Published in final edited form as:
Annu Rev Stat Appl. 2014 ; 1: 447–464. doi:10.1146/annurev-statistics-022513-115553.

Dynamic Treatment Regimes

Bibhas Chakraborty[1] and **Susan A. Murphy**[2]

[1]Department of Biostatistics, Columbia University, New York, USA, 10032

[2]Department of Statistics and Institute for Social Research, University of Michigan, Ann Arbor, USA, 48109

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 best study design as well as efficient estimation and valid inference. Due to the many novel methodological challenges it offers, this area 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 non-standard asymptotics. We reference software, whenever available. We also outline some important future directions.

Keywords

dynamic treatment regime; reinforcement learning; sequential randomization; non-regularity; Q-learning

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 demographics, results of diagnostic tests, genetic information, etc.). *Dynamic treatment regimes* (DTRs) [1, 2, 3, 4, 5, 6] 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* [7, 8, 9, 10, 11] or *treatment policies* [12, 13, 14]. 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, where a patient typically has to be treated at multiple stages, adapting the treatment (type, dosage, timing) at each stage to the evolving treatment and covariate history. DTRs underpin *clinical decision support systems* – described as a key element of the *chronic care model* [15].

A simple example of a DTR arising in the treatment of alcohol dependence is: After the patient completes an intensive outpatient program, provide the medication naltrexone along with face-to-face medical management. If within the following two months the patients experiences 2 or more heavy drinking days, then immediately augment the naltrexone with a cognitive behavioral therapy. Otherwise at the end of the two months, provide telephone disease management in addition to the naltrexone. A second example given in Rosthøj *et al*. [16] is a DTR for use in guiding warfarin dosing to control the risk of both clotting and excessive bleeding. Here the decision rules input summaries of the trajectory of 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 any). The third example, provided by Robins *et al*. [17] concerns a DTR with 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 different statistical methods for constructing the decision rules in a DTR are reviewed.

1.1 Decision Problems

Traditionally personalized medicine concerns single stage decision making. In a single-stage (non-dynamic) decision problem one observes a random vector, the first observation, O_1, then one selects an action (here a treatment action), a_1, from a set $\mathcal{A}_1$ of actions and then depending on which action was selected, observes a second observation, $O_2(a_1)$. To avoid technical details and for simplicity, here and below, we assume sufficient regularity for all statements. A decision rule, say d_1, is a mapping from the range of O_1 into $\mathcal{A}_1$. The quality of a treatment for a particular value of O_1 is evaluated in terms of its *utility*, say $r(O_1, a_1, O_2(a_1))$, for r a known function. 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*. [18] 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, $\mathcal{U}(o_1;a_1) = E[r(O_1, a_1, O_2(a_1)) | O_1 = o_1]$; this is personalized decision making since the choice of optimal treatment depends on o_1. Equivalently the optimal decision rule is given by $\arg\max_{d1} E[\mathcal{U}(O_1;d_1(O_1))]$, where the maximum is taken over all functions on the range of O_1. $E[\mathcal{U}(O_1;d_1(O_1))]$ is called the Value of the decision rule d_1.

Constructing DTRs involves solving, or estimating quantities relevant in, a multi-stage decision problem. In multi-stage decision problems, observations are interweaved with action selection; denote such a sequence by

$$O_1, a_1, O_2(a_1), a_2, O_3\left(\bar{a}_2\right), \ldots, O_k\left(\bar{a}_{K-1}\right), a_K, O_{K+1}\left(\bar{a}_K\right)$$ where $\bar{a}_j = \{a_1, \ldots, a_j\}$

and $O_{j+1}\left(\bar{a}_j\right)$ 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, $\bar{d}_K = (d_1, \ldots, d_K)$; the decision rule d_j is a mapping from the range of $\left(O_1, a_1, \ldots, O_j\left(\bar{a}_{j-1}\right)\right)$ into the jth action space,

$\mathscr{A}_j$. 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

$$E\left[\sum_{a_1\in\mathscr{A}_1} 1_{a_1=d_1(O_1)} \sum_{a_2\in\mathscr{A}_2} 1_{a_2=d_2(O_1,a_1,O_2(a_1))^r}\left(O_1, a_1, O_2\left(a_1\right), a_2, O_3\left(a_1, a_2\right)\right)\right] \quad (1)$$

(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 maximize the Value.

Constructing the optimal decision rules in multi-stage decision problems is challenging due to the time-varying or dynamic nature of this problem. Historically, an early method for solving (e.g. construct the optimal decision rules) multi-stage decision problems is *dynamic programming* (DP), which dates at least back to Bellman [19]. The primary reason why classical DP algorithms have seen little use in DTR research is due to the fact that these algorithms require complete knowledge of, or a full model for, the multivariate distribution of the data for any set of actions; this is impractical in many application areas (*curse of modeling*) [20]. Secondly, 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* [21]. But DP provides an important theoretical and conceptual foundation for research in multi-stage decision problems; in fact, as will be seen, many present day estimation methods build on classical DP algorithms, while relaxing its 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. to estimate the sequence of decision rules, one per stage, that result in the highest Value, within a class of DTRs. In each case the data used in comparing or constructing DTRs are usually from: (i) sequentially randomized studies, or (ii) longitudinal observational studies, or (iii) dynamical system models. Research based on the first source of data, that from sequentially randomized studies, is experiencing a period of rapid growth, due to 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' pioneering work [1, 2, 3, 4] 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 seen much less use 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 will be discussed in Section 3.

2.1 Sequentially Randomized Trials (SMART)

Initially, beginning with Robins' work [1, 2, 3, 4], 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 [7, 22, 11], have been implemented in practice. SMART designs involve an initial randomization of patients to available treatment actions, followed by re-randomizations at each subsequent stage of some or all of the patients to treatment actions available at that stage. The re-randomizations 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 [23]; a study involving treatment of autism among children [24, 25]; a study involving interventions for children with attention deficit hyperactivity disorder [26, 27]; a study involving treatment for pregnant drug abusers [28, 25]; and a study involving alcohol-dependent individuals [25]. For a list of some further SMARTs see the website http://methodology.psu.edu/ra/adap-treat-strat/projects.

To make the discussion more concrete, see Figure 1 for 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: cognitive behavioral therapy (CBT) or naltrexone (NTX). A participant is classified as a *non-responder* or *responder* to the initial treatment according to whether s/he does or does not experience more than two heavy drinking days during the next two months. A non-responder to NTX is re-randomized to one of the two subsequent treatment options: either a switch to CBT, or an augmentation of NTX with CBT (CBT+NTX). Similarly, a non-responder to CBT is re-randomized to either a switch to NTX, or an augmentation (CBT+NTX). Responders to the initial treatment receive telephone monitoring (TM) for an additional period of 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.

Denote the observable data trajectory for a participant in a two-stage SMART by $(O_1, A_1, O_2, A_2, O_3)$, where O_1, O_2 and 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 $Y = r(O_1, A_1, O_2, A_2, O_3)$ for r a known function. For example, in the addiction management study above, O_1 may include addiction severity and co-morbid conditions, 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 (1.1), we make a short digression into the field of causal inference. Recall that in the case of two stages, in Section (1.1), we denoted the sequence of random observations by $(O_1, a_1, O_2(a_1), a_2, O_3(a_1, a_2))$ for the selected actions (a_1, a_2). These observations are *potential outcomes* [29, 1]. *Potential outcomes* or *counterfactual outcomes* are defined as a person's outcome had s/he followed a particular treatment (sequence), possibly different from the treatment (sequence) s/he 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 to have two potential second observations, $O_2(a)$ and $O_2(a')$. However, only

one of these – the one corresponding to the treatment a participant is randomized to – will be observed. Clearly, it is not possible to observe the O_2 under both treatments a and a' 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' $\left(\mathscr{A}_1=\mathscr{A}_2=\left\{a, a'\right\}\right)$. In this case there are four sequences of potential observations, $(O_2(a), O_3(a, a))$, $(O_2(a), O_3(a, a'))$, $(O_2(a'), O_3(a', a))$, $(O_2(a'), O_3(a', a'))$; only one of these sequences will be observed on any given participant.

To connect the potential observations to the observations made during the conduct of a SMART, we make two assumptions [4]:

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

$$A_j \perp \left(O_{j+1}\left(\bar{a}_j\right), \ldots, O_K\left(\bar{a}_{K-1}\right), O_{K+1}\left(\bar{a}_K\right)\right) \Big| H_j \quad \forall_j=1, \ldots, K.$$

The consistency assumption subsumes Rubin's [30] more explanatory *Stable Unit Treatment Value Assumption* (SUTVA), which is: each participant's potential outcome is not influenced by the treatment applied to 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 $O_2 = O_2(A_1)$ and $O_3 = O_3(A_1, A_2)$. 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 O_1 and (O_1, A_1, O_2), respectively. Under this assumption, $P(O_2(a_1) \le o_2|O_1 = o_1) = P(O_2 \le o_2|O_1 = o_1, A_1 = a_1)$, and $P(O_3(a_1, a_2) \le y|O_1 = o_1, O_2(a_1) = o_2) = P(O_3 \le y|O_1 = o_1, A_1 = a_1, O_2 = o_2, A_2 = a_2)$. This implies 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 (1) can be written as

$$E\left[E\left[\sum_{a_1 \in \mathscr{A}_1} 1_{a_1=d_1(H_1)} E\left[\sum_{a_2 \in \mathscr{A}_2} 1_{a_2=d_2(H_2)} E\left[r\left(H_2, A_2, O_3\right) | H_2, A_2=a_2\right] | \right] \Big| H_1, A_1=a_1\right]\right]$$

(recall $H_1 = O_1$ and $H_2 = (O_1, A_1, O_2)$). 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 pre-specified DTRs or to estimate the optimal DTR within a certain class.

2.1.1 Some Practical Considerations in Designing a SMART— A variety of 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 [22, 11]. For example it is better to use a low dimensional summary criterion (e.g. responder/non-responder status, as used in the example addiction management SMART) instead of all intermediate outcomes (e.g. improvement of symptom severity, side-effects, adherence etc.) to restrict the class of possible treatments. 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 [11]. That is, the construction of DTRs is a developmental endeavor as opposed to confirmatory. In this sense a scientist employing a SMART design has a similar goal to Box's [31] goal of developing multicomponent treatments. Indeed the SMART can be viewed as an extension of the factorial design to the setting in which time and sequencing of treatments play a crucial role [32]. As a result often the primary hypothesis, that is, the hypothesis used to determine the sample size for the trial, concerns a main effect. However due to the multiple randomizations, a variety of interesting secondary research questions can be addressed with randomized data. Note that the SMART may or may not be powered 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 would be: "marginalizing over secondary treatments, what is the best initial treatment on average?". In other words, here the researcher wants to compare the mean primary outcome of the group of 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 non-responders to initial treatment?". Here the researcher might compare the mean primary outcome of non-responders assigned to switch with the mean primary outcome of non-responders 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 addiction management SMART there are 4 embedded DTRs, corresponding to 2 options for the first stage treatment and 2 options for the second stage treatment for nonresponders (note that there is only one option for the responders). For example, one embedded regime in this SMART is: '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 non-responder'; other embedded regimes can be described similarly. Determining appropriate sample sizes to compare two embedded DTRs in terms of a continuous outcome was considered by Murphy [11], Oetting *et al*. [33], and Dawson and Lavori [34, 35]. 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 [12, 13, 14, 36]. Relevant sample size formulae can be found in Feng and Wahed [37] and Li and Murphy [38]. 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* [39]. The term, "adaptive design" is an umbrella term used to denote a variety of trial designs that allow certain trial features to change based on accumulating data while maintaining statistical, scientific, and ethical integrity of the trial [39]. In a SMART design, each participant moves through multiple stages of treatment, while 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, the design elements such as the final sample size, randomization probabilities and treatment options are pre-specified. SMART designs involve *within-participant adaptation* of treatment, while adaptive designs involve *between-participant adaptation*. While in some settings it is possible to incorporate some adaptive elements into a SMART design [10, 40], how to optimally do this 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 the SMART is used to develop a DTR, that is, a sequence of treatments. Note that treatment allocation at any stage after the initial stage of a SMART typically depends on a participant's intermediate outcome (response/non-response). However, in a crossover trial, participants receive all the candidate treatments irrespective of their intermediate outcomes. And most importantly, it is crucial in a crossover trial to attempt to *wash out* the *carryover effects*, whereas the process of constructing a DTR involves harnessing carryover effects so as to lead to improved outcomes. That is, carryover effects such as synergistic interactions between treatments at different stages may lead to a better DTR as compared to a DTR in which there are no carryover effects.

2.2 Observational Studies

In observational studies the treatments are not randomized; in particular, the reasons why different individuals receive differing treatments or the reasons why one individual receives different treatments at different times are not known with certainty. Certainly data in which the treatments are (sequentially) randomized, when available, is preferable for making inferences concerning DTRs. However observational studies are the most common source of data for constructing DTRs and indeed 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 be partially due to the unobserved or unknown reasons why individuals receive differing treatments as opposed to the effects of the treatments. Thus to conduct inference, assumptions are required. Assumptions such as the consistency assumption and the no unmeasured confounders assumption 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. A variety of studies aimed at constructing DTRs from observational data have been undertaken. Data sources include hospital databases [16, 17, 41, 42], randomized encouragement trials [43], and cohort studies [44].

The assumption of no unmeasured confounders must be given careful consideration and thought in the observational data setting. Recall the no unmeasured confounders is the assumption that conditional on the past history, treatment received at stage *j* is independent of future potential observations and outcome: $P\left(A_j=a|H_j, O_{j+1}\left(\bar{a}_j\right), \ldots, O_K\left(\bar{a}_{K-1}\right), O_{K+1}\left(\bar{a}_K\right)\right)=P\left(A_j=a|H_j\right)$. 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* [45] or *viable* [18] 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 small number of DTRs in terms of their Value; and (b) to estimate the optimal DTR within a certain class. In the following, we review the analysis strategies for both. Throughout we assume that both the no-unmeasured confounding and consistency assumptions hold and that all DTRs considered are feasible.

Weighting is often used to address both (a) and (b). Weights or *inverse probability of treatment weights* (IPTW) were originally developed to estimate the Value of non-dynamic regimes [46, 47], but later adapted to the problem of estimating the Value of DTRs. IPTWs were used to estimate the Values of a small number of DTRs in Murphy *et al*. [48] and Wang *et al*. [18]. To see why weights might be used, consider a SMART as in Figure 1, with only one option for responding participants (e.g. telephone monitoring). Suppose that the treatment assignment probabilities at stage 1 and also for the non-responders are uniform (randomization probability is 0.5). Suppose further that 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 non-responder.” To estimate the Value

we utilize the outcome of all participants with treatment patterns consistent with this DTR. However within this group of participants there is an over-representation of responders compared to non-responders because the non-responders were subdivided in the trial but the responders were not. The IPTWs are used to adjust for over-representation of participants across the treatment patterns consistent with a given DTR. In this example, data from responders would have a weight of $\frac{1}{0.5}$ as responders have been randomized only in stage 1 (with a probability of 0.5) whereas data from non-responders would have a weight of $\frac{1}{(0.5)\,(0.5)}$ as they have been randomized twice (each with a probability of 0.5). See Wang *et al*. [18] and Nahum-Shani *et al*. [26] for detailed explanations of how IPTWs can be used to account for this over/under representation in SMARTs. Lunceford *et al*. [12], Wahed and Tsiatis [13, 14], and Miyahara and Wahed [49] use ITPW weights in estimating the Value of DTRs in the survival analysis setting. Improved versions of IPTW estimator are available in papers by Robins and colleagues [48, 17, 41, 50] and Zhang *et al*. [51].

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 impacted by the treatment. In direct methods one specifies a class of DTRs $\mathscr{D}$ (see below for an example), estimates the Value for each candidate DTR $d \in \mathscr{D}$, say $\hat{V}^d$ and then selects the DTR in $\mathscr{D}$ with maximal estimated Value.

The use of IPTWs for estimating an optimal DTR was pioneered by Robins and colleagues [17, 41]. For a simple example 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, that is, by x. In Robins *et al*. [17] the Value is parameterized as a polynomial function in x and pretreatment variables; for example, $V(x; \beta) = \beta_0 + \beta_1 x + \beta_2 x^2$. 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 β's. In the simplest setting the β's are estimated by solving an IPTW weighted estimating equation. To improve efficiency in the estimation of the β's, Robins *et al*. [17] take advantage of the fact that some individuals' treatment sequence 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 is used to estimate the Value $V(x; \beta)$ for $x = 10, 11, 12$. Operationally, the estimating equation uses three replicates of this individual's data. In the above example the individual is replicated twice to produce three replicates and the replicated outcome Y is relabeled as Y_{10}, Y_{11}, 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 their observed treatment is consistent.

The β's can be estimated by solving the weighted estimating equation

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

where the $\mathbb{P}_n$ is an average over the augmented data set (containing the replicates). Nahum-Shani *et al*. [26], in the context of SMART, provides an intuitive discussion of why replication of participants can be used to account for the fact that a participant's observed treatment is consistent with more than one DTR. The observational data setting can be more complicated; see Robins *et al*. [17] and Shortreed and Moodie [52] for detailed expositions. Related work that compares a range of candidate DTRs by incorporating a treatment-tailoring threshold can be found in Hernán *et al*. [53], Petersen *et al*. [54], van der Laan and Petersen [55], and Cotton and Heagerty [42].

Direct methods for a one-stage decision making setting (e.g. $K = 1$) has seen a great deal of research; here the single decision rule is often called an individualized decision rule. As highlighted by Qian and Murphy [56], the one stage decision making problem has a close connection with classification. Subsequently, methods based on classification [57, 58] have been proposed for estimating the decision rule. Other work in the one-stage decision setting include Cai *et al*. [59] and Imai and Ratkovicz [60].

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 later adapted to statistics. One such procedure that has become particularly popular in the DTR literature is *Q-learning* [21]. Q-learning is an *approximate dynamic programming* method – approximate because the Q-functions are approximated by the use of data and models. 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 multi-stage decision problems [61]. However, one can use more flexible models for the Q-functions, e.g. regression trees [62] or kernels [63]. The version of Q-learning considered in the DTR literature is most similar to the *fitted Q-iteration* algorithm [62] in the reinforcement learning literature.

3.2.1 Q-learning with Linear Models—For clarity, here we will define Q-functions and describe Q-learning for studies with two stages only; generalization to K (≥ 2) stages is straightforward [61]. 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 $(O_1, A_1, O_2, A_2, O_3)$; as before the histories are defined as $H_1 = O_1$ and $H_2 = (O_1, A_1, O_2)$. 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: $Q_2(H_2, A_2) = E[Y_2|H_2, A_2]$, and $Q_1(H_1, A_1) = E[Y_1 + \max_{a_2} Q_2(H_2, a_2)|H_1, A_1]$. A backwards induction argument [21] can be used 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 $d_j(h_j) = \arg\max_{aj} Q_j(h_j, a_j), j = 1, 2$. In practice, the true Q-functions are not known and hence must be estimated. Since Q-functions are conditional expectations, a natural approach to model them is via regression models. A dynamic programming (moving backwards 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 $Q_j(H_j, A_j; \beta_j, \psi_j) = \beta_j^T H_{j0} + \left(\psi_j^T H_{j1}\right) A_j$, where H_{j0} and H_{j1} are two (possibly different) features of the history H_j.

There are many versions of the Q-learning algorithm 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: $\left(\hat{\beta}_2, \hat{\psi}_2\right) = arg\, min_{\beta_2, \psi_2} \frac{1}{n} \sum_{i=1}^{n} (Y_{2i} - Q_2(H_{2i}, A_{2i}, ; \beta_2, \psi_2))^2$.
2. Stage 1 dependent variable: $\hat{Y}_{1i} = Y_{1i} + max_{a_2} Q_2\left(H_{2i}, a_2; \hat{\beta}_2, \hat{\psi}_2\right), i = 1, \ldots, n$.
3. Stage 1 regression:

   $\left(\hat{\beta}_1, \hat{\psi}_1\right) = arg\, min_{\beta_1, \psi_1} \frac{1}{n} \sum_{i=1}^{n} \left(\hat{Y}_{1i} - Q_1(H_{1i}, A_{1i}; \beta_1, \psi_1)\right)^2$.

Note that in step 2. above, the quantity $\hat{Y}_{1i}$ is a predictor of the unobserved random variable $Y_{1i} + \max_{a2} Q_2(H_{2i}, a_2)$, $i = 1, \ldots, n$. The estimated optimal DTR using Q-learning is given by $\left(\hat{d}_1, \hat{d}_2\right)$, where the stage j optimal rule is specified as

$\hat{d}_j(h_j) = arg\, max_{a_j} Q_j\left(h_j, a_1; \hat{\beta}_j, \hat{\psi}_j\right), j = 1, 2$.

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: http://methodology.psu.edu/downloads/procqlearn. Q-learning can be extended for application to observational data by incorporating appropriate adjustments to account for confounding; more precisely, this can be done 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 instead weighting the stage-specific regressions by the inverse of the propensity scores [64]. Q-learning is a version of Robins' optimal *structural nested mean model* [6] developed in the causal inference literature; see Chakraborty *et al*. [23] for a detailed discussion and derivation.

Q-learning has been generalized in a variety of ways. Lizotte *et al*. [65, 66] generalize Q-learning for use when different patients may make different tradeoffs between 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 [67, 68]; 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 there are strong biological, behavioral or social theories that can be employed to guide the formation of the nested conditional models. Once the dynamical systems model is in hand, algorithms from control theory, such as dynamic programming or constrained optimization algorithms are used to estimate the optimal DTR [69]. This is a common approach in applications in engineering, economics and business. In the clinical field there has been much less development. Bayesian methods have been employed in simple, low dimensional problems; one example is Thall *et al*. [70].

Rosenberg *et al*. [71], and Banks *et al*. [72] discussed how a variety of data sources with models based on ordinary differential equations can be used to build a dynamical systems model for use in estimating 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, in a series of presentations available at http://csel.asu.edu/node/13 and papers [69, 73], discussed how common dynamical systems models might be used to describe behaviorial dynamics and thus form the basis for DTRs involving behavioral treatments in obesity and addiction treatment. Gaweda *et al*. [74, 75] discussed the use of control theoretic approaches to anemia management in patients with end-stage renal disease. Bennett and Hauser [76] discussed a framework for simulating clinical decision making from electronic medical records data. In summary, while the dynamical systems approaches to develop DTRs are emerging, from a statistical perspective they still lag behind the other approaches presented earlier; hence this area is ripe for further development.

4 Confidence Sets

High quality measures of confidence are needed in the development of DTRs both for (i) the parameters indexing the optimal DTR; and (ii) the Value of a DTR – either a pre-specified DTR, or an estimated DTR. Inference for the Values of pre-specified regimes has been addressed by numerous authors [12, 13, 14, 9, 10]; 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 future, thus lowering the data collection burden. Second, CIs for the coefficient of the treatment variable can be used to indicate if there is insufficient support in the data to recommend one uniquely best treatment over another, thereby suggesting considerations other than the treatment effect be used to decide on treatment, e.g. cost, patient/clinician familiarity, preference etc.

Orellana *et al*. [41] discussed 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 [6] pointed out that *non-regularity* arises in the indirect estimation of DTRs. By non-regularity, 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 non-regular. 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 non-regularity. The problem can be better understood with a simple but instructive example discussed by Robins [6]; here we present a slightly modified version as presented by Chakraborty *et al*. [23]. Consider the problem of estimating $|\mu|$ based on n i.i.d. observations $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, 1)$. Note that $|\bar{X}_n|$ is the maximum likelihood estimator of $|\mu|$, where $\bar{X}_n$ is the sample average. The asymptotic distribution of $\sqrt{n}\left(|\bar{X}_n| - |\mu|\right)$ 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 μ is abrupt. Thus $|\bar{X}_n|$ is a non-regular estimator of $|\mu|$; an exact proof of non-regularity of this estimator uses local alternatives as in Leeb and Pötscher [77]. Also, for $\mu=0, lim_{n\to\infty} E\left[\sqrt{n}\left(|\bar{X}_n| - |\mu|\right)\right] = \sqrt{\frac{2}{\pi}}$. This asymptotic bias [6] is one symptom of the underlying non-regularity.

Next we review the problem of non-regularity in the context of Q-learning. Suppose we want to construct CIs for the parameters ψ_j's appearing in the model for Q-functions. In a two-stage set-up, the inference for the stage 2 parameters ψ_2 is straightforward since this falls in the standard linear regression framework. In contrast, inference for ψ_1 is complicated. Note that the stage 1 dependent variable in Q-learning for the i-th participant is $\hat{Y}_{1i}=Y_{1i}+max_{a_2} Q_2\left(H_{2i}, a_2; \hat{\beta}_2, \hat{\psi}_2\right)=Y_{1i}+\hat{\beta}_2^T H_{20,i}+|\hat{\psi}_2^T H_{21,i}|, i=1,\ldots,n$, which is a non-differentiable function of $\hat{\psi}_2$ (due to the presence of the absolute value function). Since $\hat{\psi}_1$ is a function of $\hat{Y}_{1i}, i=1,\ldots,n$, it is in turn a non-smooth function of $\hat{\psi}_2$. As a consequence, the distribution of $\sqrt{n}\left(\hat{\psi}_1 - \psi_1\right)$ does not converge uniformly over the parameter space [6]. More specifically, the asymptotic distribution of $\sqrt{n}\left(\hat{\psi}_1 - \psi_1\right)$ is normal if ϕ_2 is such that $p \triangleq P\left[H_2{:}\psi_2^T H_{21}=0\right]=0$, but is non-normal 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 [6], 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 $\phi = (\phi_1, \phi_2)$ and a hypothesis test of $\phi = \phi^0$ for each value of ϕ^0 is well behaved, then a joint $(1 - a)\%$ CI, $\mathscr{C}$ for ϕ can be constructed. This is the case in Q-learning since it is easy to construct a well-behaved hypothesis test statistic when all of the regression coefficients are set to fixed values (the test statistic is based on a quadratic form involving the estimating functions evaluated at the fixed values). Next a projected CI for ϕ_1 is given by $\bigcup_{\psi_2} \{\psi_1 : (\psi_1, \psi_2) \in \mathscr{C}\}$. Unfortunately this interval is very conservative. As a result, Robins [6] using ideas as advanced by Berger and Boos [78] 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 ϕ_1. In this context, ϕ_2 is a nuisance parameter. If the true value of ϕ_2 were known, then the asymptotic distribution of $\sqrt{n}\left(\hat{\psi}_1 - \psi_1\right)$ would be regular (in fact, normal), and standard procedures could be used to construct an asymptotically valid CI. Let $\mathscr{C}(\psi_2)$ denote a $(1 - a)\%$ asymptotic CI for ϕ_1 if ϕ_2 were known. Let $\mathscr{S}$ be a $(1 - \epsilon)\%$ asymptotic CI for ϕ_2. Then, it follows that $\bigcup_{\psi_2 \in \mathscr{S}} \{\psi_1 : \psi_1 \in \mathscr{C}(\psi_2)\}$ is a $(1 - a - \epsilon)\%$ CI for ϕ_1. To see this, note that $P\left(\psi_1 \in \bigcup_{\psi_2 \in \mathscr{S}} \mathscr{C}(\psi_2)\right) \geq 1 - \alpha + o_P(1) + P(\psi_2 \notin \mathscr{S}) = 1 - \alpha - \epsilon + o_P(1)$. Thus, this CI is the union of the CIs $\mathscr{C}(\psi_2)$ over all values $\psi_2 \in \mathscr{S}$, and is an asymptotically valid $(1 - a - \epsilon)\%$ CI for ϕ_1. The main downside of this approach is that it appears to be computationally di cult to implement; to our knowledge this CI has not yet been implemented.

4.2 Adaptive Confidence Intervals

Laber *et al*. [79] developed an *adaptive bootstrap* procedure to construct CIs for linear combinations $c^T \psi_1$, where c is a known vector. In this procedure, they decomposed the asymptotic expansion of $c^T \sqrt{n}\left(\hat{\psi}_1 - \psi_1\right)$ as $\mathbb{W}_n + \mathbb{U}_n$, where the first term $\mathbb{W}_n$ is smooth and asymptotically normally distributed, while the distribution of the second term $\mathbb{U}_n$ depends on the underlying data-generating process "non-smoothly". The adaptive confidence intervals (ACIs) are formed by first constructing smooth data-dependent upper and lower bounds on $\mathbb{U}_n$, and thereby on $c^T \sqrt{n}\left(\hat{\psi}_1 - \psi_1\right)$. The data-dependent upper/lower bounds use a *pretest* [80] that partitions the data into two sets: (i) patients for which there appears to be a treatment effect, and (ii) patients where it appears there is no treatment effect. The pretests are performed using a critical value λ_n, which is a tuning parameter of the procedure and can be varied; Laber *et al*. [79] used $\lambda_n = \log \log n$ in their analysis.

Let the upper and lower bounds on $c^T \sqrt{n}\left(\hat{\psi}_1 - \psi_1\right)$ be given by $\mathscr{U}(c)$ and $\mathscr{L}(c)$ respectively; both these quantities are functions of λ_n. Laber *et al*. [79] showed that the asymptotic distributions of $c^T \sqrt{n}\left(\hat{\psi}_1 - \psi_1\right)$, $\mathscr{U}(c)$ and $\mathscr{L}(c)$ are all equal in the regular case when $p = 0$. That is, when there is a large treatment effect for almost all patients then the bounds are asymptotically tight. However, when there is a non-null subset of patients with no treatment effect, then the asymptotic distribution of $\mathscr{U}(c)$ is stochastically larger than the asymptotic distribution of $c^T \sqrt{n}\left(\hat{\psi}_1 - \psi_1\right)$, and likewise the asymptotic distribution of $\mathscr{L}(c)$ is stochastically smaller. This adaptivity between non-regular and regular settings is a key feature of this procedure. The distributions of $\mathscr{U}(c)$ and $\mathscr{L}(c)$ are approximated using the bootstrap. Let $\hat{u}$ be the $1 - \alpha/2$ quantile of the bootstrap distribution of $\mathscr{U}(c)$, and let $\hat{l}$ be the $\alpha/2$ quantile of the bootstrap distribution of $\mathscr{L}(c)$. Then $\left(c^T\hat{\psi}_1 - \hat{u}/\sqrt{n}, c^T\hat{\psi}_1 - \hat{l}/\sqrt{n}\right)$ is the ACI for $c^T\psi_1$ Laber *et al*. [79] proved the consistency of the bootstrap in this context, and in particular that $P\left(c^T\hat{\psi}_1 - \hat{u}/\sqrt{n} \leq c^T\psi_1 \leq c^T\hat{\psi}_1 - \hat{l}/\sqrt{n}\right) \geq 1 - \alpha + o_P(1)$, 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 non-smooth functionals [81]. This method is the same as the ordinary bootstrap except that the resample size (m) satisfies: $m \to \infty$ as $n \to \infty$, but $m = o(n)$. Chakraborty *et al*. [82] proposed a data-driven method for choosing m in the context of Q-learning that is directly connected to an estimated degree of non-regularity. 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*. [82] considered a class of resample sizes of the form $m = n^{\frac{1+\eta(1-p)}{1+\eta}}$, where $\eta > 0$ is a tuning parameter. For implementation, one first needs to estimate p using a plug-in estimator, $\hat{p} = \mathbb{P}_n \mathbb{I}\left[n\left(H_{21}^T\hat{\psi}_2\right)^2 \leq \left(H_{21}^T\hat{\Sigma}_{\hat{\psi}_2}H_{21}\right)\cdot\chi^2_{1,1-\nu}\right]$, where $n^{-1}\hat{\Sigma}_{\hat{\psi}_2}$ is the plug-in estimator of the asymptotic covriance matrix of $\hat{\psi}_2$ and $\chi^2_{1,1-\nu}$ is the $(1 - \nu) \times 100$ percentile of a χ^2 distribution with 1 degree of freedom. Then the data-driven choice of the resample size is given by $\hat{m} = n^{\frac{1+\eta(1-\hat{p})}{1+\eta}}$. Note that for fixed $\hat{m}$ is a monotone decreasing function of $\hat{p}$, taking values in the interval $\left[n^{\frac{1}{1+\eta}}, n\right]$. Thus, η governs the smallest acceptable resample size. The procedure has been shown to be robust to the choice of ν. Once $\hat{m}$ is computed, a $(1 - \alpha) \times 100\%$ *m*-out-of-*n* bootstrap CI for $c^T\psi_1$ is given by $\left(c^T\hat{\psi}_1 - \hat{u}/\sqrt{\hat{m}}, c^T\hat{\psi}_1 - \hat{l}/\sqrt{\hat{m}}\right)$ where $\hat{l}$ and $\hat{u}$ are the $(\alpha/2) \times 100$ and $(1 - \alpha/2) \times 100$

percentiles of $c^T\sqrt{m}\left(\hat{\psi}_1^{(b)}-\hat{\psi}_1\right)$ respectively $\left(\hat{\psi}_1^{(b)}\right.$ is the m-out-of-n bootstrap analog of $\left.\hat{\psi}_1\right)$. This bootstrap procedure is consistent, and $P\left(c^T\hat{\psi}_1-\hat{u}/\sqrt{\hat{m}}\le c^T\psi_1\le c^T\hat{\psi}_1-\hat{l}/\sqrt{\hat{m}}\right)\ge 1-\alpha+o_P(1)$, 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 on http://cran.r-project.org/web/packages/qLearn/index.html.

See the online Supplemental Materials for a simulation study that illustrates the performance of the above approaches to forming a CI.

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 some insight can be gained by exploiting its connection with classification. As highlighted by Qian and Murphy [56] and Zhao *et al*. [58], 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 di cult due to the inherent non-smoothness; standard methods like normal approximation or usual bootstrap fail. Laber and Murphy [83] developed 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. While 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

Dynamic treatment regimes comprise an increasingly active area of current statistical research with 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 case of DTRs this requires the design and analysis of cluster-randomized SMARTs. At the design level, cluster randomization would imply increased sample size requirements due to intra-class correlation. At the analysis level, it would open up questions as to how best to incorporate random effects models or generalized estimating equations into the existing framework of estimation, how the intra-class correlation would impact the non-regularity in inference, and so on. 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 slightly different issues than similar topics in the prediction literature are under-developed, and warrant further research.

Inference in the domain of DTRs is a particularly challenging problem due to non-regularity 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. Optimality principles and statistical methods aiming to achieve optimal CIs in these non-regular problems is an open area of research. There is growing interest in confidence intervals 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. Confidence intervals for this type of parameter is as yet undeveloped.

In today's health care, there is an increasing use of sophisticated mobile devices (e.g. smart phones, actigraph units containing accelerometers, etc.) to remotely monitor patients' chronic conditions and to intervene, when needed. This 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 future research direction.

The field of DTRs is in its infancy and 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. In addition, dissemination of the newly developed methods into the medical domains and collaboration with clinical scientists will be crucial.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Chakraborty's research is supported by the NIH grant R01 NS072127-01A1. Dr. Murphy's research is supported by the NIH grant P50DA010075.

References

[1]. Robins J. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. Mathematical Modelling. 1986; 7:1393–1512.

[2]. Robins, J. Health Service Research Methodology: A Focus on AIDS. Sechrest, L.; Freeman, H.; Mulley, A., editors. NCHSR, U.S. Public Health Service; New York: 1989.

[3]. Robins J. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. Proceedings of the Biopharmaceutical Section, American Statistical Association. 1993:24–33.

[4]. Robins, J. Latent Variable Modeling and Applications to Causality: Lecture Notes in Statistics. Berkane, M., editor. Springer-Verlag; New York, NY: 1997.

[5]. Murphy S. Optimal dynamic treatment regimes (with discussions). Journal of the Royal Statistical Society. 2003; 65:331–366.Series B

[6]. Robins, J. Proceedings of the Second Seattle Symposium on Biostatistics. Lin, D.; Heagerty, P., editors. Springer; New York: 2004.

[7]. Lavori P, Dawson R. A design for testing clinical strategies: Biased adaptive within-subject randomization. Journal of the Royal Statistical Society. 2000; 163:29–38.Series A

[8]. Lavori P, Dawson R. Adaptive treatment strategies in chronic disease. Annual Review of Medicine. 2008; 59:443–453.

[9]. Thall P, Millikan R, Sung H. Evaluating multiple treatment courses in clinical trials. Statistics in Medicine. 2000; 30:1011–1128. [PubMed: 10790677]

[10]. Thall P, Sung H, Estey E. Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. Journal of the American Statistical Association. 2002; 97:29–39.

[11]. Murphy S. An experimental design for the development of adaptive treatment strategies. Statistics in Medicine. 2005a; 24:1455–1481. [PubMed: 15586395]

[12]. Lunceford J, Davidian M, Tsiatis A. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. Biometrics. 2002; 58:48–57. [PubMed: 11890326]

[13]. Wahed A, Tsiatis A. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomized designs in clinical trials. Biometrics. 2004; 60:124–133. [PubMed: 15032782]

[14]. Wahed A, Tsiatis A. Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. Biometrika. 2006; 93:163–177.

[15]. Wagner E, Austin B, Davis C, Hindmarsh M, Schaefer J, Bonomi A. Improving chronic illness care: Translating evidence into action. Health A airs. 2001; 20:64–78.

[16]. Rosthøj S, Fullwood C, Henderson R, Stewart S. Estimation of optimal dynamic anticoagulation regimes from observational data: A regret-based approach. Statistics in Medicine. 2006; 25:4197–4215. [PubMed: 16981226]

[17]. Robins JM, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. Statistics in Medicine. 2008; 27:4678–4721. [PubMed: 18646286]

[18]. Wang L, Rotnitzky A, Lin X, Millikan R, Thall P. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. Journal of the American Statistical Association. 2012; 107:493–508. [PubMed: 22956855]

[19]. Bellman, R. Dynamic programming. Princeton University Press; Princeton: 1957.

[20]. Kulkarni K, Gosavi A, Murray S, Grantham K. Semi-markov adaptive critic heuristics with application to airline revenue management. Journal of Control Theory and Applications. 2011; 9:421–430.

[21]. Sutton, R.; Barto, A. Reinforcement learning: An introduction. MIT Press; Cambridge: 1998.

[22]. Lavori P, Dawson R. Dynamic treatment regimes: Practical design considerations. Clinical Trials. 2004; 1:9–20. [PubMed: 16281458]

[23]. Chakraborty B, Murphy S, Strecher V. Inference for non-regular parameters in optimal dynamic treatment regimes. Statistical Methods in Medical Research. 2010; 19:317–343. [PubMed: 19608604]

[24]. Kasari, C. Developmental and augmented intervention for facilitating expressive language (ccnia). National Institutes of Health; Bethesda, MD: 2009. http://clinicaltrials.gov/ct2/show/NCT01013545?term=kasari\&rank=5

[25]. Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy S. A SMART design for building individualized treatment sequences. The Annual Review of Clinical Psychology. 2011

[26]. Nahum-Shani I, Qian M, Almiral D, Pelham W, Gnagy B, et al. Experimental design and primary data analysis methods for comparing adaptive interventions. Psychological Methods. 2012a; 17:457–477. [PubMed: 23025433]

[27]. Nahum-Shani I, Qian M, Almiral D, Pelham W, Gnagy B, et al. Q-learning: A data analysis method for constructing adaptive interventions. Psychological Methods. 2012b; 17:478–494. [PubMed: 23025434]

[28]. Jones, H. Reinforcement-based treatment for pregnant drug abusers (HOME II). National Institutes of Health; Bethesda, MD: 2010. http://clinicaltrials.gov/ct2/show/NCT01177982?term=jones+pregnant\&rank=9

[29]. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974; 66:688–701.

[30]. Rubin D. Discussion of "Randomized analysis of experimental data: The Fisher randomization test" by D. Basu. Journal of the American Statistical Association. 1980; 75:591–593.

[31]. Box, G.; Hunter, W.; Hunter, J. Statistics for experimenters: An introduction to design, data analysis, and model building. Wiley; New York: 1978.

[32]. Murphy S, Bingham D. Screening experiments for developing dynamic treatment regimes. Journal of the American Statistical Association. 2009; 184:391–408. [PubMed: 20589222]

[33]. Oetting, A.; Levy, J.; Weiss, R.; Murphy, S. Causality and Psychopathology: Finding the Deter minants of Disorders and their Cures. Shrout, P.; Shrout, P.; Keyes, K.; Ornstein, K., editors. American Psychiatric Publishing, Inc.; Arlington, VA: 2011.

[34]. Dawson R, Lavori P. Sample size calculations for evaluating treatment policies in multi-stage designs. Clinical Trials. 2010; 7:643–652. [PubMed: 20630903]

[35]. Dawson R, Lavori P. E cient design and inference for multistage randomized trials of individualized treatment policies. Biostatistics. 2012; 13:142–152. [PubMed: 21765180]

[36]. Feng W, Wahed A. Supremum weighted log-rank test and sample size for comparing two-stage adaptive treatment strategies. Biometrika. 2008; 95:695–707.

[37]. Feng W, Wahed A. Sample size for two-stage studies with maintenance therapy. Statistics in Medicine. 2009; 28:2028–2041. [PubMed: 19382105]

[38]. Li Z, Murphy S. Sampe size formulae for two-stage randomized trials with survival outcomes. Biometrika. 2011 DOI: 10.1093/biomet/asr019.

[39]. Coffey C, Levin B, Clark C, Timmerman C, Wittes J, et al. Overview, hurdles, and future work in adaptive designs: Perspectives from an nih-funded workshop. Clinical Trials. 2012; 9:671–680. [PubMed: 23250942]

[40]. Thall P, Wathen J. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. Statistics in Medicine. 2005; 24:1947–1964. [PubMed: 15806621]

[41]. Orellana L, Rotnitzky A, Robins JM. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part I: Main content. The International Journal of Biostatistics. 2010a; 6

[42]. Cotton C, Heagerty P. A data augmentation method for estimating the causal effect of adherence to treatment regimens targeting control of an intermediate measure. Statistics in Bioscience. 2011; 3:28–44.

[43]. Moodie E, Platt R, Kramer M. Estimating response-maximized decision rules with applications to breastfeeding. Journal of the American Statistical Association. 2009; 104:155–165.

[44]. Van der Laan MJ, Petersen ML. Statistical learning of origin-specific statically optimal individualized treatment rules. The International Journal of Biostatistics. 2007a; 3

[45]. Robins J. Correcting for non-compliance in randomized trials using structural nested mean models. Communications in Statistics. 1994; 23:2379–2412.

[46]. Robins, J. Statistical Models in Epidemiology, the Environment, and Clinical Trials. Halloran, ME.; Berry, D., editors. Vol. 116 of IMA. Springer; New York, NY: 1999.

[47]. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000; 11:550–60. [PubMed: 10955408]

[48]. Murphy S, der Laan MV, Robins J, CPPRG. Marginal mean models for dynamic regimes. Journal of the American Statistical Association. 2001; 96:1410–1423. [PubMed: 20019887]

[49]. Miyahara S, Wahed A. Weighted kaplanmeier estimators for two-stage treatment regimes. Statistics in Medicine. 2010; 29:2581–2591. [PubMed: 20799259]

[50]. Orellana L, Rotnitzky A, Robins JM. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part II: Proofs and additional results. The International Journal of Biostatistics. 2010b; 6

[51]. Zhang B, Tsiatis A, Laber E, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012a; 68:1010–1018. [PubMed: 22550953]

[52]. Shortreed S, Moodie E. Estimating the optimal dynamic antipsychotic treatment regime: Evidence from the sequential-multiple assignment randomized CATIE Schizophrenia Study. Journal of the Royal Statistical Society. 2012; 61:577–599. [PubMed: 23087488] Series C

[53]. Hernán MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. Basic & Clinical Pharmacology & Toxicology. 2006; 98:237–242. [PubMed: 16611197]

[54]. Petersen ML, Deeks SG, Van der Laan MJ. Individualized treatment rules: Generating candidate clinical trials. Statistics in Medicine. 2007; 26:4578–4601. [PubMed: 17450501]

[55]. Van der Laan MJ, Petersen ML. Causal effect models for realistic individualized treatment and intention to treat rules. The International Journal of Biostatistics. 2007b; 3

[56]. Qian M, Murphy S. Performance guarantees for individualized treatment rules. Annals of Statistics. 2011; 39:1180–1210. [PubMed: 21666835]

[57]. Zhang B, Tsiatis A, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. Stat. 2012b; 1:103–114. [PubMed: 23645940]

[58]. Zhao Y, Zeng D, Rush A, Kosorok M. Estimating individual treatment rules using outcome weighted learning. Journal of the American Statistical Association. 2012; 107:1106–1118. [PubMed: 23630406]

[59]. Cai T, Tian L, Wong P, Wei L. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011; 12:270–282. [PubMed: 20876663]

[60]. Imai K, Ratkovicz M. Estimating treatment effect heterogeneity in randomized program evaluation. Annals of Applied Statistics. 2013; 7:443–470.

[61]. Murphy S. A generalization error for Q-learning. Journal of Machine Learning Research. 2005b; 6:1073–1097. [PubMed: 16763665]

[62]. Ernst D, Geurts P, Wehenkel L. Tree-based batch mode reinforcement learning. Journal of Ma chine Learning Research. 2005; 6:503–556.

[63]. Ormoneit D, Sen S. Kernel-based reinforcement learning. Machine Learning. 2002; 49:161–178.

[64]. Moodie E, Chakraborty B, Kramer M. Q-learning for estimating optimal dynamic treatment rules from observational data. Canadian Journal of Statistics. 2012; 40:629–645. [PubMed: 23355757]

[65]. Lizotte, D.; Bowling, M.; Murphy, S. Twenty-Seventh International Conference on Machine Learning (ICML). Omnipress; Haifa, Israel: 2010.

[66]. Lizotte D, Bowling M, Murphy S. Linear fitted-Q iteration with multiple reward functions. Journal of Machine Learning Research. 2012; 13:3253–3295. [PubMed: 23741197]

[67]. Zhao Y, Zeng D, Socinski M, Kosorok M. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. Biometrics. 2011; 67:1422–1433. [PubMed: 21385164]

[68]. Goldberg Y, Kosorok M. Q-learning with censored data. The Annals of Statistics. 2012; 40:529–560.

[69]. Rivera D, Pew M, Collins L. Using engineering control principles to inform the design of adaptive interventions: A conceptual introduction. Drug and Alcohol Dependence. 2007; 88:S31–S40. [PubMed: 17169503]

[70]. Thall PF, Logothetis C, Pagliaro LC, Wen S, Brown MA, et al. Adaptive therapy for androgen-independent prostate cancer: A randomized selection trial of four regimens. Journal of the National Cancer Institute. 2007; 99:1613–1622. [PubMed: 17971530]

[71]. Rosenberg E, Davidian M, Banks H. Using mathematical modeling and control to develop structured treatment interruption strategies for hiv infection. Drug and Alcohol Dependence. 2007; 88:S41–S51. [PubMed: 17276624]

[72]. Banks H, Jang T, Kwon H. Feedback control of hiv antiviral therapy with long measurement time. International Journal of Pure and Applied Mathematics. 2011; 66:461–485.

[73]. Navarro-Barrientos J, Rivera D, Collins L. A dynamical model for describing behavioural interventions for weight loss and body composition change. Mathematical and Computer Modelling of Dynamical Systems. 2011; 17:183–203. [PubMed: 21673826]

[74]. Gaweda A, Muezzinoglu M, Arono G, Jacobs A, Zurada J, Brier M. Individualization of pharmacological anemia management using reinforcement learning. Neural Networks. 2005; 18:826–834. [PubMed: 16109475]

[75]. Gaweda A, Jacobs A, Arono G, Brier M. Model predictive control of erythropoietin administrateion in the anemia of esrd. American Journal of Kidney Diseases. 2008; 51:71–79. [PubMed: 18155535]

[76]. Bennett C, Hauser K. Artificial intelligence framework for simulating clinical decision-making: A markov decision process approach. Artificial Intelligence in Medicine. 2012

[77]. Leeb H, Pötscher B. The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. Econometric Theory. 2003; 19:100–142.

[78]. Berger R, Boos D. P values maximized over a confidence set for the nuisance parameter. Journal of the American Statistical Association. 1994; 89:1012–1016.

[79]. Laber E, Qian M, Lizotte D, Murphy S. Statistical inference in dynamic treatment regimes. arXiv:1006.5831v2 [stat.ME]. 2011

[80]. Olshen R. The conditional level of the f-test. Journal of the American Statistical Association. 1973; 68:692–698.

[81]. Shao J. Bootstrap sample size in nonregular cases. Proceedings of the American Mathematical Society. 1994; 122:1251–1262.

[82]. Chakraborty B, Laber E, Zhao Y. Inference for optimal dynamic treatment regimes using an adaptive *m*-out-of-*n* bootstrap scheme. Biometrics. 2013 In press.

[83]. Laber E, Murphy S. Adaptive confidence intervals for the test error in classification. Journal of the American Statistical Association. 2011; 106:904–913. [PubMed: 22053123]

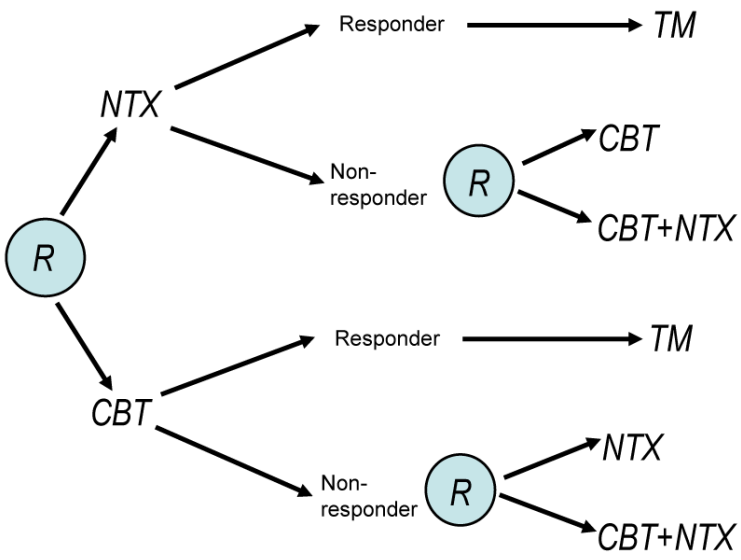

Figure 1.
Hypothetical SMART design schematic for the addiction management example (an "R" within a circle denotes randomization at a critical decision point).