Dynamic Treatment Regimes

Min Qian[1,*], Inbal Nahum-Shani[2] and Susan A. Murphy[1]

[1] Department of Statistics, University of Michigan

439 West Hall, 1085 South University Ave., Ann Arbor, MI, 48109

[2] The Methodology Center, Pennsylvania State University

204 E. Calder Way, Suite 400, State College, PA, 16801

[*] Corresponding author

Email: minqian@umich.edu, *Fax:* 734-763-4676, *Phone:* 734-276-4305

Abstract

In recent years, treatment and intervention scientists increasingly realize that individual heterogeneity in disorder severity, background characteristics and co-occurring problems translates into heterogeneity in response to various aspects of any treatment program. Accordingly, research in this area is shifting from the traditional "one-size-fits-all" treatment to dynamic treatment regimes, which allow greater individualization in programming over time. A dynamic treatment regime is a sequence of decision rules that specify how the dosage and/or type of treatment should be adjusted through time in response to an individual's changing needs, aiming to optimize the effectiveness of the program. In the chapter we review the Sequential Multiple Assignment Randomized Trials (SMART), which is an experimental design useful for the development of dynamic treatment regimes. We compare the SMART approach with other experimental approaches and discuss data analyses methods for constructing a high quality dynamic treatment regime as well as other secondary analyses.

1 Introduction

Recent research (see Lavori and Dawson 2000, 2004) stresses the need to take into account patients' heterogeneity in need for treatment when developing intervention programs. In order to improve patient care the type of treatment and the dosage should vary by patients. Additionally, in many cases, the need for treatment may change over time, yielding repeated opportunities to adapt the intervention. For example, patients with mental illnesses (e.g. depression, drug-abuse, alcoholism, etc) often respond differently to treatment and also tend to experience repeated cycles of cessation and relapse (see McLellan 2002; Fava et al. 2003 for examples). Therefore, the clinical management of mental illnesses requires that clinicians make a sequence of treatment decisions, where the first step is aimed at stabilizing the patient and the following steps are directed to preventing relapse in the long term. Dynamic treatment regimes operationalize this sequential decision making. A dynamic treatment regime is a sequence of decision rules, one per treatment decision, that provide the mechanism by which patient's values on key characteristics, called tailoring variables are translated into dosage amount and type. Instead of delivering the same type and dosage of treatment to every patient, a dynamic treatment regime assigns different treatment types/dosages across patients and within each patient across time according to the patient's values on tailoring variables. The term 'dynamic treatment regimes' is also known as 'adaptive treatment strategies' (Lavori and Dawson 2000; Murphy 2005a), 'multi-stage treatment strategies' (Thall et al. 2002; Thall and Wathen 2005), 'treatment policies' (Lunceford et al. 2002; Wahed and Tsiatis 2004, 2006) or 'individualized treatment rules' (Petersen et al. 2007; van der Laan and Petersen 2007). All are aimed at constructing a sequence of decision rules that when implemented will produce the best long term outcome.

Better understanding of dynamic treatment regimes can be gained by considering the following example. This example demonstrates a sequential decision making problem in the area of clinical science. It will be used throughout the chapter.

Addiction management example:

Suppose in planning the treatment for alcohol dependent patients we are particularly interested in making two critical decisions. First we must decide what is the best initial treatment for an alcohol

dependent patient. For example, we may consider two possible treatment options: opiate-antagonist Naltrexone (NTX) and Combined Behavioral Intervention (CBI; Miller 2004). Second, we must decide what is the best subsequent treatment for non-improving patients (i.e. non-responders) and improving patients (i.e. responders). For example, if a patient is a non-responder to NTX, we need to decide whether to augment NTX with CBI (NTX+CBI) or switch the treatment to CBI. If a patient is a non-responder to CBI, we need to decide whether to augment CBI with NTX (CBI+NTX) or switch to NTX. If a patient is a responder, we will refer the patient to a 12-step program (Alcoholics Anonymous 2001), but we need to decide if it is worthwhile to augment the 12-step program with Telephone Disease Management (TDM; Oslin et al. 2003) for an additional period of six months.

The above example is inspired by the ExTENd trial conducted by David Oslin from the University of Pennsylvania (personal communication) and the COMBINE trial conducted by COMBINE Study Research Group (2003). In the context of this example, two simple dynamic treatment regimes may be:

- Regime A: Treat patients with NTX first; then provide CBI for non-responders and refer to the 12-step program for responders.
- Regime B: Treat patients with CBI first; then provide NTX+CBI for non-responders and refer to the 12-step program for responders.

The above two dynamic treatment regimes tailor the subsequent treatment to each patient using the patient's response status to the initial treatment. More complex dynamic treatment regimes may use a patient's pretreatment information (e.g. medical history, severity of addiction and co-occurring disorders) to choose the initial treatment and/or use intermediate outcomes (e.g. the patient's response status, side effects and adherence to the initial treatment) to choose the subsequent treatment.

In order to further clarify the concept of dynamic treatment regimes, in the next section we use the potential outcome framework to define dynamic treatment regimes and the optimal dynamic treatment regime. In section 3, we introduce the SMART design proposed by Lavori and Dawson (2000, 2004) and Murphy (2005a), and discuss the motivation behind it. In section 3.3, we compare the SMART design with adaptive experimental designs. In section 4, we discuss commonly used methods for

developing the optimal dynamic treatment regime using data collected from a SMART trial. More specifically, we focus on Q-learning which is a well-known regression-based method (Murphy 2005b). In section 5, we discuss other analyses related to the Q-learning algorithm. Finally, we summarize the chapter and discuss challenges and open questions in section 6.

2 Potential outcomes framework

Potential outcomes were introduced in Neyman (1923) to analyze causal effects of time-independent treatments in randomized studies. Extensions of Neyman's work to the analysis of causal effects of time-independent treatments in observational studies can be found in Rubin (1978). A formal theory of causal inference was proposed in Robins (1986, 1987) to assess the direct and indirect effects of time-varying treatments from experimental and observational data. In this section, we use potential outcome models to provide a framework for developing the optimal dynamic treatment regime. Later we explain how inference concerning potential outcomes can be made, using data from experimental trials.

For simplicity, assume there are only two decision points. The extension to multiple decision points is straight forward. Denote the treatment decision at the initial and secondary stage by a_1 and a_2, respectively. Let S_1 denote the pretreatment information. For each fixed value of the treatment sequence (a_1, a_2), we conceptualize potential outcomes denoted by $S_2(a_1)$ and $Y(a_1, a_2)$, where $S_2(a_1)$ is the intermediate outcome (ongoing information) that would have been observed prior to the secondary decision point if the initial treatment assignment were a_1, and $Y(a_1, a_2)$ is the primary outcome (large values are desirable) that an individual would have if he/she were assigned the treatments (a_1, a_2). In this context, a *dynamic treatment regime* is a sequence of decision rules, (d_1, d_2), where d_1 takes S_1 as input and outputs a treatment a_1 and d_2 takes $(S_1, a_1, S_2(a_1))$ as input and outputs a treatment a_2.

Let $\mathcal{A}$ be the collection of all possible treatment sequences. Then the set of all potential outcomes is $\{(S_1, S_2(a_1), Y(a_1, a_2)) : (a_1, a_2) \in \mathcal{A}\}$ (S_1 is included for completeness). The mean primary outcome for regime (d_1, d_2), denoted by $\mu_{(d_1,d_2)}$, is defined as the average primary outcome that would be

observed if the entire study population were assigned (d_1, d_2). Mathematically, this is

$$\mu_{(d_1,d_2)} = E[Y(a_1, a_2)_{a_1=d_1(S_1),a_2=d_2(S_1,a_1,S_2(a_1))}], \tag{1}$$

where the expectation is taken with respect to the multivariate distribution of $(S_1, S_2(a_1), Y(a_1, a_2))$ for the treatment sequence determined by (d_1, d_2). The goal is to develop a dynamic treatment regime that leads to the maximal $\mu_{(d_1,d_2)}$ (as compared to other possible dynamic treatment regimes). This dynamic treatment regime is called the *optimal dynamic treatment regime*. Note that (1) can be written as a repeated expectation

$$\mu_{(d_1,d_2)} = E\left[E\left[E[Y(a_1, a_2)|S_1, S_2(a_1)]_{a_2=d_2(S_1,a_1,S_2(a_1))}|S_1\right]_{a_1=d_1(S_1)}\right].$$

If we knew the distribution of the potential outcomes $(S_1, S_2(a_1), Y(a_1, a_2))$ for each treatment pattern $(a_1, a_2) \in \mathcal{A}$, classical algorithms based on backwards induction (dynamic programming) (Bellman 1957) could be used to find the optimal sequence of decision rules. The optimal secondary decision rule d_2^* maximizes the mean primary outcome at the second decision point, i.e.

$$d_2^*(s_1, a_1, s_2) \in \arg\max_{a_2} E[Y(a_1, a_2)|S_1 = s_1, S_2(a_1) = s_2].$$

Note that we use "$\in$" in the above formula since there may be multiple treatments that maximize the mean primary outcome given $S_1 = s_1$ and $S_2(a_1) = s_2$. The optimal initial decision rule d_1^* then chooses the treatment that maximizes the mean primary outcome given that d_2^* is followed at the second decision point, i.e.

$$d_1^*(s_1) \in \arg\max_{a_1} E\Big[\max_{a_2} E[Y(a_1, a_2)|S_1, S_2(a_1)]\Big|S_1 = s_1\Big].$$

The sequence of decision rules (d_1^*, d_2^*) is the optimal dynamic treatment regime. And $\mu_{(d_1^*,d_2^*)}$ is the optimal mean primary outcome. Note that the treatment options at each decision point may depend on a patient's ongoing information and/or previous treatment. In the addiction management example, the treatment options at the second decision point depend on a patient's initial treatment

and/or whether the patient had improved or not. In the above formulae the maximization at each decision point is taken over all possible treatment options at that point.

In general the multivariate distribution of the potential outcomes for each of the possible treatment patterns is unknown; thus we can not directly use the arguments given above to construct the optimal dynamic treatment regime. Accordingly, Murphy (2005a) proposed experimental trials by which data can be obtained and used for the formulation of decision rules. In the next section, we introduce this Sequential Multiple Assignment Randomized Trials (SMART) and discuss the motivation for developing this approach for the formulation of dynamic treatment regimes.

3 SMART design

In a SMART trial, each subject may be randomly assigned to treatments several times. More specifically, this is a multi-stage randomized trial, in which each subject progresses through stages of treatments and is randomly assigned to treatments at each stage. This type of design was first introduced by Lavori and Dawson (2000), and was named 'biased coin adaptive within-subject' (BCAWS) design. TenHave et al. (2003) compared the BCAWS design with other designs. Lavori and Dawson (2004) discussed practical considerations for the design. Murphy (2005a) proposed the general framework of the SMART design.

Trials in which each subject is randomized multiple times have been widely used, especially in cancer research (see for example, CALGB study 8923 for treating elderly patients with primary acute myelogenous leukemia (Stone et al. 1995)). Precursors of SMART trials include the CATIE trial for antipsychotic medications in patients with Alzheimer's (Schneider et al. 2001) and STAR*D for treatment of depression (Lavori et al. 2001; Fava et al. 2003). In recent years, a number of SMART trials have been conducted. These include phase II trials at MD Anderson for treating cancer (Thall et al. 2000), the ExTENd trial concerning alcohol dependence by David Oslin from the University of Pennsylvania (personal communication) and the ADHD trial conducted by William Pelham from the University at Buffalo, SUNY (personal communication).

To make the discussion more concrete we consider a SMART trial based on our addiction management example (see Figure 1). In this trial, each subject is randomly assigned to one of two possible

initial treatments (CBI or NTX). Then in the next two months clinicians record heavy drinking, adherence, side effects and other intermediate outcomes. If at any time during this two-month period the subject experiences a third heavy drinking day, he/she is classified as a non-responder to the initial treatment. As soon as the subject is classified as a non-responder he/she is re-randomized to one of the two subsequent treatments, depending on his/her initial treatment assignment: NTX+CBI or CBI alone for NTX non-responders; NTX+CBI or NTX alone for CBI non-responders. However, if the subject is able to avoid more than two heavy drinking days during the two-month period, he/she is considered as a responder to the initial treatment. In this case, the subject is re-randomized to one of the following two possible treatments for an additional period of six months: 12-step program or 12-step program+TDM. The goal of the study is to maximize the number of non-heavy drinking days over a 12 month study period.

[Figure 1 near here]

This experimental approach is motivated by several disadvantages of traditional "single-stage" experimental designs discussed in the following section.

3.1 Motivation for the SMART design

Most randomized trials are used to compare single stage treatments. To ascertain the best treatment sequence, an alternative approach to SMART is to conduct multiple randomized trials; each trial compares available treatment options at each stage based on results from previous trials and/or based on historical trials and the available literature. For example, instead of the SMART trial for our addiction management study, the researcher may conduct two single-stage randomized trials. The first trial compares the initial treatments (CBI versus NTX). Based on the results of the first trial, the researcher chooses the best treatment and moved on to the second trial where all subjects are initially offered with the chosen treatment and then responders are randomized to one of the two possible conditions: 12-step program or 12-step program+TDM; non-responders are randomized to one of the two possible conditions: augment current treatment with another type of treatment or switch to

another type of treatment. However, this approach has at least three disadvantages as compared to a SMART trial when used to optimize dynamic treatment regimes.

First, this approach may fail to detect possible delayed effects that are cases where a treatment has an effect that is less likely to occur unless it is followed by a particular subsequent treatment. For example in the addiction management study, the first trial may indicate that NTX performs the same as CBI on average. And, the second trial, in which NTX was chosen to be the initial treatment based on the results from the previous trial (since CBI is more expensive then NTX), may indicate that CBI along performs as well as NTX+CBI for non-responders (and avoids further side effects due to NTX for non-responders who experience side effects) and the 12-step program is as effective as 12-step+TDM for responders. In that case, the researcher may conclude that regime A described in section 1 is the best treatment sequence. However, it is possible that NTX was found to be as effective as CBI during the first trial because for many subjects CBI works best if used over a longer period of time. If the researcher had chosen CBI instead of NTX as the initial treatment for the second trial, then NTX+CBI might be much more effective than NTX alone for non-responders and the 12-step program may be sufficient for responders. That is, regime B described in section 1 might have been more effective than regime A though NTX may initially appear to be as effective as CBI. Hence using the single-stage experimental approach may reduce the ability of the researcher to detect the delayed effects and eventually led to the wrong conclusion as to the most effective regime.

Second, although the results of the first trial may indicate that one treatment is initially less effective than the other, the former treatment may elicit diagnostic information that permits the researcher to better match the subsequent treatment to each subject, and thus improve the primary outcome.

Third, subjects who enroll and remain in the single-stage trials may be different than those who enroll and remain in a SMART trial. This is a type of cohort effects (see Murphy et al. 2007). Consider a one stage randomized trial in which CBI is compared with NTX. For subjects who are not improving, there are no other options besides dropping out of the study. However, in a SMART trial non-responding subjects know that their treatmens will be altered. Consequently non-responding subjects may be less likely to drop out from a SMART trial relative to a one stage randomized trial.

Thus the choice of initial treatment based on the single-stage approach may be based on a sample that is less representative of the study population relative to the SMART trial.

From the above discussion, we see that conducting separate trials for different stages and examining treatment options at each stage separately from other stages may fail to detect delayed effects and diagnostic effects, and may result in deleterious cohort effects. As a result, the developed sequence of treatment decisions may not be the best choice. This has led researchers to consider designs in which each individual is randomized multiple times, one per critical decision, so as to be able to link different stages in the decision process, and improve the ability to develop dynamic treatment regimes in which sequential treatments are synergetic.

3.2 Design aspects

Denote the observable data for a subject in a SMART trial by (S_1, A_1, S_2, A_2, Y), where S_1 and S_2 are the pretreatment information and intermediate outcomes, A_1 and A_2 are the randomly assigned initial and secondary treatments, and Y is the primary outcome of the subject, respectively. For example, in the addiction management study (see Figure 1), S_1 may include addiction severity and co-morbid conditions, S_2 may include the subject's 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. Under Robins' consistency assumption (i.e. a subject's treatment assignment does not effect other subjects' outcomes; see Robins and Wasserman (1997)), the potential outcomes are connected to the subject's data by $S_2 = S_2(A_1)$ and $Y = Y(A_1, A_2)$.

The treatment randomization probabilities in a SMART trial are allowed to depend on past history (i.e. past information and treatment). That is the randomization probabilities for A_1 and A_2 may depend on S_1 and (S_1, A_1, S_2), respectively. Thus data from a SMART trial satisfies the "sequential ignorability" assumption (i.e. at each decision point the current treatments are assigned independently of potential future responses to treatment, conditional on the history of treatments and response to date (Robins 2004)). Under the "sequential ignorability" assumption, the conditional distributions of the potential outcomes are the same as the corresponding conditional distributions of the observable

data. That is,

$$P(S_2(a_1) \leq s_2 | S_1 = s_1) = P(S_2 \leq s_2 | S_1 = s_1, A_1 = a_1)$$

and $$P(Y(a_1, a_2) \leq y | S_1 = s_1, S_2(a_1) = s_2) = P(Y \leq y | S_1 = s_1, A_1 = a_1, S_2 = s_2, A_2 = a_2).$$

Thus the mean primary outcome of a dynamic treatment regime, (d_1, d_2), can be written as a function of the multivariate distribution of the observable data:

$$\mu_{(d_1,d_2)} = E\left[E\left[E[Y|S_1, A_1, S_2, A_2 = d_2(S_1, A_1, S_2)]|S_1, A_1 = d_1(S_1)\right]\right].$$

Hence we can evaluate the effect of a dynamic treatment regime or estimate the optimal dynamic treatment regime using data from a SMART trial. In particular, define the Q-function at the secondary decision point by

$$Q_2^*(s_1, a_1, s_2, a_2) = E[Y|S_1 = s_1, A_1 = a_1, S_2 = s_2, A_2 = a_2], \tag{2}$$

and the Q-function at the initial decision point by

$$Q_1^*(s_1, a_1) = E\Big[\max_{a_2} Q_2^*(S_1, A_1, S_2, a_2)|S_1 = s_1, A_1 = a_1\Big]. \tag{3}$$

Each Q-function measures the quality of the current treatment for patients with the specified past information and treatment assignments. Then the optimal decision rule at each decision point chooses treatment that maximizes the corresponding Q-function.

To power a SMART trial, we need to specify a primary research question. This research question may concern some components of dynamic treatment regimes (e.g. testing the main effect of the second stage treatment) or the whole regimes (e.g. comparing effects of two dynamic treatment regimes). A good primary research question should be both scientifically important and helpful in developing a dynamic treatment regime. For example in the addiction management study an interesting primary research question would be "on average what is the best subsequent treatment for responders to initial treatment". That is, we want to compare the mean primary outcomes of two groups of responders

(12-step versus 12-step+TDM). Standard test statistics (Hoel 1984) and sample size formula (Jennison and Turnbull 2000) for a large sample comparison of two means can be used in this case. Define the standardized effect size δ as the standardized difference in mean primary outcomes between two groups (Cohen 1988), i.e.

$$\delta = \frac{E(Y|Response, A_2 = \text{12-step}) - E(Y|Response, A_2 = \text{12-step+TDM})}{\sqrt{[Var(Y|Response, A_2 = \text{12-step}) + Var(Y|Response, A_2 = \text{12-step+TDM})]/2}}.$$

Let γ denote the overall initial response rate. Suppose the randomization probability is 1/2 for each treatment option at the secondary decision point. Standard calculation yields a sample size formula for the two sided test with power $1 - \beta$ and size α:

$$n = 4(z_{\alpha/2} + z_\beta)^2 \delta^{-2} \gamma^{-1},$$

where $z_{\alpha/2}$ and z_β are the standard normal $(1 - \alpha/2)$ percentile and $(1 - \beta)$ percentile, respectively. To use the formula, one needs to postulate the overall initial response rate γ.

Alternatively researchers may be more interested in primary research questions that are related to dynamic treatment regimes. In this case, Murphy (2005a) advocated that the primary research questions should involve the comparison of two dynamic treatment regimes beginning with different initial stage treatment options. This would allow researchers to decide which of the possible initial stage treatment options are worthy of further investigation. In the addiction management study, we may want to compare regime A with regime B defined in section 1. Test statistics and sample size formulae for this type of research question have been provided in Murphy (2005a) and Oetting et al. (2007). In the following, we review the formulae in the context of our example.

Let $p_1(a_1|S_1)$ and $p_2(a_2|S_1, A_1, S_2)$ be the randomization probability at the initial and secondary decision point, respectively. For any dynamic treatment regime of interest (d_1, d_2), assume

$$P(p_1(d_1|S_1)p_2(d_2|S_1, A_1, S_2) > 0) = 1. \tag{4}$$

Assumption (4) implies that treatments specified by regime A and B at any decision point for any given past history (i.e. past information and treatments) have positive probabilities of being assigned.

Murphy et al. (2001) showed that an unbiased estimator of $\mu_{(d_1,d_2)}$ (the mean primary outcome for regime (d_1, d_2)) is

$$\hat{\mu}_{(d_1,d_2)} = \frac{\mathbb{P}_n\left[\frac{1_{A_1=d_1(S_1)}1_{A_2=d_2(S_1,A_1,S_2)}}{p_1(d_1|S_1)p_2(d_2|S_1,A_1,S_2)}Y\right]}{\mathbb{P}_n\left[\frac{1_{A_1=d_1(S_1)}1_{A_2=d_2(S_1,A_1,S_2)}}{p_1(d_1|S_1)p_2(d_2|S_1,A_1,S_2)}\right]}, \tag{5}$$

where n is the sample size, $\mathbb{P}_n f = \sum_{i=1}^{n} f(X_i)/n$ in which X_i is a vector of observations for the i^{th} subject and f is a given function, and 1_Ω is an indicator function which equals 1 if event Ω occurs and 0 otherwise. A consistent estimator of the variance of $\sqrt{n}\hat{\mu}_{(d_1,d_2)}$ is

$$\hat{\tau}^2_{(d_1,d_2)} = \mathbb{P}_n\Big(\Big[\frac{1_{A_1=d_1(S_1)}1_{A_2=d_2(S_1,A_1,S_2)}}{p_1(d_1|S_1)p_2(d_2|S_1,A_1,S_2)}(Y - \hat{\mu}_{(d_1,d_2)})\Big]^2\Big).$$

The comparison of regime A with regime B can be obtained by comparing the subgroup of subjects in the trial whose treatment assignments are consistent with regime A with the subgroup of subjects in the trial whose treatment assignments are consistent with regime B. Note that there is no overlap between these two subgroups since a subject's initial treatment assignment can be consistent with only one of the regimes (A or B). The test statistic

$$Z = \frac{\sqrt{n}(\hat{\mu}_A - \hat{\mu}_B)}{\sqrt{\hat{\tau}^2_A + \hat{\tau}^2_B}} \tag{6}$$

has an asymptotic standard normal distribution under the null hypothesis $\mu_A = \mu_B$ (Murphy 2005a). The standardized effect size for addressing this question is defined as $\delta = (\mu_A - \mu_B)/\sqrt{(\sigma^2_A + \sigma^2_B)/2}$, where σ^2_A and σ^2_B are the variances of the primary outcomes for regime A and B, respectively. Suppose the randomization probability for each treatment option is 1/2 at each decision point. Equation (10) in Murphy (2005a) implies that $Var(\sqrt{n}\hat{\mu}_A) = 4\sigma^2_A$ and $Var(\sqrt{n}\hat{\mu}_B) = 4\sigma^2_B$ in large samples. Using a large sample approximation, the required sample size for the two sided test $\big(H_0 : \mu_A = \mu_B \quad \text{v.s.} \quad H_1 : \mu_A - \mu_B = \delta\sqrt{(\sigma^2_A + \sigma^2_B)/2}\big)$ with power $1 - \beta$ and size α is

$$n = 8(z_{\alpha/2} + z_\beta)^2\delta^{-2}.$$

A detailed derivation of the sample size formula in a similar context can be found in Oetting et al. (2007). Oetting et al. (2007) also discussed additional research questions and the corresponding test statistics and sample size formulae under different working assumptions. A web application that calculates the required sample size for sizing a study designed to discover the best dynamic treatment regime using a SMART design for continuous outcomes can be found at http://methodology.psu.edu/index.php/ra/adap-treat-strat/smart.

Formulae for the randomization probabilities that create equal sample sizes across all dynamic treatment regimes have been provided in Murphy (2005a). This was motivated by the classical large sample comparison of means for which, given equal variances, the power of a test is maximized by equal sample sizes. Let $k_1(S_1)$ be the number of treatment options at the initial decision point with pretreatment information S_1 and $k_2(S_1, A_1, S_2)$ be the number of treatment options at the secondary decision point with past history (S_1, A_1, S_2), respectively. Murphy's formulae yield

$$p_2(a_2|S_1, A_1, S_2) = k_2(S_1, A_1, S_2)^{-1}$$
$$p_1(a_1|S_1) = \frac{E[k_2(S_1, A_1, S_2)^{-1}|S_1, A_1 = a_1]^{-1}}{\sum_{b=1}^{k_1(S_1)} E[k_2(S_1, A_1, S_2)^{-1}|S_1, A_1 = b]^{-1}}. \quad (7)$$

If k_2 does not depend on S_2, the above formulae can be directly used. In our example, there are two initial treatment options and two secondary treatment options for each subject, i.e. $k_1(S_1) = 2$ and $k_2(S_1, A_1, S_2) = 2$ for all possible combinations of (S_1, A_1, S_2). Thus (7) yields a randomization probability of 1/2 for each treatment option at each decision point. In general working assumptions concerning the distribution of S_2 given (S_1, A_1) are needed in order to use the formulae. See Murphy (2005a) for more details.

Some principles and practical considerations are as follows (for more details see Lavori and Dawson (2004), Murphy (2005a) and Murphy et al. (2007)). First, Murphy (2005a) proposed that the primary research question should consider at most simple dynamic treatment regimes so as to simplify the sample size formulae. In our addiction management study, we consider regimes where the initial decision rule is a constant (i.e. does not depend on an individual's pre-treatment information) and the secondary decision rule depends only on the individual's initial treatment and his/her response status. Second, both Lavori and Dawson (2004) and Murphy (2005a) pointed out that, when designing

the trial, the class of treatment options at each decision point should be restricted (only) by ethical, scientific or feasibility considerations. Lavori and Dawson (2004) demonstrated how to constrain treatment options and thus decision rules using the STAR*D example (Lavori et al. 2001; Fava et al. 2003). Yet, Murphy (2005a) warns against undue restriction of the class of the decision rules. Our addiction management example reflects this notion. Although we might have reason to believe that non-adherent non-responders to NTX should receive different treatment from adherent non-responders to NTX, we do not provide these two groups different treatment options if we are uncertain that such restriction is necessary. Finally, the SMART trial should be viewed as one trial among a series of randomized trials intended to develop and/or refine a dynamic treatment regime. It should eventually be followed by a randomized control trial that compares the developed regime and an appropriate control (Murphy 2005a; Murphy et al. 2007). Note that like traditional randomized trials, SMART trials may involve standard problems such as dropout, incomplete assessments, etc.

3.3 SMART design versus adaptive experimental designs

The SMART design introduced in the previous section involves stages of treatment. Some adaptive experimental designs also utilize stages of experimentation (Berry 2002, 2004). However, the SMART design is quite different from adaptive experimental designs.

An adaptive experimental design is "a multistage study design that uses accumulating data to decide how to modify aspects of the study without undermining the validity and integrity of the trial" (Dragalin 2006). Chow and Chang (2008) summarized various types of adaptive experimental designs. For example, a response adaptive design modifies the randomization schedules based on prior subjects' observed data at interim in order to increase the probability of success for future subjects (see Berry et al. 2001 for an example). A group sequential design allows premature stopping of a trial due to safety, futility and/or efficacy with options of additional adaptations based on results of interim analysis (see Pampallona and Tsiatis 1994 for an example). A sample size re-estimation design involves the recalculation of sample size based on study parameters (e.g. revised effect size, conditional power, nuisance parameters) obtained from interim data (see Banerjee and Tsiatis 2006 for an example). In general, the aim of adaptive experimental designs is to improve the quality, speed and efficiency of

clinical development by modifying one or more aspects of a trial.

With the above definition, the difference between standard SMART design and adaptive experimental designs is rather straight forward. In a SMART design, each subject moves through multiple stages of treatment. On the other hand in most adaptive experimental designs each stage involves different subjects. That is, each subject only participates in one stage of treatment. In both cases randomization occurs at each stage. The goal of a SMART trial is to develop a dynamic treatment regime that could benefit future patients. Many adaptive experimental designs (e.g. response adaptive randomization) try to provide most efficacious treatment to each subject in the trial based on the knowledge at the time that subject is randomized. In a SMART trial the design elements such as the final sample size, randomization probabilities and treatment options are specified prior to conducting the trial. On the other hand, in an adaptive experimental design the final sample size, randomization probabilities and treatment options may be altered during the conduct of the trial.

There are some studies in which the adaptive experimental design was combined with the SMART design. For example, Thall et al. (2002) provided a statistical framework for an "outcome adaptive design" in a multi-course treatments setting in which two SMART trials are involved. Each trial used one half of the subjects. If the data from the first trial show a particular treatment sequence to be inferior to the others within a subgroup of subjects, then that treatment sequence option is dropped within that subgroup in the second trial. At the end the best treatment sequence for each subgroup is selected. Thall and Wathen (2005) considered a similar but more flexible design where the randomization criteria for each subject at each stage depends on the data from all subjects previously enrolled. Thall and his colleagues were able to apply such a strategy because subject outcomes in each SMART trial are observed quickly. In many other settings, obtaining subject's outcomes may take a long time (e.g. 12 months in our addiction management study). Thus, adaptation based on interim data is less feasible. How to optimally combine adaptive experimental design with the SMART design is worthy of further investigation.

4 Optimal dynamic treatment regimes

In the previous section we discussed issues concerning experimental data and some primary analyses for developing dynamic treatment regimes. In the current section we discuss useful methods for estimating the optimal dynamic treatment regime using experimental data.

4.1 Simple dynamic treatment regimes

In some cases researchers would like to consider relatively simple dynamic treatment regimes. For example in the addiction management study, the primary research question may be the comparison of two simple dynamic treatment regimes (regime A and B). Each regime specifies one initial treatment for all patients and assigns one treatment for initial treatment responders and another for non-responders. There are totally eight such simple dynamic treatment regimes (see Figure 1).

When there are only a few dynamic treatment regimes, we can estimate the mean primary outcome for each regime using the estimator in (5) and select the best one. Consider regime A in the addiction management example. When the randomization probability for each treatment option at each decision point is 1/2, assumption (4) is satisfied. $\hat{\mu}_A$ is simply the average of Y over subjects whose treatment assignments are consistent with regime A. The estimation of μ_A can be improved by considering doubly robust estimator (Robins 2000), which may result in smaller variance (see Murphy et al. 2001).

4.2 Dynamic treatment regimes involving covariates

In general, there may be a large number of possible dynamic treatment regimes. The initial treatment decision may depend on pretreatment information S_1. The secondary treatment may vary according to an individual's pretreatment information S_1, initial treatment A_1 and ongoing information S_2. In fact, there could be infinite number of dynamic treatment regimes. For example, suppose there are two treatment options, 1 and -1, at each decision point and the data are collected from a SMART trial in which the randomization probability is 1/2 for each treatment option at each decision point. Often one uses summaries of past history (i.e. past information and treatment assignments) to form decision rules. Let H_{11} be a vector summary of S_1 and H_{21} be a vector summary of (S_1, A_1, S_2), respectively. We may be interested in finding the best dynamic treatment regime of the form $\{d_1(H_{11}) = sign(\psi_{10} +$

$H_{11}^T\psi_{11})$ and $d_2(H_{21}) = sign(\psi_{20} + H_{21}^T\psi_{21})\big\}$, where ψs are the parameters, $sign(x) = 1$ if $x > 0$ and -1 otherwise. ψ_{10} is the main effect of the initial treatment. Each component in ψ_{11} is the interaction effect of the initial treatment and the corresponding component in H_{11}. In our addiction management example, H_{11} could be a two-dimensional vector including addiction severity and an indicator of presence/absence of a co-occurring disorder, and H_{21} could be a four-dimensional vector including the initial treatment, response status, adherence to the initial treatment and a measure of side effects. In this case, (5) is equivalent to

$$\hat{\mu}_{(d_1,d_2)} = \frac{\mathbb{P}_n\big[1_{A_1(\psi_{10}+H_{11}^T\psi_{11})>0}1_{A_2(\psi_{20}+H_{21}^T\psi_{21})>0}Y\big]}{\mathbb{P}_n\big[1_{A_1(\psi_{10}+H_{11}^T\psi_{11})>0}1_{A_2(\psi_{20}+H_{21}^T\psi_{21})>0}\big]}. \tag{8}$$

Selecting the dynamic treatment regime (i.e. the ψs) that maximizes (8) is computationally intractable since the objective function (8) is nonconcave in the ψs. An additional problem is that, if H_{11} and H_{21} are of high dimension, the regime that maximizes (8) is subject to overfitting the data and may yield a poor mean primary outcome among all dynamic treatment regimes under consideration. This is because we try to maximize $\hat{\mu}_{(d_1,d_2)}$ instead of $\mu_{(d_1,d_2)}$ over (d_1, d_2). Denote the maximizer of $\hat{\mu}_{(d_1,d_2)}$ by $(\tilde{d}_1, \tilde{d}_2)$. When the sample size is relatively small as compared to the complexity of the dynamic treatment regimes, the resulting regime $(\tilde{d}_1, \tilde{d}_2)$ may fit the data well but the regime may not be close to the regime that maximizes $\mu_{(d_1,d_2)}$.

An alternative approach to estimate the optimal dynamic treatment regime is Q-learning, which can be viewed as a generalization of regression to multi-stage decision making. This method estimates the optimal decision rules by learning the Q-functions defined in section 3.2. There are many variants of Q-learning (Watkins 1989; Sutton and Barto 1998; Ormoneit and Sen 2002; Lagoudakis and Parr 2003; Ernst et al. 2005). Below we review Q-learning with function approximation as described in Murphy (2005b).

Let $\mathcal{Q}_1$ be the approximation space for the initial stage Q-function Q_1^* defined in (3) and $\mathcal{Q}_2$ be the approximation space for the second stage Q-function Q_2^* defined in (2), respectively. For example, assume there are two treatment options at each decision point; $a_1 \in \{-1, 1\}$ and $a_2 \in \{-1, 1\}$. We

may consider linear approximation spaces for the Q-functions,

$$\begin{aligned} \mathcal{Q}_2 =& \Big\{ Q_2(h_{20}, h_{21}, a_2; \boldsymbol{\theta}_2) = \phi_{20} + h_{20}^T \phi_{21} + (\psi_{20} + h_{21}^T \psi_{21}) a_2 : \boldsymbol{\theta}_2 = (\phi_{20}, \phi_{21}^T, \psi_{20}, \psi_{21}^T) \in \Theta_2 \Big\} \\ \text{and} \quad \mathcal{Q}_1 =& \Big\{ Q_1(h_{10}, h_{11}, a_1; \boldsymbol{\theta}_1) = \phi_{10} + h_{10}^T \phi_{11} + (\psi_{10} + h_{11}^T \psi_{11}) a_1, \boldsymbol{\theta}_1 = (\phi_{10}, \phi_{11}^T, \psi_{10}, \psi_{11}^T) \in \Theta_1 \Big\} \end{aligned} \tag{9}$$

where h_{20} and h_{21} are vector summaries of (s_1, a_1, s_2), h_{10} and h_{11} are vector summaries of s_1, and Θ_1 and Θ_2 are the parameter spaces. Note that we use upper case letters to denote random variables or data for subjects in the SMART trial and lower case letters to denote the values of the variables.

Since Q_2^* is the conditional mean of Y given past information and treatment assignments (S_1, A_1, S_2, A_2), we can estimate the second stage parameter $\boldsymbol{\theta}_2$ using least squares,

$$\hat{\boldsymbol{\theta}}_2 = \arg \min_{\boldsymbol{\theta}_2 \in \Theta_2} \mathbb{P}_n [Y - Q_2(H_{20}, H_{21}, A_2; \boldsymbol{\theta}_2)]^2$$

Similarly, since Q_1^* is the conditional mean of $\max_{a_2} Q_2^*$ given (S_1, A_1), a least squares estimator of the initial stage parameter $\boldsymbol{\theta}_1$ (with Q_2^* estimated by $Q_2(H_{20}, H_{21}; \hat{\boldsymbol{\theta}}_2)$) is

$$\hat{\boldsymbol{\theta}}_1 = \arg \min_{\boldsymbol{\theta}_1 \in \Theta_1} \mathbb{P}_n [\max_{a_2} Q_2(H_{20}, H_{21}, a_2; \hat{\boldsymbol{\theta}}_2) - Q_1(H_{10}, H_{11}; \boldsymbol{\theta}_1)]^2 \tag{10}$$

(see Tsitsiklis and Van Roy 1996 for this estimation method in a similar context). The estimated optimal dynamic treatment regime then uses treatments that maximize the estimated Q-functions. In the linear model example, the estimated optimal regime is

$$\begin{aligned} & \hat{d}_2(h_{21}) = sign(\hat{\psi}_{20} + h_{21}^T \hat{\psi}_{21}) \\ \text{and} \qquad & \hat{d}_1(h_{11}) = sign(\hat{\psi}_{10} + h_{11}^T \hat{\psi}_{11}). \end{aligned} \tag{11}$$

Note that sometimes researchers may consider different treatment options for different subgroups of subjects. For example in the addiction management study, different secondary treatment options are offered each of the three subgroups: initial treatment responders (subgroup 1), NTX non-responders (subgroup 2) and CBI non-responders (subgroup 3). In this case, we code $A_2 = 1$ for one treatment

option and $A_2 = -1$ for the other treatment option in each subgroup. We can use different linear models, say $Q_2^{(j)}, j =$ subgroup $1, \ldots, 3$, for the three subgroups of subjects, respectively. The final model for Q_2^* can be written as $\sum_{j=1}^{3} Q_2^{(j)} 1_{(j)}$, where $1_{(j)}$ is 1 if the subject belongs to the j^{th} subgroup and 0 otherwise. Since each $Q_2^{(j)}$ is a linear model, the model for Q_2^* fits the linear framework (9).

In Q-learning we modeled the Q-functions Q_1^* and Q_2^*. However, only part of the Q-function is relevant for the construction of the decision rules. This can be easily seen from the above linear model example where the estimated decision rule at each decision point only depends on the interaction part in the corresponding linear model (e.g. the decision rules (11) only depend on ψs not ϕs). In general, each Q_t^* $(t = 1, 2)$ can be written as $Q_t = g_t^* + \max_{a_t} Q_t^*$ $(g_t^* = Q_t^* - \max_{a_t} Q_t^*)$. g_t^* is called the advantage function (Baird 1994) at the t^{th} decision point. It measures the gain in performance obtained by following treatment a_t as compared to following the best treatment at the t^{th} decision point. Since $\max_{a_t} Q_t^*$ does not contain a_t, we only need to model g_t^* instead of modeling Q_t^*. g_t^* may include many fewer variables than the corresponding Q_t^* since it contains only variables in the interaction terms in Q_t^*. Estimation of only the advantage functions was first proposed in Murphy (2003), along with a least squares estimation method. Robins (2004) provided a refined estimating equation to gain efficiency. Chakraborty and Murphy (2009) showed that under appropriate conditions Q-learning with linear models is algebraically equivalent to an inefficient version of Robins' method.

5 Other analyses

5.1 Inference

It is crucial to attach measures of confidence (e.g. standard errors, confidence intervals, etc.) to the estimated dynamic treatment regimes. Furthermore since collecting patient information in order to apply the decision rules may be relatively expensive in clinical practice, researchers may be interested in assessing whether certain patient variables are necessary for making the decision. For example in the addiction management example, suppose we are to make the initial decision based on a linear decision rule "treat patients with NTX if $\psi_{10} + \psi_{11} \cdot$ "addiction severity" > 0, and treat patients with CBI otherwise". It may be possible to simplify the decision rule by removing the variable "addiction severity" if the data do not provide sufficient evidence that this variable is necessary. In this case it

would be useful to assess the extent to which the variable "addiction severity" is important in the data set (say via inference on the associated parameter ψ_{11}). To further justify whether the two initial treatments have different effects on the primary outcome, we may also want to test if $\psi_{10} = 0$.

Inferential methods have been discussed in Robins (2004), Moodie and Richardson (2007) and Chakraborty and Murphy (2009). Robins (2004) noted that the treatment effect parameters at any stage prior to the last can be non-regular. As a consequence, it is difficult to provide valid confidence intervals for the optimal dynamic treatment regime estimated from Q-learning or Robins' method. To explain meaning of the non-regularity in this context, we discuss the inference for ψ_{11}. Consider the linear model as described in (9) with H_{11} being the variable "addiction severity". Then the Q-learning estimator of ψ_{11} is given by solving (10). Note that $\hat{\psi}_{11}$ is a function of $\max_{a_2} Q_2(H_{20}, H_{21}, a_1; \hat{\boldsymbol{\theta}}_2)$. With the linear parameterization, $\max_{a_2} Q_2(H_{20}, H_{21}, a_1; \hat{\boldsymbol{\theta}}_2)$ equals $\hat{\phi}_{20} + H_{20}^T\hat{\phi}_{21} + |\hat{\psi}_{20} + H_{21}^T\hat{\psi}_{21}|$, which is non-differentiable at the point $\hat{\psi}_{20} + H_{21}^T\hat{\psi}_{21} = 0$. Due to the non-differentiability, it can be shown (see Robins 2004 and Moodie and Richardson 2007) that the asymptotic distribution of $\sqrt{n}(\hat{\psi}_{11} - \psi_{11})$ is normal if $P(\psi_{20} + H_{21}^T\psi_{21} = 0) = 0$ and non-normal otherwise ($\psi_{20} + H_{21}^T\psi_{21} = 0$ implies that there is no second stage treatment effect for patients with past history H_{21}). Thus ψ_{11} is a "non-regular" parameter and $\hat{\psi}_{11}$ is a "non-regular" estimator of ψ_{11} (see Bickel et al. (1993) for a more precise definition of non-regularity). The practical consequence of this is that whenever $\psi_{20} + H_{21}^T\psi_{21}$ is close to zero, both confidence intervals of ψ_{11} based on formulae derived from Taylor series arguments and confidence intervals based on bootstrap will perform poorly (Moodie and Richardson 2007; Chakraborty and Murphy 2009).

Several approaches have been discussed to deal with non-regularity. Robins (2004) constructed a score method that provides a conservative uniform asymptotic confidence interval for ψ_{11}. Other methods reduced bias in the estimation of ψ_{11} due to non-regularity by substituting $|\hat{\psi}_{20} + H_{21}^T\hat{\psi}_{21}|$ in (10) with other quantities. Moodie and Richardson (2007) proposed a hard-threshold estimator in which $|\hat{\psi}_{20} + H_{21}^T\hat{\psi}_{21}|$ is replaced by 0 if it is below a threshold. This method reduced the bias of $\hat{\psi}_{11}$ when $|\hat{\psi}_{20} + H_{21}^T\hat{\psi}_{21}|$ is close to zero. However, it is unclear how to select the threshold value. Chakraborty and Murphy (2009) used a soft-threshold estimator, where $|\hat{\psi}_{20} + H_{21}^T\hat{\psi}_{21}|$ in (10) is replaced by $|\hat{\psi}_{20} + H_{21}^T\hat{\psi}_{21}|\big(1 - \frac{\lambda}{|\hat{\psi}_{20}+H_{21}^T\hat{\psi}_{21}|}\big)^+$ with tuning parameter λ chosen by an empirical Bayes

method. Simulations in Chakraborty and Murphy (2009) provided evidence that, in the non-regular setting in which $P(\psi_{20}+H_{21}^T\psi_{21}=0)>0$, the use of bootstrap confidence intervals along with the soft-threshold estimator (and hard-threshold estimator in some cases) reduced bias due to non-regularity and gave correct coverage rate. When $P(\psi_{20}+H_{21}^T\psi_{21}=0)=0$, bootstrap confidence intervals based on the original Q-learning estimator performed the best, but the percentile bootstrap with the soft-threshold estimator also performed reasonably well. Theoretical optimality of this method is unclear and worth further investigation. Instead of providing confidence sets for ψ_{10} and ψ_{11}, Lizotte et al. (2009) proposed a "voting method", the vote for a treatment is an estimation of the probability that the treatment would be selected as the best treatment if the trial were conducted again. The non-regularity problem in this method is addressed using a hard-threshold estimator. However, this approach is relatively new and untested, hence further investigation and refinements are needed.

5.2 Modeling

The inference problem discussed in the previous section is based on the parametric or semi-parametric modeling of the Q-functions. Note that the approximation for the Q-functions together with the definition of the estimated decision rules as the argmax of the estimated Q-functions places implicit restrictions on the set of regimes that can be considered. More specifically with a given approximation space for the Q-functions the set of regimes under consideration is $\mathcal{D}=\{(d_1,d_2): d_t\in \arg\max_{a_t} Q_t, Q_t\in \mathcal{Q}_t, t=1,2.\}$. Thus at least implicitly the goal becomes the estimation of the best regime in the space $\mathcal{D}$. However, problems occur if the approximation space for Q_1^* and Q_2^* does not contain the true Q-functions. In particular, when the approximation is poor, the mean primary outcome of the estimated regime, $\mu_{(\hat{d}_1,\hat{d}_2)}$, may not be close to $\max_{(d_1,d_2)\in\mathcal{D}}\mu_{(d_1,d_2)}$ even in large samples (Tsitsiklis and Van Roy 1997). That is, $\mu_{(\hat{d}_1,\hat{d}_2)}$ may not be a consistent estimator of $\max_{(d_1,d_2)\in\mathcal{D}}\mu_{(d_1,d_2)}$ when the approximation space does not contain the true Q-functions. The potential bias (i.e. inconsistency) will be eliminated if the approximation space provides a sufficiently good approximation for Q_1^* and Q_2^*, but then the estimated Q-functions will have large variances due to the small sample size. Consequently, $\mu_{(\hat{d}_1,\hat{d}_2)}$ will have a large variance as well. Thus selecting an appropriate approximation space is the key to success. Ormoneit and Sen (2002) used a sequence of kernel based approximation spaces

and made assumptions on the target function to guarantee a sufficiently rich approximation. Another promising avenue is to use model selection techniques in regression/classification with Q-learning to get a good trade off between bias and variance.

6 Discussion and open questions

Dynamic treatment regimes are a new approach to treatment design. These treatment designs adapt the treatment to patient characteristics and outcomes.

In section 3 we reviewed the SMART design. To conduct a SMART trial where there is a plethora of clinical decisions and treatment options, we need to pre-specify study components, such as how many critical decisions are of interest, what are the best time-points to make decisions (which can also be viewed as part of the critical decisions), and what treatment options should be investigated at each decision point, etc. A new screening experimental design proposed in Murphy and Bingham (2008) can be used to identify promising components and screening out negligible ones. As discussed in Murphy et al. (2007), another approach is to use the MOST paradigm developed in Collins et al. (2005). This paradigm advocates the use of a series of experimental trials to prospectively determine active components for future investigation. How to effectively integrate the SMART design into the MOST paradigm is an area for future research.

In section 4, we reviewed several methods for estimating the optimal dynamic treatment regimes. In fact, many other methods were omitted. For example, Robins (1986) proposed a G-computation formula to estimate the mean primary outcome of a dynamic treatment regime based on modeling the conditional distributions of the data. Lunceford et al. (2002) and Wahed and Tsiatis (2004, 2006) provided semiparametric estimators of the survival function for a given regime in cancer trials. Those estimates can be used to select the best regime among a small set of dynamic treatment regimes. Thall and colleagues (Thall et al. 2000, 2002, 2007) developed several likelihood based methods, both Bayesian and frequentist, for selecting the best dynamic treatment regime. Instead of estimating the optimal dynamic treatment regimes, van der Laan and his colleagues (van der Laan et al. 2005; Petersen et al. 2007; van der Laan and Petersen 2007) aimed at ascertaining a "statically" optimal dynamic treatment regime. At each stage the "statically" optimal dynamic treatment regime only

uses current available information to make all future decisions. The "statically" optimal dynamic treatment regime is not truly optimal in a multi-stage setting (see Petersen et al. (2007) for a detailed comparison of the optimal dynamic treatment regime and the "statically" optimal dynamic treatment regime).

In section 5, we discussed several open questions concerning data analyses including inference and modeling. Still, there many other challenges and open questions in this regard. For example, instead of inference, variable selection techniques can be used to assess whether particular patient variables are necessary for decision making. In fact variable selection is very important here since data collected in clinical trials are often of high dimension while only a few variables are likely to be useful in selecting the best sequence of treatments. Although variable selection techniques developed for prediction can be used here, such methods may miss variables that are useful for decision making. Gunter et al. (2007) developed a variable selection method for decision making in the case of single decision point and empirically showed that their method is better than Lasso (Tibshirani 1996) in decision making. However, theoretical properties of their method have not been developed and the extension to multi-stage decision making is needed. Another open question is how to construct a dynamic treatment regime when there are multiple primary outcomes (e.g. functionality, side effects, cost, etc). Thall et al. (2008) proposed a Bayesian procedure for finding the best dosage in a single stage setting involving bivariate (efficacy, toxicity) outcomes. In general, to develop a good dynamic treatment regime and conduct other secondary analyses, the construction of a high quality composite outcome is very important. Other challenges include feature construction, dealing with missing data, etc. All these issues are worthy of future research.

The current chapter illustrates the benefits of dynamic treatment regimes and the SMART experimental approach for constructing and evaluating dynamic treatment regimes. We also present methods for estimating the optimal dynamic treatment regime, and discuss challenges associated with these methods as well as their potentials for intervention scientists aiming to develop dynamic treatment regimes. Overall, this topic deserves continued research attention and a prominent role in intervention research.

Acknowledgements

We acknowledge support for this work from NIH grants RO1 MH080015 and P50 DA10075.

References

Alcoholics Anonymous 2001. Chapter 5: How It Works. *Alcoholics A nonymous (4th edition ed.).* Alcoholics Anonymous World Services.

Baird, L. C. 1994. Reinforcement learning in continuous time: advantage updating. *IEEE International Conference on Neural Networks,* 4: 2448-2453.

Banerjee, A. and Tsiatis, A. A. 2006. Adaptive two-stage designs in phase II clinical trials. *Statistics in Medicine,* 25(19):3382-3395.

Bellman, R. E. 1957. *Dynamic Programming*. Princeton University Press.

Berry, D. A., Mueller, P., Grieve, A. P., Smith, M., Parke, T., Blazek, R., Mitchard, N. and Krams, M. 2001. Adaptive Bayesian designs for dose-ranging drug trials. In:*Gatsonis C, Carlin B, Carriquiry A, editors. Case Studies in Bayesian Statistics,* Volume V. 99-181, New York: Springer-Verlag.

Berry D. A. 2002. Adaptive clinical trials and Bayesian statistics (with discussion). *Pharmaceutical Report, by the American Statistical Association.*

Berry, D. A. 2004. Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science,* 19:175-187.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. 1993. *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press.

Chakraborty, B. and Murphy, S. A. 2009. Inference for nonregular parameters in optimal dynamic treatment regimes. *To appear in the special issue on "Clinical Trials in Mental Health" of the journal "Statistical Methods in Medical Research".*

Chow, S. C. and Chang, M. 2008. Adaptive design methods in clinical trials - a review. *Orphanet Journal of Rare Diseases*, 3:11.

Cohen J. 1988. Statistical Power Analysisfor the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Collins, L. M., Murphy, S. A., Nair, V. and Strecher, V. 2005. A strategy for optimizing and evaluating behavioral intervention. *Annals of Behavioral Medicine,* 30:65-73.

COMBINE Study Research Group 2003. Testing combined pharmacotherapies and behavioral interventions in alcohol dependence: rationale and methods. *Alcoholism Clinical and Experimental Research,* 27:1107-1122.

Dragalin, V. 2006. Adaptive designs: terminology and classification. *Drug Information Journal*, 40:425-435.

Ernst, D., Geurts, P. and Wehenkel, L. 2005. Tree-based batch mode reinforcement learning, *Journal of Machine Learning Research,* 6:503-556.

Fava, M., Rush, A. J., Trivedi, M. H., Nierenberg, A. A., Thase, M. E., Sackeim, H. A., Quitkin, F. M., Wisniewski, S., Lavori, P. W., Rosenbaum, J. F. and Kupfer, D. J. 2003. Background and rationale for the sequenced treatment alternatives to relieve depression (STAR*D) study. *Psychiatr Clin North Am.* 26(2):457-494.

Gunter, L. L., Zhu, J. and Murphy, S. A. 2007. Variable selection for optimal decision making. *Poceedings of the 11th Conference on Artificial Intelligence in Medicine, AIME 2007, LNCS/LNAI 4594,* 149-154.

Hoel, P. 1984. *Introduction to Mathematical Statistics.* 5th ed. New York: John Wiley and Sons.

Jennison, C. and Turnbull, B. 2000. *Group Sequential Methods with Applications to Clinical Trials.* Boca Raton, FL city: Chapman & Hall.

Lagoudakis, M. G. and Parr, R. 2003. Least-squares policy iteration, *Journal of Machine Learning Research,* 4:1107-1149.

Lavori, P. W. and Dawson, R. 2000. A design for testing clinical strategies: biased individually tailored within-subject randomization. *Journal of the Royal Statistical Society A*, 163:29-38.

Lavori, P. W., Rush, A. J., Wisniewski, S. R., Alpert, J., Fava, M., Kupfer, D. J., Nierenberg, A., Quitkin, f. M., Sackeim, H. A., Thase, M. E. and Trivedi, M. 2001. Strengthening clinical effectiveness trials:equipoise-stratified randomization. *Biological Psychiatry*, 48:605-614.

Lavori, P. W. and Dawson, R. 2004. Dynamic treatment regimes: practical design considerations. *Clinical Trials*, 1:9-20.

Lizotte, D. J., Laber, E. and Murphy, S. A. 2009. Assessing confidence in policies learned from sequential randomized trials. *Submitted.*

Lunceford, J. K., Davidian, M. and Tsiatis, A. A. 2002. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics,* 58: 48-57.

McLellan, A. T. 2002. Have we evaluated addiction treatment correctly? Implications from a chronic care perspective. *Addiction*, 97:249-252.

Miller, W. R. (ed.) 2004. COMBINE Monograph Series, Combined Behavioral Intervention Manual: A Clinical Research Guide for Therapists Treating People With Alcohol Abuse and Dependence. *DHHS Publication No. (NIH)* 04-5288, volume 1. NIAAA, Bethesda, MD.

Moodie, E. E. M. and Richardson, T. S. 2007. Bias Correction in Non-Differentiable Estimating Equations for Optimal Dynamic Regimes. *COBRA Preprint Series. Article 17.*

Murphy, S. A., van der Laan, M. J., Robins, J. M. and CPPRG 2001. Marginal mean models for dynamic regimes. *Journal of American Statistical Association*, 96:1410-1423.

Murphy, S. A. 2003. Optimal Dynamic Treatment Regimes. *Journal of the Royal Statistical Society, Series B (with discussion)*, 65(2):331-366.

Murphy, S. A. 2005a. An Experimental Design for the Development of Adaptive Treatment Strategies. *Statistics in Medicine*, 24:1455-1481.

Murphy, S. A. 2005b. A Generalization Error for Q-Learning. *Journal of Machine Learning Research,* 6(Jul):1073-1097.

Murphy, S. A., Lynch, K. G., Oslin, D., Mckay, J. R. and TenHave, T. 2007. Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence*, 88s:s24-s30.

Murphy, S. A. and Bingham, D. 2008. Screening Experiments for Developing Dynamic Treatment Regimes. *To Appear in JASA.*

Neyman, J. 1923. On the application of probability theory to agricultural experiments. Translated in *Statistical Science*, 5:465-480 (1990).

Oetting, A. I., Levy, J. A., Weiss, R. D. and Murphy, S. A. 2007. Statistical methodology for a SMART Design in the development of adaptive treatment strategies. *To appear in Causality and Psychopathology: Finding the Determinants of Disorders and their Cures (P.E. Shrout, Ed.) Arlington VA: American Psychiatric Publishing, Inc.*

Ormoneit, D. and Sen, S. 2002. Kernel-based reinforcement learning. *Machine Learning,* 49(2-3):161-178.

Oslin, D. W., Sayers, S., Ross, J., Kane, V., TenHave, T., Conigliaro, J. and Cornelius, J. 2003. Disease management for depression and at-risk drinking via telephone in an older population for veterans. *Psychosomatic Medicine,* 65:931-937.

Pampallona, S and Tsiatis, A. A. 1994. Group sequential designs for one and two sided hypothesis test- ing with provision for early stopping in favour of the null hypothesis. *Journal of Statistical Planning and Inference,* 42:19-35.

Petersen, M. L., Deeks, S. G. and van der Laan, M. J. 2007. Individualized Treatment Rules: Generating Candidate Clinical Trials. *Statistics in Medicine,* 26(25): 4578-4601.

Robins, J. M. 1986. A new approach to causal inference in mortality studies with sustained exposure periods -application to control of the healthy worker survivor effect. *Computers and Mathematics with Applications,* 14:1393-1512.

Robins, J. M. 1987. Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods -application to control of the healthy worker survivor effect." *Computers and Mathematics with Applications,* 14:923-945.

Robins, J. M. and Wasserman, L. 1997. Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence,* eds. D. Geiger and P. Shenoy, San Francisco: Morgan Kaufmann.

Robins, J. M. 2000. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999,* 6-10.

Robins, J. M. 2004. Optimal structural nested models for optimal sequential decisions. *Proceedings of the Second Seattle Symposium on Biostatistics. In D.Y. Lin, P. Haegerty eds. Lecture notes in Stastitics.* New York: Springer.

Rubin, D. B. 1978. Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, 6:34-58.

Schneider, L. S., Tariot, P. N., Lyketsos, C. G., Dagerman, K. S., Davis, K. L., Davis, S., Hsiao, J. K., Jeste, D. V., Katz, I. R., Olin, J. T., Pollock, B. G., Rabins, P. V., Rosenheck, R. A., Small, G. W., Lebowitz, B. and Lieberman, J. A. 2001. National Institute of Mental Health clinical antipsychotic trials of intervention effectiveness (CATIE) alzheimer disease trial methodology *American Journal of Geriatric Psychiatry,* 9(4):346-360.

Stone, R. M., Berg, D. T., George, S. L., Dodge, R. K., Paciucci, P. A., Schulman, P., Lee, E. J., Moore, J. O., Powell, B. L. and Schiffer, C. A. 1995. Granulocyte Macrophage colony-stimulating factor after initial chemotherapy for elderly patients with primary acute myelogenous leukemia. *The New England Journal of Medicine,* 332: 1671-1677.

Sutton, R. S. and Barto, A. G. 1998. Reinforcement Learning: An Introduction. TheMIT Press, Cambridge, Mass.

TenHave, T. R., Coyne, J., Salzer, M. and Katz, I. 2003. Research to improve the quality of care for depression:alternatives to the simple randomized clinical trial. *General Hospital Psychiatry,* 25:115-123.

Thall, P. F., Millikan, R. E. and Sung, H. G. 2000. Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine,* 19:1011-1028.

Thall, P. F., Sung, H. G. and Estey, E. H. 2002. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the American Statistical Association,* 97:29-39.

Thall, P. F. and Wathen, J. K. 2005. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Statistics in Medicine,* 24:1947-1964.

Thall, P. F., Wooten, L. H., Logothetis, C. J., Millikan, R. and Tannir, N. M. 2007. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in Medicine,* 26:4687-4702.

Thall, P. F., Nguyen, H. and Estey, E.H. 2008. Patient-specific dose-finding based on bivariate outcomes and covariates. *Biometrics,* 64(4):1126-1136.

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B,* 32: 135-166.

Tsitsiklis, J. N. and Van Roy, B. 1996. Feature-based methods for large scale dynamic programming, *Machine Learning,* 22:59-94.

Tsitsiklis, J. N. and Van Roy, B. 1997. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control,* 42(5): 674-690.

van der Laan, M. J., Petersen, M. L. and Joffe, M. M. 2005. History-Adjusted Marginal Structural Models and Statically-Optimal Dynamic Treatment Regimens. *The International Journal of Biostatistics,* 1(1):Article 4.

van der Laan, M. J. and Petersen, M. L. 2007. Statistical Learning of Origin-Specific Statically Optimal Individualized Treatment Rules. *The International Journal of Biostatistics,* 3(1).

Wahed, A. S. and Tsiatis, A. A. 2004. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics,* 60:124-133.

Wahed, A. S. and Tsiatis, A. A. 2006. Semiparametric efficient estimation of survival distribution for treatment policies in two-stage randomization designs in clinical trials with censored data. *Biometrika,* 93: 163-177.

Watkins, C. J. C. H. 1989. Learning from Delayed Rewards. Ph.D. thesis, Cambridge University.

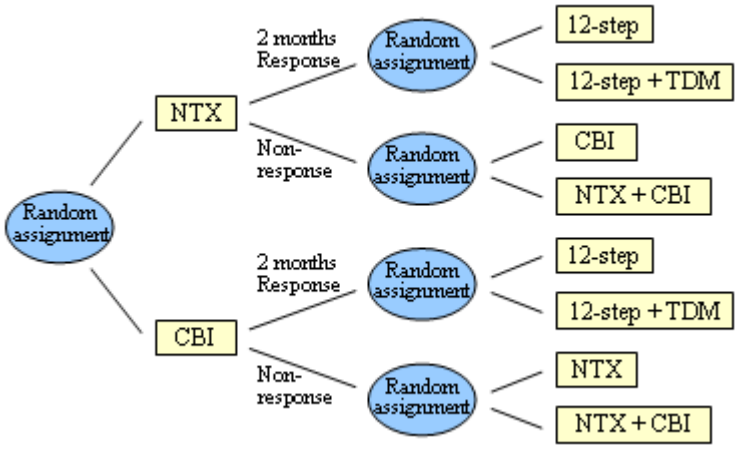

Figure 1: The SMART design for the addiction management study.