).



### 3.1. Direct Methods for Estimating an Optimal DTR

For notational simplicity, let  $d$  denote the DTR,  $\bar{d}_K$ , in the following. Recall from Section 1 that the value of a DTR is the mean of the utility, marginalized over all observations that might be affected by the treatment. In direct methods, one specifies a class of DTRs  $\mathcal{D}$  (see below for an example), estimates the value for each candidate DTR  $d \in \mathcal{D}$ , e.g.,  $\hat{V}^d$ , and then selects the DTR in  $\mathcal{D}$  with maximal estimated value.

Robins and colleagues (Orellana et al. 2010a, Robins et al. 2008) pioneered the use of IPTWs for estimating an optimal DTR. For a simple example, we consider DTRs that use a risk score to indicate when to initiate treatment. At the clinic visit at which the risk score is greater than or equal to  $x$ , treatment is initiated. The value varies by DTR, i.e., by  $x$ . In Robins et al. (2008), the value is parameterized as a polynomial function in  $x$  and in pretreatment variables, e.g.,  $V(x; \beta) = \beta_0 + \beta_1 x + \beta_2 x^2$ . The optimal DTR is to initiate treatment when the risk score is greater than or equal to  $x_0$ , where  $x_0 = \arg \max_x V(x, \beta)$ . To estimate the optimal DTR, we need estimators of the  $\beta$ s. In the simplest setting, we estimate the  $\beta$ s by solving an inverse probability of treatment-weighted estimating equation. To improve efficiency in the estimation of the  $\beta$ s, Robins et al. (2008) take advantage of the fact that some individuals' treatment sequences will be consistent with more than one DTR. For example, if the individual initiates treatment with a risk score of 12, and at the prior office visits, the individual's risk score was always lower than 10, then this individual has a treatment sequence consistent with  $x = 10, 11$ , and 12. To improve efficiency, this individual's data are used to estimate the value,  $V(x; \beta)$ , for  $x = 10, 11$ , and 12. Operationally, the estimating equation uses three replicates of this individual's data. In the above example, the individual's data are copied twice to produce three replicates, and the replicated outcome  $Y$  is relabeled as  $Y_{10}$ ,  $Y_{11}$ , and  $Y_{12}$  ( $Y_{10} = Y_{11} = Y_{12}$ ). In general, the number of replicates of an individual's data is equal to the number of DTRs with which his or her observed treatment is consistent.

We can estimate the  $\beta$ s by solving

$$0 = \mathbb{P}_n \left[ \sum_x w_{d^x, \pi} \cdot \frac{\partial}{\partial \beta} V(x; \beta) (Y_x - V(x; \beta)) \right],$$

where  $\mathbb{P}_n$  is an average over the augmented data set (containing the replicates). Nahum-Shani et al. (2012a), in the context of SMART, provide an intuitive discussion of why replication of participants can be used to account for a participant's observed treatment being consistent with more than one DTR. The observational data setting can be more complicated (see Robins et al. 2008 and Shortreed & Moodie 2012 for detailed expositions). Related work that compares a range of candidate DTRs by incorporating a treatment-tailoring threshold can be found in Cotton & Heagerty (2011), Hernán et al. (2006), Petersen et al. (2007), and van der Laan & Petersen (2007a).

Direct methods for a one-stage decision-making setting (e.g.,  $K = 1$ ) have seen a great deal of research; here, the single-decision rule is often called an individualized decision rule. As Qian & Murphy (2011) highlight, the one-stage decision-making problem has a close connection with classification. Subsequently, researchers have proposed methods based on classification for estimating the decision rule (Zhang et al. 2012a, Zhao et al. 2012). Other work in the one-stage decision setting includes Cai et al. (2011) and Imai & Ratkovic (2013).

### 3.2. Indirect Methods for Estimating an Optimal DTR

Indirect approaches to estimating the optimal DTR are commonly employed when scientists wish to consider decision rules that may depend on multiple covariates or depend on covariates in a complex manner. In the indirect approach, the stage-specific conditional mean outcomes (called

Q-functions) or contrasts thereof are modeled first, and then the optimal decision rules are found via maximization of these estimated conditional means or contrasts. These methods were originally developed in the reinforcement learning literature within computer science but were later adapted to statistics. One such procedure that has become particularly popular in the DTR literature is Q-learning (Sutton & Barto 1998). Q-learning is an approximate DP method—approximate because data and models are used to approximate the Q-functions. In its simplest incarnation, Q-learning uses linear models for the Q-functions and can be viewed as an extension of least squares regression to multistage decision problems (Murphy 2005a). However, one can use more flexible models for the Q-functions, e.g., regression trees (Ernst et al. 2005) or kernels (Ormoneit & Sen 2002). The version of Q-learning considered in the DTR literature is most similar to the fitted Q-iteration algorithm (Ernst et al. 2005) in the reinforcement learning literature.

**3.2.1. Q-learning with linear models.** For clarity, here we define Q-functions and describe Q-learning for studies with only two stages; generalization to  $K(\geq 2)$  stages is straightforward (Murphy 2005a). For simplicity, assume that the data come from a SMART with two possible treatments at each stage,  $A_j \in \{-1, 1\}$ , and that the treatment is randomized with known randomization probabilities. The data from a SMART involving  $n$  subjects will consist of  $n$  data trajectories of the form  $(\mathbf{O}_1, A_1, \mathbf{O}_2, A_2, \mathbf{O}_3)$ . As before, the histories are defined as  $H_1 = \mathbf{O}_1$  and  $H_2 = (\mathbf{O}_1, A_1, \mathbf{O}_2)$ . The study can have either a single terminal utility (primary outcome),  $Y$ , observed at the end of stage 2, or two stage-specific utilities,  $Y_1$  and  $Y_2$ , adding up to the primary outcome,  $Y = Y_1 + Y_2$  (in general,  $Y$  can be any known function of the data). The interest lies in estimating a two-stage DTR  $(d_1, d_2)$ , with  $d_j(H_j) \in \{-1, 1\}$ .

The optimal Q-functions for the two stages are defined as  $Q_2(H_2, A_2) = E[Y_2 \mid H_2, A_2]$  and  $Q_1(H_1, A_1) = E[Y_1 + \max_{a_2} Q_2(H_2, a_2) \mid H_1, A_1]$ . We can use a backward induction argument (Sutton & Barto 1998) to prove that the optimal treatment at a particular stage is given by the value of the action that maximizes the associated Q-function. In particular, if these two Q-functions were known, the optimal DTR  $(d_1, d_2)$  would be  $d_j(b_j) = \arg \max_{a_j} Q_j(b_j, a_j)$ ,  $j = 1, 2$ . In practice, the true Q-functions are unknown and hence must be estimated. Because Q-functions are conditional expectations, a natural approach to modeling them is via regression models. A DP (moving backward through the stages) approach is used to estimate the parameters. Consider linear regression models for the Q-functions. Let the stage  $j$  ( $j = 1, 2$ ) Q-function be modeled as  $Q_j(H_j, A_j; \beta_j, \psi_j) = \beta_j^T H_{j0} + (\psi_j^T H_{j1}) A_j$ , where  $H_{j0}$  and  $H_{j1}$  are two (possibly different) features of the history,  $H_j$ .

Many versions of the Q-learning algorithm exist, depending on whether there are parameters that are common across the stages and depending on the form of the dependent variable used in the stage 1 regression. One form for the Q-learning algorithm consists of the following steps:

1. Stage 2 regression:  $(\hat{\beta}_2, \hat{\psi}_2) = \arg \min_{\beta_2, \psi_2} \frac{1}{n} \sum_{i=1}^n (Y_{2i} - Q_2(H_{2i}, A_{2i}; \beta_2, \psi_2))^2$ .
2. Stage 1 dependent variable:  $\hat{Y}_{1i} = Y_{1i} + \max_{a_2} Q_2(H_{2i}, a_2; \hat{\beta}_2, \hat{\psi}_2)$ ,  $i = 1, \dots, n$ .
3. Stage 1 regression:  $(\hat{\beta}_1, \hat{\psi}_1) = \arg \min_{\beta_1, \psi_1} \frac{1}{n} \sum_{i=1}^n (\hat{Y}_{1i} - Q_1(H_{1i}, A_{1i}; \beta_1, \psi_1))^2$ .

In step 2, the quantity  $\hat{Y}_{1i}$  is a predictor of the unobserved random variable  $Y_{1i} + \max_{a_2} Q_2(H_{2i}, a_2)$ ,  $i = 1, \dots, n$ . The estimated optimal DTR using Q-learning is given by  $(\hat{d}_1, \hat{d}_2)$ , where the stage  $j$  optimal rule is specified as  $\hat{d}_j(b_j) = \arg \max_{a_j} Q_j(b_j, a_j; \hat{\beta}_j, \hat{\psi}_j)$ ,  $j = 1, 2$ .

Q-learning (with  $K = 2$ ) has been implemented in the R package `qLearn`, freely available from <http://cran.r-project.org/web/packages/qLearn/index.html>, and in the SAS procedure `QLEARN`, located at <http://methodology.psu.edu/downloads/procqllearn>. Q-learning can be extended for application to observational data by incorporating appropriate adjustments to account for confounding; more precisely, we can make this adjustment either by including all the

measured confounders—or simply the propensity score as a proxy for all measured confounders—in the models for Q-functions, or by weighting the stage-specific regressions by the inverse of the propensity scores (Moodie et al. 2012). Q-learning is a version of Robins’s (2004) optimal structural nested mean model developed in the causal inference literature (see Chakraborty et al. 2010 for a detailed discussion and derivation).

Q-learning has been generalized in several ways. Lizotte et al. (2010, 2012) generalize Q-learning for use when different patients may make different trade-offs among multiple outcomes, and thus a data analysis of one composite outcome is insufficient. Q-learning has also been generalized to settings in which  $Y$  is a (possibly censored) survival time (Goldberg & Kosorok 2012, Zhao et al. 2011). Both these papers provide a Q-learning method with the aim of maximizing a truncated survival time.

**3.2.2. Approaches based on dynamical systems models.** An alternate indirect approach to estimating an optimal DTR is to use dynamical systems models. By dynamical systems models, we mean a time-ordered sequence of nested conditional models (each model conditions on past data) for the multivariate distribution of the data. In this approach, one first develops a dynamical systems model; this model may be constructed using expert opinion or may be estimated using observational or sequentially randomized data sets. Indeed, these types of models are quite attractive when strong biological, behavioral, or social theories exist to guide the formation of the nested conditional models. Once the dynamical systems model is in hand, algorithms from control theory, such as DP or constrained optimization algorithms, are used to estimate the optimal DTR (Rivera et al. 2007). This approach is common in applications in engineering, economics, and business. The clinical field has seen much less development. Bayesian methods have been employed in simple, low-dimensional problems (one example is Thall et al. 2007).

Rosenberg et al. (2007) and Banks et al. (2011) discuss how different data sources with models based on ordinary differential equations can be used to build a dynamical systems model to estimate an optimal DTR in AIDS treatment. In this setting, the treatment is a continuous dose of antiviral therapy, and the optimal DTR is chosen to bring the dynamical system to its steady state. Rivera and colleagues (Navarro-Barrientos et al. 2011, Rivera et al. 2007), in a series of presentations (available at <http://csel.asu.edu/node/13>) and papers, discuss how common dynamical systems models might be used to describe behavioral dynamics and thus form the basis for DTRs that involve behavioral techniques in obesity and addiction treatment. Gaweda et al. (2005, 2008) examined the use of control-theoretic approaches to anemia management in patients with end-stage renal disease. Bennett & Hauser (2012) discuss a framework for simulating clinical decision making from electronic medical records data. In summary, although the dynamical systems approaches to developing DTRs are emerging, from a statistical perspective, they still lag behind the other approaches presented earlier. This area is ripe for further growth.

## 4. CONFIDENCE SETS

High-quality measures of confidence are needed in the development of DTRs for both (a) the parameters indexing the optimal DTR and (b) the value of a DTR—either a prespecified or estimated DTR. Numerous authors (Lunceford et al. 2002; Thall et al. 2000, 2002; Wahed & Tsiatis 2004, 2006) have addressed inference for the values of prespecified regimes; however, there is little work on inference for the value of an estimated regime. We return to this problem after discussing the construction of confidence intervals (CIs) for the parameters indexing the optimal regime. Measures of confidence for these parameters are important for the following reasons. First, if the CIs for some of these parameters contain zero, then the corresponding patient variables

need not be collected in the future, thus lowering the data collection burden. Second, CIs for the coefficient of the treatment variable can be used to indicate whether insufficient evidence in the data exists to suggest that one treatment is best, and therefore considerations other than the treatment effect, e.g., cost, patient/clinician familiarity, and/or preference, should be used to decide on treatment.

Orellana et al. (2010a) discuss the construction of confidence sets for parameters indexing the optimal DTR when direct methods of estimation using IPTW are employed. These confidence sets are based on standard Taylor series arguments and are asymptotically valid under a set of smoothness assumptions. Robins (2004) points out that nonregularity arises in the indirect estimation of DTRs. By nonregularity, we mean that the asymptotic distribution of the estimator of the treatment effect parameter does not converge uniformly over the parameter space (see below for further details). Indeed, the treatment effect parameters at any stage prior to the last can be nonregular. This phenomenon has practical consequences, including bias in estimation and poor frequentist properties of Wald-type or other standard CIs in small samples. Any inference technique that aims to provide good frequentist properties such as nominal Type I error and/or nominal coverage of CIs in small samples has to address this problem of nonregularity. Robins (2004) discusses a simple but instructive example that can help us better understand the problem; here, we present a slightly modified version, as presented by Chakraborty et al. (2010). Consider the problem of estimating  $|\mu|$  based on  $n$  independent and identically distributed observations  $X_1, \dots, X_n$  from  $\mathcal{N}(\mu, 1)$ .  $|\bar{X}_n|$  is the maximum likelihood estimator of  $|\mu|$ , where  $\bar{X}_n$  is the sample average. The asymptotic distribution of  $\sqrt{n}(|\bar{X}_n| - |\mu|)$  for any  $\mu \neq 0$  is a standard normal, whereas for  $\mu = 0$ , it is nonnormal; that is, the change in the distribution as a function of  $\mu$  is abrupt. Thus,  $|\bar{X}_n|$  is a nonregular estimator of  $|\mu|$ ; an exact proof of the nonregularity of this estimator uses local alternatives as in Leeb & Pötscher (2003). Also, for  $\mu = 0$ ,  $\lim_{n \rightarrow \infty} E[\sqrt{n}(|\bar{X}_n| - |\mu|)] = \sqrt{\frac{2}{\pi}}$ . This asymptotic bias (Robins 2004) is one symptom of the underlying nonregularity.

Next we review the problem of nonregularity in the context of Q-learning. Suppose we want to construct CIs for the parameters  $\psi_j$  appearing in the model for Q-functions. In a two-stage setup, the inference for the stage 2 parameters  $\psi_2$  is straightforward because this inference falls in the standard linear regression framework. In contrast, inference for  $\psi_1$  is complicated. The stage 1-dependent variable in Q-learning for the  $i$ th participant is  $\hat{Y}_{1i} = Y_{1i} + \max_{a_2} Q_2(H_{2i}, a_2; \hat{\beta}_2, \hat{\psi}_2) = Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}|$ ,  $i = 1, \dots, n$ , which is a nondifferentiable function of  $\hat{\psi}_2$  (owing to the presence of the absolute value function). Because  $\hat{\psi}_1$  is a function of  $\hat{Y}_{1i}$ ,  $i = 1, \dots, n$ , it is in turn a nonsmooth function of  $\hat{\psi}_2$ . As a consequence, the distribution of  $\sqrt{n}(\hat{\psi}_1 - \psi_1)$  does not converge uniformly over the parameter space (Robins 2004). More specifically, the asymptotic distribution of  $\sqrt{n}(\hat{\psi}_1 - \psi_1)$  is normal if  $\psi_2$  is such that  $p \triangleq P[H_2 : \psi_2^T H_{21} = 0] = 0$  but is nonnormal if  $p > 0$ , and this change in the distribution happens abruptly. Below we present several different approaches to address the problem.

## 4.1. Adjusted Projection Confidence Intervals

As discussed in Robins (2004), a joint CI for all of the parameters (in our two-stage example, both the first- and second-stage regression coefficients) can be formed by inverting hypothesis tests. That is, if the parameters are  $\psi = (\psi_1, \psi_2)$  and a hypothesis test of  $\psi = \psi^0$  for each value of  $\psi^0$  is well behaved, then a joint  $(1 - \alpha)\%$  CI, say,  $\mathcal{C}$  for  $\psi$ , can be constructed. This case applies in Q-learning because constructing a well-behaved hypothesis test statistic when all of the regression coefficients are set to fixed values is easy (the test statistic is based on a quadratic form involving the estimating functions evaluated at the fixed values). Next, a projected CI for  $\psi_1$

is given by  $\cup_{\psi_2} \{\psi_1 : (\psi_1, \psi_2) \in \mathcal{C}\}$ . Unfortunately, this interval is very conservative. As a result, Robins (2004), using ideas advanced by Berger & Boos (1994), adjusts the usual projection CI. We discuss this idea in the context of the two-stage Q-learning method presented above.

Recall that we are interested in a CI for  $\psi_1$ . In this context,  $\psi_2$  is a nuisance parameter. If the true value of  $\psi_2$  were known, then the asymptotic distribution of  $\sqrt{n}(\hat{\psi}_1 - \psi_1)$  would be regular (in fact, normal), and standard procedures could be used to construct an asymptotically valid CI. Let  $\mathcal{C}(\psi_2)$  denote a  $(1 - \alpha)\%$  asymptotic CI for  $\psi_1$  if  $\psi_2$  were known. Let  $\mathcal{S}$  be a  $(1 - \varepsilon)\%$  asymptotic CI for  $\psi_2$ . Then, it follows that  $\cup_{\psi_2 \in \mathcal{S}} \{\psi_1 : \psi_1 \in \mathcal{C}(\psi_2)\}$  is a  $(1 - \alpha - \varepsilon)\%$  CI for  $\psi_1$ . Importantly,  $P(\psi_1 \in \cup_{\psi_2 \in \mathcal{S}} \mathcal{C}(\psi_2)) \geq 1 - \alpha + o_P(1) + P(\psi_2 \notin \mathcal{S}) = 1 - \alpha - \varepsilon + o_P(1)$ . Thus, this CI is the union of the CIs  $\mathcal{C}(\psi_2)$  over all values  $\psi_2 \in \mathcal{S}$  and is an asymptotically valid  $(1 - \alpha - \varepsilon)\%$  CI for  $\psi_1$ . The main downside of this approach is that it appears to be computationally difficult to implement; to our knowledge, this CI has not yet been implemented.

## 4.2. Adaptive Confidence Intervals

Laber and colleagues (E. Laber, D. Lizotte, M. Qian, S. Murphy, submitted) develop an adaptive bootstrap procedure to construct CIs for linear combinations  $\mathbf{c}^T \psi_1$ , where  $\mathbf{c}$  is a known vector. In this procedure, they decompose the asymptotic expansion of  $\mathbf{c}^T \sqrt{n}(\hat{\psi}_1 - \psi_1)$  as  $\mathbb{W}_n + \mathbb{U}_n$ , where the first term,  $\mathbb{W}_n$ , is smooth and asymptotically normally distributed, and the distribution of the second term,  $\mathbb{U}_n$ , depends on the underlying data-generating process in a nonsmooth manner. The adaptive confidence intervals (ACIs) are formed by first constructing smooth data-dependent upper and lower bounds on  $\mathbb{U}_n$  and thereby on  $\mathbf{c}^T \sqrt{n}(\hat{\psi}_1 - \psi_1)$ . The data-dependent upper and lower bounds use a pretest (Olshen 1973) that partitions the data into two sets: (a) patients for whom a treatment effect appears to exist and (b) patients for whom a treatment effect does not appear to exist. The pretests are performed using a critical value  $\lambda_n$ , which is a tuning parameter of the procedure and can be varied; Laber and colleagues (E. Laber, D. Lizotte, M. Qian, S. Murphy, submitted) use  $\lambda_n = \log \log n$  in their analysis.


Let the upper and lower bounds on  $\mathbf{c}^T \sqrt{n}(\hat{\psi}_1 - \psi_1)$  be given by  $\mathcal{U}(\mathbf{c})$  and  $\mathcal{L}(\mathbf{c})$ , respectively; both these quantities are functions of  $\lambda_n$ . Laber and colleagues (E. Laber, D. Lizotte, M. Qian, S. Murphy, submitted) show that the asymptotic distributions of  $\mathbf{c}^T \sqrt{n}(\hat{\psi}_1 - \psi_1)$ ,  $\mathcal{U}(\mathbf{c})$ , and  $\mathcal{L}(\mathbf{c})$  are all equal in the regular case when  $p = 0$ . That is, when there is a large treatment effect for almost all patients, the bounds are asymptotically tight. However, when there is a nonnull subset of patients with no treatment effect, the asymptotic distribution of  $\mathcal{U}(\mathbf{c})$  is stochastically larger than the asymptotic distribution of  $\mathbf{c}^T \sqrt{n}(\hat{\psi}_1 - \psi_1)$ , and likewise the asymptotic distribution of  $\mathcal{L}(\mathbf{c})$  is stochastically smaller. This adaptivity between nonregular and regular settings is a key feature of this procedure. We can approximate the distributions of  $\mathcal{U}(\mathbf{c})$  and  $\mathcal{L}(\mathbf{c})$  using the bootstrap. Let  $\hat{u}$  be the  $1 - \alpha/2$  quantile of the bootstrap distribution of  $\mathcal{U}(\mathbf{c})$ , and let  $\hat{l}$  be the  $\alpha/2$  quantile of the bootstrap distribution of  $\mathcal{L}(\mathbf{c})$ . Then  $(\mathbf{c}^T \hat{\psi}_1 - \hat{u}/\sqrt{n}, \mathbf{c}^T \hat{\psi}_1 - \hat{l}/\sqrt{n})$  is the ACI for  $\mathbf{c}^T \psi_1$ . Laber and colleagues (E. Laber, D. Lizotte, M. Qian, S. Murphy, submitted) prove the consistency of the bootstrap in this context, and in particular that  $P(\mathbf{c}^T \hat{\psi}_1 - \hat{u}/\sqrt{n} \leq \mathbf{c}^T \psi_1 \leq \mathbf{c}^T \hat{\psi}_1 - \hat{l}/\sqrt{n}) \geq 1 - \alpha + o_P(1)$ , where the probability statement is with respect to the bootstrap distribution. Furthermore, if  $p = 0$ , then the above inequality can be strengthened to equality. This result shows that the adaptive bootstrap method can be used to construct valid, though potentially conservative, CIs regardless of the underlying parameters of the generative model. This method is implemented in the SAS procedure QLEARN: <http://methodology.psu.edu/downloads/procqlearn>.

## 4.3. *m*-Out-of-*n* Bootstrap Confidence Intervals

The *m*-out-of-*n* bootstrap is a tool for producing valid CIs for nonsmooth functionals (Shao 1994). This method is the same as the ordinary bootstrap, except that the resample size (*m*) satisfies these

requirements:  $m \rightarrow \infty$  as  $n \rightarrow \infty$ , but  $m = o(n)$ . Chakraborty et al. (2013) propose a data-driven method for choosing  $m$  in the context of Q-learning, a method that is directly connected to an estimated degree of nonregularity. This method is adaptive in that it leads to the usual  $n$ -out-of- $n$  bootstrap in regular settings ( $p = 0$ ) and the  $m$ -out-of- $n$  bootstrap otherwise.

In this approach, Chakraborty et al. (2013) consider a class of resample sizes of the form  $m = n^{[1+\eta(1-p)]/(1+\eta)}$ , where  $\eta > 0$  is a tuning parameter. For implementation, one first needs to estimate  $p$  using a plug-in estimator,  $\hat{p} = \mathbb{P}_n \mathbb{I}[n(H_{21}^T \hat{\psi}_2)^2 \leq (H_{21}^T \hat{\Sigma}_{\hat{\psi}_2} H_{21}) \cdot \chi_{1,1-\nu}^2]$ , where  $n^{-1} \hat{\Sigma}_{\hat{\psi}_2}$  is the plug-in estimator of the asymptotic covariance matrix of  $\hat{\psi}_2$ , and  $\chi_{1,1-\nu}^2$  is the  $(1 - \nu) \times 100$  percentile of a  $\chi^2$  distribution with 1 degree of freedom. Then the data-driven choice of the resample size is given by  $\hat{m} = n^{[1+\eta(1-p)]/(1+\eta)}$ . For fixed  $n$ ,  $\hat{m}$  is a monotonically decreasing function of  $\hat{p}$ , taking values in the interval  $[n^{(1/1+\eta)}, n]$ . Thus,  $\eta$  governs the smallest acceptable resample size. The procedure is robust to the choice of  $\nu$ . Once  $\hat{m}$  is computed, a  $(1 - \alpha) \times 100\%$   $m$ -out-of- $n$  bootstrap CI for  $\mathbf{c}^T \psi_1$  is given by  $(\mathbf{c}^T \hat{\psi}_1 - \hat{u}/\sqrt{\hat{m}}, \mathbf{c}^T \hat{\psi}_1 - \hat{l}/\sqrt{\hat{m}})$ , where  $\hat{l}$  and  $\hat{u}$  are the  $(\alpha/2) \times 100$  and  $(1 - \alpha/2) \times 100$  percentiles of  $\mathbf{c}^T \sqrt{\hat{m}}(\hat{\psi}_1^{(b)} - \hat{\psi}_1)$ , respectively [ $\hat{\psi}_1^{(b)}$  is the  $m$ -out-of- $n$  bootstrap analog of  $\hat{\psi}_1$ ]. This bootstrap procedure is consistent, and  $P(\mathbf{c}^T \hat{\psi}_1 - \hat{u}/\sqrt{\hat{m}} \leq \mathbf{c}^T \psi_1 \leq \mathbf{c}^T \hat{\psi}_1 - \hat{l}/\sqrt{\hat{m}}) \geq 1 - \alpha + o_p(1)$ , where the probability statement is with respect to the bootstrap distribution. Furthermore, if  $p = 0$ , then the procedure possesses the adaptive property in that the above inequality is an equality. The method has been implemented in the R package *qLearn* at <http://cran.r-project.org/web/packages/qLearn/index.html>. We refer the reader to the **Supplemental Appendix** for a simulation study that illustrates the performance of the above approaches to forming a CI (follow the **Supplemental Material** link from the Annual Reviews home page at <http://www.annualreviews.org>).

 Supplemental Material

#### 4.4. Confidence Intervals for the Value of an Estimated DTR

The topic of constructing CIs for the value of an estimated DTR has not been adequately addressed in the literature yet, but we can gain some insight by exploiting its connection with classification. As Qian & Murphy (2011) and Zhao et al. (2012) highlight, the value of a DTR can be expressed in a similar form as the misclassification error rate in a weighted classification problem. Thus, constructing a CI for the value of an estimated DTR is equivalent to constructing a CI for the test error of an estimated weighted classifier. Unfortunately, even in an unweighted classification problem, constructing a CI for the test error is difficult because of inherent nonsmoothness; standard methods like normal approximation or usual bootstrap fail. Laber & Murphy (2011) develop a method for constructing such CIs using smooth data-dependent upper and lower bounds on the test error; this method is similar to the ACI method described in Section 4.2. Although intuitively one can expect that this method could be successfully adapted for the value of an estimated DTR, more targeted research is needed to extend and fine-tune the procedure to the current setting.

### 5. DISCUSSION AND THE FUTURE

DTRs make up an increasingly active area of current statistical research and have received much interest from the clinical science community. SMART studies are increasing in number, indicating that, for some time, the design of and data analysis for these trials will provide a steady source of new statistical problems. For example, many interventions are administered in group settings; in the case of DTRs, this type of administration requires the design and analysis of cluster-randomized SMARTs. At the design level, cluster randomization would imply increased sample size requirements because of intraclass correlation. At the analysis level, it would open up questions



such as how best to incorporate random effects models or generalized estimating equations into the existing framework of estimation and how the intraclass correlation would affect the nonregularity in inference. Furthermore, the development of statistical methods that can be used in the analysis of longitudinal observational data sets will likely continue to be necessary in this area. In either case, methods for variable selection and model checking in the context of constructing data-driven DTRs, both of which pose issues slightly different from those of similar topics in the prediction literature, are underdeveloped and warrant further research.

Inference in the domain of DTRs is a particularly challenging problem because of the nonregularity of the estimators under certain underlying longitudinal data distributions. This challenge occurs both when the targets of inference are the parameters indexing the optimal DTR and when the target is the value of an estimated DTR. In these nonregular problems, methods for developing optimal CIs represent an open area of research. Interest is growing in CIs for other parameters. One example is data-dependent parameters, such as the first-stage regression coefficients that would result in a future study in which the estimated second-stage decision rule is used to assign treatment. CIs for this type of parameter are as yet undeveloped.

Today's health care increasingly uses sophisticated mobile devices (e.g., smart phones, actigraph units containing accelerometers) to remotely monitor patients' chronic conditions and to intervene when needed. This increased use is an instance in which methods from online reinforcement learning in the infinite horizon setting may be useful. Development of sound estimation and inference techniques for such a setting is an important research direction.

The field of DTRs is in its infancy but is quickly evolving. These methods and trial designs hold much promise for informing sequential decision making in health care. To achieve this promise, many of the problems discussed above require further efforts on the part of the statistical community. Dissemination of the newly developed methods into the medical domains and collaboration with clinical scientists will be crucial.

## 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

Dr. Chakraborty's research is supported by National Institutes of Health (NIH) grant R01 NS072127-01A1. Dr. Murphy's research is supported by NIH grant P50DA010075.

## LITERATURE CITED

- Banks H, Jang T, Kwon H. 2011. Feedback control of HIV antiviral therapy with long measurement time. *Int. J. Pure Appl. Math.* 66:461–85
- Bellman R. 1957. *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press
- Bennett C, Hauser K. 2012. Artificial intelligence framework for simulating clinical decision-making: a Markov decision process approach. *Artif. Intell. Med.* 57:9–19
- Berger R, Boos D. 1994. *P* values maximized over a confidence set for the nuisance parameter. *J. Am. Stat. Assoc.* 89:1012–16
- Box G, Hunter W, Hunter J. 1978. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. New York: Wiley
- Cai T, Tian L, Wong P, Wei L. 2011. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* 12:270–82



- Chakraborty B, Laber E, Zhao Y. 2013. Inference for optimal dynamic treatment regimes using an adaptive  $m$ -out-of- $n$  bootstrap scheme. *Biometrics* 69:714–23
- Chakraborty B, Murphy S, Strecher V. 2010. Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.* 19:317–43
- Coffey C, Levin B, Clark C, Timmerman C, Wittes J, et al. 2012. Overview, hurdles, and future work in adaptive designs: perspectives from an NIH-funded workshop. *Clin. Trials* 9:671–80
- Cotton C, Heagerty P. 2011. A data augmentation method for estimating the causal effect of adherence to treatment regimens targeting control of an intermediate measure. *Stat. Biosci.* 3:28–44
- Dawson R, Lavori P. 2010. Sample size calculations for evaluating treatment policies in multi-stage designs. *Clin. Trials* 7:643–52
- Dawson R, Lavori P. 2012. Efficient design and inference for multistage randomized trials of individualized treatment policies. *Biostatistics* 13:142–52
- Ernst D, Geurts P, Wehenkel L. 2005. Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.* 6:503–56
- Feng W, Wahed A. 2008. Supremum weighted log-rank test and sample size for comparing two-stage adaptive treatment strategies. *Biometrika* 95:695–707
- Feng W, Wahed A. 2009. Sample size for two-stage studies with maintenance therapy. *Stat. Med.* 28:2028–41
- Gaweda A, Jacobs A, Aronoff G, Brier M. 2008. Model predictive control of erythropoietin administration in the anemia of ESRD. *Am. J. Kidney Dis.* 51:71–79
- Gaweda A, Muezzinoglu M, Aronoff G, Jacobs A, Zurada J, Brier M. 2005. Individualization of pharmacological anemia management using reinforcement learning. *Neural Netw.* 18:826–34
- Goldberg Y, Kosorok M. 2012. Q-learning with censored data. *Ann. Stat.* 40:529–60
- Hernán MA, Lanoy E, Costagliola D, Robins JM. 2006. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic Clin. Pharmacol. Toxicol.* 98:237–42
- Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7:443–70
- Jones H. 2010. *Reinforcement-based treatment for pregnant drug abusers (HOME II)*. ClinicalTrials.gov database, updated October 19, 2012, accessed July 24, 2013, Natl. Inst. Health, Bethesda, MD. <http://clinicaltrials.gov/ct2/show/NCT01177982>
- Kasari C. 2009. *Developmental and augmented intervention for facilitating expressive language (CCNLA)*. ClinicalTrials.gov database, updated Apr. 26, 2012, accessed July 24, 2013. Natl. Inst. Health, Bethesda, MD. <http://clinicaltrials.gov/ct2/show/NCT01013545>
- Kulkarni K, Gosavi A, Murray S, Grantham K. 2011. Semi-Markov adaptive critic heuristics with application to airline revenue management. *J. Control Theory Appl.* 9:421–30
- Laber E, Murphy S. 2011. Adaptive confidence intervals for the test error in classification. *J. Am. Stat. Assoc.* 106:904–13
- Lavori P, Dawson R. 2000. A design for testing clinical strategies: biased adaptive within-subject randomization. *J. R. Stat. Soc. A* 163:29–38
- Lavori P, Dawson R. 2004. Dynamic treatment regimes: practical design considerations. *Clin. Trials* 1:9–20
- Lavori P, Dawson R. 2008. Adaptive treatment strategies in chronic disease. *Annu. Rev. Med.* 59:443–53
- Leeb H, Pötscher B. 2003. The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econom. Theory* 19:100–42
- Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy S. 2011. A “SMART” design for building individualized treatment sequences. *Annu. Rev. Clin. Psychol.* 8:21–48
- Li Z, Murphy S. 2011. Sample size formulae for two-stage randomized trials with survival outcomes. *Biometrika* 98:503–18
- Lizotte D, Bowling M, Murphy S. 2010. Efficient reinforcement learning with multiple reward functions for randomized clinical trial analysis. In *Twenty-Seventh International Conference on Machine Learning (ICML)*, pp. 695–702. Haifa, Israel: Omnipress
- Lizotte D, Bowling M, Murphy S. 2012. Linear fitted-Q iteration with multiple reward functions. *J. Mach. Learn. Res.* 13:3253–95
- Lunceford J, Davidian M, Tsiatis A. 2002. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics* 58:48–57

- Miyahara S, Wahed A. 2010. Weighted Kaplan-Meier estimators for two-stage treatment regimes. *Stat. Med.* 29:2581–91
- Moodie E, Chakraborty B, Kramer M. 2012. Q-learning for estimating optimal dynamic treatment rules from observational data. *Can. J. Stat.* 40:629–45
- Moodie E, Platt R, Kramer M. 2009. Estimating response-maximized decision rules with applications to breastfeeding. *J. Am. Stat. Assoc.* 104:155–65
- Murphy S. 2003. Optimal dynamic treatment regimes. *J. R. Stat. Soc. B* 65:331–66
- Murphy S. 2005a. A generalization error for Q-learning. *J. Mach. Learn. Res.* 6:1073–97
- Murphy S. 2005b. An experimental design for the development of adaptive treatment strategies. *Stat. Med.* 24:1455–81
- Murphy S, Bingham D. 2009. Screening experiments for developing dynamic treatment regimes. *J. Am. Stat. Assoc.* 184:391–408
- Murphy S, van der Laan M, Robins JM, Conduct Probl. Prev. Res. Group. 2001. Marginal mean models for dynamic regimes. *J. Am. Stat. Assoc.* 96:1410–23
- Nahum-Shani I, Qian M, Almira D, Pelham W, Gnagy B, et al. 2012a. Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychol. Methods* 17:457–77
- Nahum-Shani I, Qian M, Almira D, Pelham W, Gnagy B, et al. 2012b. Q-learning: a data analysis method for constructing adaptive interventions. *Psychol. Methods* 17:478–94
- Navarro-Barrientos J, Rivera D, Collins L. 2011. A dynamical model for describing behavioural interventions for weight loss and body composition change. *Math. Comput. Model. Dyn. Syst.* 17:183–203
- Oetting A, Levy J, Weiss R, Murphy S. 2011. Statistical methodology for a SMART design in the development of adaptive treatment strategies. In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, ed. PE Shrout, KM Keyes, K Ornstein, pp. 179–205. New York: Oxford Univ. Press
- Olshen R. 1973. The conditional level of the F-test. *J. Am. Stat. Assoc.* 68:692–98
- Orellana L, Rotnitzky A, Robins JM. 2010a. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part I: main content. *Int. J. Biostat.* 6:8
- Orellana L, Rotnitzky A, Robins JM. 2010b. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part II: proofs and additional results. *Int. J. Biostat.* 6:9
- Ormoneit D, Sen S. 2002. Kernel-based reinforcement learning. *Mach. Learn.* 49:161–78
- Petersen ML, Deeks SG, van der Laan MJ. 2007. Individualized treatment rules: generating candidate clinical trials. *Stat. Med.* 26:4578–601
- Qian M, Murphy S. 2011. Performance guarantees for individualized treatment rules. *Ann. Stat.* 39:1180–210
- Rivera D, Pew M, Collins L. 2007. Using engineering control principles to inform the design of adaptive interventions: a conceptual introduction. *Drug Alcohol Depend.* 88:S31–40
- Robins J. 1986. A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Math. Model.* 7:1393–512
- Robins J. 1989. The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*, ed. L Sechrest, H Freeman, A Mulley, pp. 113–59. New York: Natl. Cent. Health Serv. Res. Health Care Technol.
- Robins J. 1993. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. *Proc. Biopharm. Sect. Am. Stat. Assoc.*, pp. 24–33. Alexandria, VA: Am. Stat. Assoc.
- Robins J. 1994. Correcting for non-compliance in randomized trials using structural nested mean models. *Commun. Stat.* 23:2379–412
- Robins J. 1997. Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, ed. M Berkane, pp. 69–117. New York: Springer
- Robins J. 1999. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, ed. ME Halloran, D Berry, pp. 95–134. New York: Springer
- Robins J. 2004. Optimal structural nested models for optimal sequential decisions. *Proc. Seattle Symp. Biostat.*, 2nd, ed. D Lin, P Heagerty, pp. 189–326. New York: Springer

- Robins JM, Hernán MA, Brumback B. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11:550–60
- Robins JM, Orellana L, Rotnitzky A. 2008. Estimation and extrapolation of optimal treatment and testing strategies. *Stat. Med.* 27:4678–721
- Rosenberg E, Davidian M, Banks H. 2007. Using mathematical modeling and control to develop structured treatment interruption strategies for HIV infection. *Drug Alcohol Depend.* 88:S41–51
- Rosthøj S, Fullwood C, Henderson R, Stewart S. 2006. Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach. *Stat. Med.* 25:4197–215
- Rubin D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66:688–701
- Rubin D. 1980. Discussion of “Randomized analysis of experimental data: the Fisher randomization test” by Basu D. *J. Am. Stat. Assoc.* 75:591–93
- Shao J. 1994. Bootstrap sample size in nonregular cases. *Proc. Am. Math. Soc.* 122:1251–62
- Shortreed S, Moodie E. 2012. Estimating the optimal dynamic antipsychotic treatment regime: evidence from the sequential-multiple assignment randomized CATIE Schizophrenia Study. *J. R. Stat. Soc. C* 61:577–99
- Sutton R, Barto A. 1998. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press
- Thall P, Millikan R, Sung H. 2000. Evaluating multiple treatment courses in clinical trials. *Stat. Med.* 30:1011–28
- Thall P, Sung H, Estey E. 2002. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *J. Am. Stat. Assoc.* 97:29–39
- Thall P, Wathen J. 2005. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Stat. Med.* 24:1947–64
- Thall PF, Logothetis C, Pagliaro LC, Wen S, Brown MA, et al. 2007. Adaptive therapy for androgen-independent prostate cancer: a randomized selection trial of four regimens. *J. Natl. Cancer Inst.* 99:1613–22
- van der Laan MJ, Petersen ML. 2007a. Causal effect models for realistic individualized treatment and intention to treat rules. *Int. J. Biostat.* 3:3
- van der Laan MJ, Petersen ML. 2007b. Statistical learning of origin-specific statically optimal individualized treatment rules. *Int. J. Biostat.* 3:6
- Wagner E, Austin B, Davis C, Hindmarsh M, Schaefer J, Bonomi A. 2001. Improving chronic illness care: translating evidence into action. *Health Aff.* 20:64–78
- Wahed A, Tsiatis A. 2004. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomized designs in clinical trials. *Biometrics* 60:124–33
- Wahed A, Tsiatis A. 2006. Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. *Biometrika* 93:163–77
- Wang L, Rotnitzky A, Lin X, Millikan R, Thall P. 2012. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *J. Am. Stat. Assoc.* 107:493–508
- Zhang B, Tsiatis A, Davidian M, Zhang M, Laber E. 2012a. Estimating optimal treatment regimes from a classification perspective. *Stat* 1:103–14
- Zhang B, Tsiatis A, Laber E, Davidian M. 2012b. A robust method for estimating optimal treatment regimes. *Biometrics* 68:1010–18
- Zhao Y, Zeng D, Rush A, Kosorok M. 2012. Estimating individual treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.* 107:1106–18
- Zhao Y, Zeng D, Socinski M, Kosorok M. 2011. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* 67:1422–33