Chapter 23: Design and Analysis Methods for Developing Personalized Treatment Rules

Emily L. Butler and Michael R. Kosorok
University of North Carolina
Chapel Hill, NC

1 Introduction

Precision or personalized medicine is the practice of tailoring healthcare plans to patients as individualized as possible. In an ideal setting caregivers are able to account for patient history, genetic information, response to previous treatments, and other relevant factors when making treatment decisions throughout the course of the disease. These treatments would be assigned based on how each patient is progressing at each stage and which sequence of treatments is best earns the optimal expected long term response. These treatment plans have been coined dynamic treatment regimes (DTRs) and estimating DTRs is the main goal of precision medicine for patients with a chronic disease. These treatment plans are unique in the sense that they are not predetermined at the initial doctor's visit as a general treatment plan for all patients but are determined at each follow up visit, which makes the treatment plan dynamic through time. The evidence based treatment decisions are determined by the patient's response to the course of treatment and uses prognostic and treatment information to define their history [Collins et al., 2007]. DTRs have four components: decision points (time points at which the decisions are made), tailoring variables (patient's prognostic information used to make treatment decisions), intervention components (the type or dose/intensity/duration of the treatment), and decision rules (a function that links the tailoring variables to the intervention options at decision points) [Collins et al., 2014].

In the statistical framework, DTRs are developed to create mathematically dependent treatment plans that mimic how clinicians treat patients who need chronic medical treatment. Initial interest in this method of treating patients began in the mid-1980s when researchers were evaluating two stage dynamic treatment strategies for cancer patients; in the 1990s, the idea of a two stage customized treatment plan expanded when psychiatrists were interested in expanding to k-stage treatment plans; and by the early 2000s, clinical trials and treatment strategies were implemented by doctors studying DTRs in substance abuse and mental health research [Lavori and Dawson, 2014]. Once the investigator has chosen a selection of potential DTRs, they need to be evaluated in a special clinical trial called a Sequential Multiple Assignment Randomization Trials (SMART) [Murphy, 2005b]. The development of SMARTs began when traditional clinical trial designers were looking to identify an intermediate outcome between randomization and the primary outcome and make an appropriate reaction to this outcome. The treatment plan can be adjusted based on the patient's progress or just carry-over the patient's to a new set of treatments at the following stage.

It is important to highlight the distinction between DTRs and SMARTs. A DTR is a treatment strategy that tells the caregiver which adaptive, sequential treatment plan will most likely lead to the highest probability of success. A SMART design is an experimental trial with the purpose of determining which DTRs are optimal for which patients. A SMART is able to evaluate the various DTRs because it compares them by randomizing patients to a set preselected treatments at each phase. A SMART can have two goals: compare a small number of pre-specified DTRs embedded in the SMART design or construct new DTRs which may not be naturally embedded in the design.

The material presented in this literature review is meant to be a survey of current statistical methods used to estimate DTRS in the context of using data collected in a SMART trial. Kosorok and Moodie [2016] published a book titled Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine which covers this and similar topics in more detail. In this book, the authors discuss select topics such as SMART designs, observational studies, reinforcement learning and various methods of estimating DTRs. Please refer to this books for an in depth description of a number of useful methods for DTRs.

This review covers a brief introduction on how to design a SMART and what statistical methodologies exist to make inference from this data, with particular emphasis on recent developments in the area. Section 2 will highlight how a SMART design fits into the overall development of a treatment strategy, the importance of pilot studies, practical considerations such as sample size, calculations, and handling missing data, as well as the future of data collection. Section 3 introduces methodological results for the single stage paradigm and section 4 provides methodological results for the multiple stage paradigm. Section 5 highlights related topics, such as variable selection and developing DTRs for when there are multiple outcomes and observational data.

2 Study Design

A SMART is a clinical trial design that helps determine which sequence of treatments is best for which type of patients. They investigate the best sequencing of intervention components, what tailoring variables should be used, when and how frequently should these tailoring variables be assessed, and should one treatment be assigned or should patients have the ability to choose from a list of options. This design is novel in that the treatments are assessed as an entire treatment sequence and not isolated by each individual phase of the trial. This involves multiple randomizations over time where each randomization corresponds to a treatment decision. Additionally, this design can distinguish between responders and non-responders by assigning rescue or maintenance therapies [Collins et al., 2007, 2014]. Pragmatically, another benefit of this design is that SMARTs require a smaller sample size than their RCT counterparts because of their efficient use of subjects throughout stages. Allowing for a rescue therapy for patients experiencing extreme adverse events will deter dropout as well as better capture their response profile overtime.

Consider a generic example for illustration. A 2 stage SMART enrolls 200 participants with equal randomization. It assigns 100 patients to treatment A and 100 patients to treatment B at stage 1. At the end of stage 1 the patient's status is assessed and the they are classified as either responsive or non-responsive, where what defines responsive or non-responsive should be determined a priori by the investigator. Each of the patients is then re-randomized according to their response classification. For instance, patients who responded to treatment A are assigned to stay on treatment A at stage 2, while the patients who did not respond to treatment A are randomized to treatment C or D. A similar situation could arise for treatment B where responders stay on treatment B and non-responders are randomized to treatments E or F. In this scenario, there are 6 DTRs: {A, A}, {A, C}, {A, D} {B, B}, {B, E}, {B, F}. Alternatively, responsive patients could also be re-randomized. There could be more than 2 treatment options at each stage, or the non-responsive patients could switch from A to B or vice versa. The randomization also does not need to be balanced, and the stages could be extended beyond only 2. In this example data would be collected at 3 time points: baseline, time 1 (after the first stage of the study but before the patient is re-randomized to the second treatment), and time 2 (after the second stage or at the end of the study). Generally speaking, randomization does not need to depend on responder status, although this is the case in this example.

There are numerous reasons it is advantageous to use a SMART. For example, it has the desirable quality of making better use of the pre-determined sample size and can answer more clinical questions than a RCT. In a two phase SMART, similar to RCT, the first phase of the trial can be assessed by comparing the mean outcomes between the first two lines of treatment and the second phase can compare the effect of the treatment options for responders and non-responders, regardless of their first line of treatment. This increases the power of the hypothesis tests, since they recycle the patients in the second phase of the study. In addition to hypothesis testing as in RCT, SMARTs can also be used to generate hypotheses. Most importantly, the design then has the ability to compare the embedded treatment regimes that pool information across multiple experimental conditions by, again, recycling patients through the trial [Collins et al., 2014]. While the main goal of a SMART is to mine data used to develop DTRs, they have other uses, such as discovering which treatments work best sequentially to obtain an improved outcome, investigating the interplay between trajectories of the patients disease progression and treatment sequences, comparing different treatment sequences, and investigating the benefit of both prognostic information and clinical data in determining individualized treatments [Almirall et al., 2012].

As with most experimental designs, it is important to begin with an information gathering pilot study. A good pilot study is great practice for implementing a larger design, gives critical value estimates, is important for sample size estimation, and provides a preliminary look at the utility of the proposed treatments. The novelty of SMARTs raises feasibility concerns which makes a pilot study even more crucial for designing an effective and efficient SMART. Almirall et al. [2012] highlights the following important topics that researchers must be aware of when designing a SMART and how solutions can be elicited through a pilot study: (1) determining primary outcome. This will be used to assess response status (if applicable), when treatment decisions should be made, what criterion is used to make the decisions, frequency of assessment, how sensitive this measure is, justification of its use, and feasible application in clinical treatment; (2) deciding which tailoring variables should be collected, such as baseline characteristics or time varying measures; (3) deciding how to control for missing data. This should be guided by how it would be handled in clinical practice; (4) deciding between up-front randomization (randomization at the beginning of the trial) or real-time randomization (randomization at each decision point, which allows for clinical information to be used in randomization); (5) highlighting the difference between research assessments for data analysis to develop adaptive treatment strategies and assessments of the adaptive treatment strategies used to inform the sequential treatment assessment; (6) identifying concerns clinicians have regarding sequences of treatments offered and assessment of what determines response versus nonresponse; (7) assessing for patient acceptability; (8) testing the language of consent forms; (9) illuminating unanticipated tailoring variables that would be useful in the subsequent SMART.

The pilot study provides crucial information to increase the probability of success of a SMART. As previously stated, one goal of a SMART is to identify embedded DTRs not embedded in the original design. A great example of this is the analysis of SMART data collected from a study for the treatment of advanced prostate cancer. Wang et al. [2012] created a new method to compare dynamic treatment regimes and along the way, changed the definition of the DTRs after the trial ended. This analysis is different than previous analyses using the same data because the authors changed the definition of viable DTRs based on what had been predetermined as a "missing observation". The protocol required that patients were re-randomized to a new treatment and if they did not complete it, they were classified as missing. However, the treatment plan determined by the protocol was not feasible for patients with toxicity or disease progression. They altered one of the DTRs to include those patients who had to leave the study because of toxicity and specified that the second treatment was the recovery treatment they were given after leaving the trial. The viable DTRs were now defined by efficacy, toxicity and disease progression. The authors also redefined this endpoint because it needed to quantify the health experience of the patient over a pre-specified fixed period, not just their final tumor size or toxicity.

While there is still methodological work to be done or expanded upon, which we will see later, there is still research to be done at the design phase. One clear gap in the literature is a universal sample size formula. While there are sample size calculation publications for SMART designs, but there is no universally accepted calculation for general use. One of these developments is the development of upper bound sample size estimates for censored data. Unfortunately, this sample size formula does not have great generalizability properties because the upper bounds are based on a Kaplan-Meier estimate and the log-rank statistic. Li and Murphy [2011] developed a sample size formula for a two stage SMART for failure time outcomes. The difficulty in such a calculation stems from the variances of the common test statistics. These test statistics depend on the joint distribution of the time, early response determination, and the primary failure time, which are likely to be dependent. The sample size is derived using upper bounds on the variances in place of the usual variances and, hence, the resulting formula only requires the same assumptions of a traditional single stage randomized clinical trial. Using the upper bounds of the variances, the proposed samples size formulas for the Kaplan-Meier estimate (n_K) and the log rank statistic (n_L) are

$$\begin{aligned} n_K &\leq \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma_B^2}{\left\{\bar{F}_1(\tau) - \bar{F}_2(\tau)\right\}^2} \\ n_L &\leq \left(\frac{1}{pq} + \frac{1}{(1-p)q}\right) \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\xi^2 \int_0^t \bar{F}_c(t) \mathrm{d}F_1(t)} \end{aligned}$$

where Z is the z-score for a standard normal distribution, α is the type I error, β is the type II error, $\bar{F}$ is the survival function, τ is the time at the end of the study, $p = p(A_1 = 1)$ and $q = p(A_2 = 1|R = 1)$ are the randomization probabilities, R is the indicator of randomization, ξ is the log hazard ratio. We define

$$\sigma_B^2 = \frac{\bar{F}_1^2(\tau)}{pq}\int_0^\tau \frac{\mathrm{d}\Lambda_1(t)}{\bar{F}_1(t)\bar{F}_c(t)} + \frac{\bar{F}_2^2(\tau)}{(1-p)q}\int_0^\tau \frac{\mathrm{d}\Lambda_2(t)}{\bar{F}_2(t)\bar{F}_c(t)}$$

where Λ is the cumulative hazard function. This sample size calculation has been proven to provide the desired power if the hazards of the alternative are proportional. This sample size calculation is most notable because the nature of chronic diseases allows the outcome, or surrogate outcome, to be thought of in terms of failure, even if death is not the primary endpoint. This means that this sample size formula will be applicable in many settings.

Another important research topic is developing strategies that prevent, and methodology that controls, for missing data. The construct of a SMART requires numerous randomizations and multiple treatment prescriptions which presents unique challenges when analyzing incomplete data. Imputation strategies for handling missing data collected from SMARTs is an understudied area at this time. Shortreed et al. [2014] presented the following five missing data issues: (1) transition between treatment stages does not always occur at pre-specified times, but instead can be determined by a patient outcome; (2) some outcome variables are irregularly spaced while some tailoring variables are collected at regularly scheduled study visits; (3) observing some tailoring variables is dependent on a patient's history, which results in structural missingness for the data-dependent portion of the observed data. (4) individuals are simply lost to follow up leaving the treatment stage; (5) some individuals are lost to follow up entering the treatment stage. Their proposed solution is a flexible imputation strategy that is a time ordered, nested, conditional imputation method which exploits the nearly monotone pattern of missing data found in this type of longitudinal study. It ensures that a complete multivariate prediction distribution exists while obtaining desirable traits for inference across longitudinal outcomes. Assuming missingness at random, this method works best when the data is imputed with a pseudo-Gibbs sampler, which applies repeated iterations through the model. Multiple imputation is one of many strategies used when working with missing data, and the type of strategy often depends on the structure of the data and the nature of the missingness. This method was not compared to other imputation strategies such as inverse probability weighting or likelihood methods, and there is no contingency plan when the missingness is not monotone. It is clear the presented work is exciting and promising progress, but more headway is still needed.

An exciting area of expansion is data collection and treatment allocation using mobile technology. There is strong interest is the ability to increase access to fast, accurate care through mobile technologies that include cell phones, sensors and monitors. The goal is development of evidence based Just in Time Adaptive Interventions (JITAIs) that collects real time data from patients and uses that data to inform the real time delivery of intervention options, such as treatment, dose and timing of care [Nahum-Shani, 2013]. For example, trying to intervene in heavy drinking and smoking, a mobile phone would be administered and participants would be prompted 3 times a day to assess their smoking urge, affect, and drinking behaviors. Urge management interventions would be delivered only if the individual reports the urge to smoke at a specific time. Anytime during the day the user can text either lapse or crave, and a series of encouraging text messages will be sent back to their cell phone. Another example is managing eating disorders. When treating college women with eating disorders, the subject would be provided with a cell phone which receives 5 prompts regarding mood, eating behaviors, exposure, etc. When she reports what is considered a negative mood she is recommended to use one of the treatments provided via a CD. In all instances, the interventions are adapted and delivered through a mobile medium such that patient information can be obtained at any time and responses can be administered at any time. The variety of potential interventions includes reach out interventions, behavioral strategies, cognitive strategies, and goal setting. Tailoring variables can be collected actively (self-reported via prompting or user initiated) or passively (activity level, location, social media activities, number of ignored recommended interventions, etc.). The decision points vary depending on the goal of the treatment plan and can include a random prompt, user requested help, or indication of specific experiences. The decision rules can be deterministic (if the patient reports more than X then give them this, otherwise give them that) or stochastic (determining the probability of an intervention). The corresponding thresholds can be determined and optimized using reinforcement learning (which will be covered in later sections). Even though this concept is in the early stages of development, there are obvious feasibility issues for this kind of treatment implementation, such as cost and monitoring adherence. Innovative methods like this have the ability to change the way patients are treated and can serve as a guide for future treatment and collection methods.

A SMART is an innovative approach to efficiently collect data which accurately estimates DTRs for specific types of patients. The basic design structure has been created, trials are ongoing in the clinical setting and new advances are being developed day by day, but more work needs to be done. The flexibility of the design makes developing broad techniques difficult, but the need and the talent is there to continue to make advancements in what has become an extremely timely, interesting and practical trial design.

3 Analysis Techniques: Single Stage

The list of methodology presented here and in the subsequent section is neither complete nor representative of all available methods, but a simple summary of recent methods employed across a broad range. The purpose is to introduce popular techniques, highlight advancements and display a plethora of methodological options applicable in multiple areas of interest.

Important notation must be introduced so that an individualized treatment rule (ITR) for the single stage paradigm can be properly defined. An ITR differs from a DTR in that it is the personalized rule for a single treatment setting while a DTR is the sequence of decision rules for a multiple treatment setting. Assuming the data is collected from a single stage two arm trial, the treatments will be annotated as $A \in \{-1, 1\}$. These are independent of the patient prognostic variables denoted as $\boldsymbol{X} = (X_1, \ldots, X_p)^T$, where $\boldsymbol{X}$ is a p-dimensional matrix. The observed clinical outcome, Y, can be considered a reward function where larger values are desired. The ITR is a map from the prognostic variable space, $\boldsymbol{X}$, to the treatment space, A, and the optimal ITR is the A which maximizes the expected reward. The distribution of $(\boldsymbol{X}, A, Y)$ is denoted by P with the respective expectation denoted as E. The distribution of $(\boldsymbol{X}, A, Y)$ given the ITR, D (i.e. that $A = D(\boldsymbol{X})$), is denoted as P^D

and the corresponding expectation as E^D. The expected reward under D is

$$V(D) = E^D(Y) = E\left[\frac{I\left\{A = D(\boldsymbol{X})\right\}}{A\pi + \frac{1-A}{2}} Y\right]$$

where $\pi = P(A = 1)$. This $V(D)$ is referred to as the value function for a given D. The optimal ITR, denoted D^*, is estimated as:

$$D^* \in argmax_D V(D) = argmax_D E\left[\frac{I\left\{A = D(\boldsymbol{X})\right\}}{A\pi + \frac{1-A}{2}} Y\right]$$

and is considered the treatment rule, D, which maximizes the value function $V(D)$. The optimal treatment regime is defined as the one that maximizes the average expected outcome [Zhao et al., 2012].

One way to estimate ITRs is to restructure the estimation procedure into a classification problem where the optimal classifier corresponds to the optimal treatment decision. The optimal classifier can be found by estimating the Bayes classifier, which is the one that minimizes the expected weighted misclassification error. This framework allows for estimation of mean outcomes under existing methods such as regression estimation, inverse probability weighted estimation (IPWE) or augmented inverse probability weighted estimation (AIPWE) [Zhang et al., 2012a]. The class of treatment decisions is data driven because it is chosen by minimizing the L1 the expected weighted misclassification error and does not need to be pre-specified.

Define the contrast function as

$$C(\boldsymbol{X}) = \mu(1, \boldsymbol{X}) - \mu(-1, \boldsymbol{X})$$

which can be thought of as the mean difference between treatment options for a given set of prognostic variables. The optimal ITR estimation problem can be transformed into a weighted classification problem such that the optimal treatment rule D^* is found by

$$D^* = argmax_D E\left[D(\boldsymbol{X})C(\boldsymbol{X})\right] = argmax_D E\left(|C(\boldsymbol{X})|\left[I\left\{C(\boldsymbol{X}) > 0\right\} - D(\boldsymbol{X})\right]^2\right)$$

This means that the optimal treatment rule, D^*, is found to be the one that maximizes $E\left(|C(\boldsymbol{X})|\left[I\left\{C(\boldsymbol{X}) > 0\right\} - D(\boldsymbol{X})\right]^2\right)$, which is a weighted classification problem. Each subject belongs to two classes such that class $Z = 1$ contains those subjects who would benefit more from treatment $A = 1$ as opposed to treatment $A = -1$, e.g. $\mu(1, \boldsymbol{X}) > \mu(-1, \boldsymbol{X})$, and $Z = 0$ the opposite. Each observation is also given a weight, $W = |C(\boldsymbol{X})|$, which is the loss that would incur from misclassification. Hence, the optimal ITR is the expected weighted misclassification error under the classification rule $D(\boldsymbol{X})$. Within this classification construct, the problem then decomposes into two critical steps. First, one must construct a suitable estimator of the contrast function using regression, and then invert this to find the estimated optimal treatment rules with an interpretable form using classification methods. This can be extended to the multiple stage scenario as well. This classification prospective falls under the machine learning umbrella. Machine learning, most specifically reinforcement learning, has recently been implemented since it sidesteps the problem of completely modeling the underlying generation model as is necessary in some estimation techniques. Reinforcement learning is a dynamic programming system that dictates which actions are recommended to optimize the expectation of a given reward.

Qian and Murphy [2011] propose a modification of this which first estimates the conditional mean response using L_1 penalized least squares (L_1-PLS) with a rich linear model and then uses that to derive the estimated treatment rule. If the conditional mean is modelled correctly, this method consistently estimates the optimal treatment rule. The finite sample upper bounds of the difference between the mean response from the optimal treatment rule and the mean response from the estimated treatment rule holds even if the linear model for the conditional mean response is incorrect. If the part of the conditional mean model involving the treatment effect is correct then the upper bounds imply that the estimated treatment rule is consistent. These upper bounds can also inform how to choose the tuning parameters involved in the L_1-penalty to create the best rate of convergence. To obtain the ITR the estimated prediction error is minimized then the conditional mean model is maximized over the treatment A. To control for overfitting, L_1 penalized least squares is implemented since the L_1 penalty innately does variable selection. The resulting treatment rules are cheaper to implement and easier to interpret.

The forgoing methods are considered indirect methods of estimation. Indirect estimation refers to techniques that first estimate a quantity reflecting the conditional distribution of the outcome, such as conditional mean, and then uses the resulting model to deduce the optimal ITR [Laber et al., 2014b]. Indirect estimation can be desirable because the initial estimation regarding the outcome can be built using traditional statistical modeling techniques. Unfortunately, optimal ITR estimation requires that the conditional outcome be modeled correctly. Indirect estimation methods often and easily experience model misspecification because of the difficulty of modeling high dimensional, time dependent factors. In high dimensional situations the two-step procedure of estimation and maximization equations can be poor fits. In contrast, direct methods of estimation are solutions to the problems proposed by these other techniques which directly achieves this maximization without requiring the initial estimation step be done with indirect approaches. This direct class of methods immediately estimates the value function for all pre-specified treatment rules and then obtains the optimal treatment rule by maximizing the estimator. Direct estimation methods tend to produce treatment regime estimates that are more precise than indirect methods in the single stage setting due to the associated reduced bias [Zhao et al., 2012].

Zhang et al. [2012b] approaches estimating dynamic treatment rules by assuming a posited regression model. This defines the class of treatment rules while recognizing that it is possible for the model to be misspecified. The optimal treatment regime is estimated by directly maximizing the estimator for the overall population mean outcome under all possible specified treatment plans using a suitable inverse probability weighted estimator. When using observational data, this estimator has the ability to control for possible confounders by estimating propensity scores and exploiting the predicted outcome, which ensures precision of the estimate. Let D^* be the optimal treatment decision, which is the one that corresponds to the largest value of $E\left[Y^*(D)\right]$, where

$$Y^*(D) = Y^*(1)D(\boldsymbol{X}) + Y^*(-1)\left\{1 - D(\boldsymbol{X})\right\}$$

is the potential outcome. The potential outcome is the outcome that would be observed if a randomly chosen patient were to receive treatment regime D. Consider treatment rules of the form $D_\eta(\boldsymbol{X}) = D(\boldsymbol{X}, \eta)$ in the class of all possible treatment rules which is indexed by η and will contain D^* if $\mu(A, \boldsymbol{X}; \beta)$ the posited regression model is correctly specified. Therefore, estimating $\eta^* = argmax_\eta E\left[Y^*(D_\eta)\right]$ and defining $D^*_\eta = D(\boldsymbol{X}, \eta^*)$ will provide an estimator for D^*. To estimate $E\left[Y^*(D_\eta)\right]$, an IPWE or a doubly robust AIPWE can be employed. This estimator is directly maximized in η to obtain an η^* and hence $\widehat{D}^*_\eta(\boldsymbol{X}) = D(\boldsymbol{X}, \widehat{\eta}^*)$. This can easily be extended to the multiple decision situation by estimating $Q(\eta) = E\left[Y^*(D_\eta)\right]$ as a function of η.

Direct methods of estimation can also be restructured into a classification problem which can utilize computer science techniques. This looks at the data by comparing the difference between subjects with observed high and low rewards so that the determination of the actual treatment decisions is associated with the actual treatments received for the different groups. This method is referred to as outcome weighted learning (OWL or O-learning). Developed by Zhao et al. [2012], O-learning is a nonparametric approach which directly optimizes the value function, $V(D)$, where each subject is weighted proportional to their clinical outcome divided by the propensity score, which is the probability of receiving the assigned treatment given the covariates. In the case of a clinical trial, the propensity simplifies to the constant probably of receiving the assigned treatment. Finding the D^* that maximizes $V(D) = E\left[\frac{I\{A = D(\boldsymbol{X})\}}{A\pi + \frac{1-A}{2}} Y\right]$ is equivalent to finding the D^* that minimizes $\bar{V}(D) = E\left[\frac{I\{A \neq D(\boldsymbol{X})\}}{A\pi + \frac{1-A}{2}} Y\right]$ which sets the stage to view this as a weighted classification error. Minimizing the previous expected value can be approximated by minimizing

$$n^{-1} \sum_{i=1}^{n} \frac{Y_i}{A_i\pi + \frac{1-A_i}{2}} I\left[A_i \neq sign\left\{f(\boldsymbol{X}_i)\right\}\right]$$

to find the optimal f^* and then setting

$$D^*(\boldsymbol{X}) = sign\left\{f^*(\boldsymbol{X})\right\}$$

since $D^*(\boldsymbol{X})$ can always be represented as $sign\left\{f^*(\boldsymbol{X})\right\}$. This implies the goal is to find a decision rule which chooses treatments based on their specific prognostic variables. On average, patients with large rewards will be recommended the same treatment that they actually received while patients with small rewards will receive the opposite. This is considered 0-1 loss in the machine learning scenario and is difficult to minimize due to non-convexity and discontinuity. This problem is alleviated by transforming the problem, using a surrogate for the 0-1 loss, so that the goal becomes minimizing

$$n^{-1}\sum_{i=1}^{n}\frac{Y_i}{A_i\pi + \frac{1-A_i}{2}}\left\{1 - A_i f(\boldsymbol{X})\right\}^+ + \lambda||f||^2$$

where $\boldsymbol{X}^+ = max(\boldsymbol{X}, 0)$, and $||f||$ is the norm of f. Therefore, this problem is now a weighted classification problem that can be solved using support vector machine methods.

O-learning has many applications but there are also many ways to extend this line of thinking. Chen et al. presented a one stage clinical trial design for penalized dose finding using a robust analysis method based on the O-learning framework. The method converts the individualized dose selection problem into a penalized weighted regression with truncated L_1 loss. The dose level is assumed to be found on a continuum and a non-trivial extension of O-learning for binary treatments is proposed. The dose finding problem becomes a weighted regression with random outcome where the individual responses are the weights. In the linear case, this framework has the goal of minimizing the loss plus penalty of the form

$$\min_f\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{R_i l_\phi\left\{A_i - f(\boldsymbol{X}_i)\right\}}{2\phi_n p(A_i|\boldsymbol{X}_i)} + \lambda_n\big|\big|f\big|\big|^2\right\}$$

where $\phi = \phi_n$ is non-random parameter in real space, λ_n controls the severity of the penalty on f, $l_\phi\left\{A_i - f(\boldsymbol{X}_i)\right\} = \min\left(\frac{|A-f(\boldsymbol{X})|}{\phi}, 1\right)$, $R^*(a)$ is the potential outcome and $R = \int I(A = a)R^*(a)p(a|x)\mathrm{d}a$. The complexity of $f(x)$ is penalized to prevent overfitting. This function is nonconvex and hence difficult to optimize, so an adaptive difference convex

(DC) algorithm is implemented [Tao and An, 1997]. Considering a linear loss function, the objective function is

$$S = \frac{\lambda_n}{2}||w||_2^2 + \frac{1}{\phi_n}\sum_{i=1}^{n} R_i \min\left(\frac{|A_i = D(\boldsymbol{X}_i)|}{\phi_n}, 1\right)$$

where λ_n is now the tuning parameter. This algorithm minimizes the sequence of convex sub-problems with the intent of solving the original non-convex minimization problem. Therefore, the convex sub-problem becomes a weighted penalized median regression problem. Ultimately, the algorithm concludes when $||w^{t+1} - w^t||$ is smaller than some pre-specified constant, where $w = \sum_{i \in T}(\alpha_i - \bar{\alpha}_i)x_i$. Expanding to the nonlinear framework, the decision function then becomes a function of w and some unknown transformation on $\boldsymbol{X}$. A Gaussian kernel is used to construct a dual problem for nonlinear learning that is solved using quadratic programming. To practically implement this procedure, a nonconvex loss function and a DC algorithm for optimization is employed.

Another common and extremely relevant application of O-learning is estimating ITRs for censored data. Realistically many chronic diseases measure short term success of a treatment as a failure or success. It is desirable to develop methods of estimating treatment regimes that are applicable to survival analysis because it has clear relevance to personalized medicine. When considering censored data, notation is slightly altered. The value function is redefined as

$$V(D) = E^D(T) = E[T|\boldsymbol{X}, A = D(\boldsymbol{X})] = E\left[\frac{I\left\{A = D(\boldsymbol{X})\right\}}{A\pi + \frac{1-A}{2}}T\right]$$

where $T = min(\tau, \tilde{T})$ where $\tilde{T}$ is the survival time and τ is the end of the study [Zhao et al., 2015]. Even though the outcome is redefined, the optimal treatment rule is still the treatment rule which maximizes the value function. The goal is to estimate D^* using censored data following the OWL framework. There are two approaches to estimation. First, one can maximize the estimator of the average survival time. To account for right censoring, the estimated mean survival time is reassigned as the weighted misclassification rate. These weights are comprised of both the observed outcome and the inverse probability of censored weights. To offset bias from a misspecified censoring model, a second method, a doubly robust variation of outcome weighted learning, is formulated. In both instances, the treatment rule is consistent for the optimal rule when the model for either the survival times or censoring times is correctly specified. Note: it is not required that both models be correctly specified. A convex relaxation idea from support vector machines is invoked for construction of the necessary estimation algorithm.

The methodological techniques available for estimating ITRs in the single stage scenario encompass a broad spectrum. Some of these techniques have been extended to apply to the estimation of DTRs but not all have, making this an important area of future work. Because of the nature of sequential decision making, some of the associated techniques cannot easily be extended beyond the single stage setting, so it is important to continue making progress in both areas.

4 Analysis Techniques: Multiple Stages

There is a lot of interest and value in creating techniques which accurately estimate optimal DTRs because the multiple stage scenario most similarly mimics the natural course of a chronic disease. Considering patients often need multiple treatments, individuals respond differently to different treatments at different points in their progression and the longevity of the disease can be unknown, these techniques are important for an adequate treatment plan. It is important to note that each method requires assumptions on the design structure, which may or may not be flexible for each method. Design considerations should be made in line with the chosen analysis technique to ensure all assumptions are met.

In order to properly define DTRs in this setting, notation will be presented that expands on that which is used in the single stage paradigm. Consider a trial with T decision points. For $t = 1, \ldots, T$, let $A_t \in \{-1, 1\}$ be the dichotomous treatment assignment at the t^{th} stage and $\boldsymbol{X}_t$ be the patients' prognostic variables before the t^{th} decision point but after the A_{t-1} treatment assignment. The outcome, or reward, at the t^{th} stage is Y_t where larger values are assumed more desirable. Y_t is assumed to depend on all previous prognostic information $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$, all treatment history $(A_1, \ldots, A_t)$ and previous outcomes $(Y_1, \ldots, Y_{t-1})$. The overall outcome of interest is the total reward $\sum_{t=0}^{T} Y_t$. The DTR is a set of sequential decision rules $D = (D_1, \ldots, D_t)$ which is a map from total patient history, $\boldsymbol{H}_t = (\boldsymbol{X}_1, A_1, \ldots, A_{t-1}, \boldsymbol{X}_t)$, to the treatment space. The value function is then defined as

$$V(D) = E^D \left[\sum_{t=1}^{T} Y_t \right]$$

where E^D is the expectation under the measure P^D which is the probiility distribution for distribution of $(\boldsymbol{X}_1, A_1, Y_1, \ldots, \boldsymbol{X}_T, A_T, Y_T, \boldsymbol{X}_{T+1})$. The value function is the expected long term benefit if the population were to follow regimen D and can also be defined as

$$V(D) = E^D \left[\frac{\left(\sum_{t=1}^{T} Y_t \right) \prod_{t=1}^{T} I \left\{ A_t = D_t(\boldsymbol{H}_t) \right\}}{\prod_{t=1}^{T} \pi_t(A_t, \boldsymbol{H}_t)} \right].$$

Similar to the single stage situation, the value that maximizes the value function $V(D)$

$$D^* \in argmax_D V(D)$$

is the optimal DTR D^* [Zhao et al., 2014].

One of the most popular indirect methods of estimation is a computer science method called Q-learning [Watkins and Dayan, 1992]. Q-learning is a form of reinforcement learning and is a dynamic programming procedure that uses backwards recursion to solve the complex Bellman equation more efficiently using regression models. In the two stage setting, Q-functions are defined as

$$\begin{aligned} Q_2(\boldsymbol{h}_2, a_2) &= E\left[Y | \boldsymbol{H}_2 = \boldsymbol{h}_2, A_2 = a_2 \right] \\ Q_1(\boldsymbol{h}_1, a_1) &= E\left[\max_{a_2} Q_2(\boldsymbol{h}_2, a_2) | \boldsymbol{H}_1 = \boldsymbol{h}_1, A_1 = a_1 \right]. \end{aligned}$$

The Q-functions are conditional expectations where $Q_2(\boldsymbol{h}_2, a_2)$ evaluates the quality of choosing treatment a_2 for patients with history $\boldsymbol{h}_2$ and $Q_1(\boldsymbol{h}_1, a_1)$ evaluates the quality of choosing treatment a_1 for patients with history $\boldsymbol{h}_1$ assuming that the best second stage intervention has chosen. This can be extended to more than 2 phases such that

$$Q_t(\boldsymbol{h}_t, a_t) = E\left[\max_{a_t} Q_{t+1}(\boldsymbol{h}_{t+1}, a_{t+1}) | \boldsymbol{H}_t = \boldsymbol{h}_t, A_t = a_t\right]$$

would evaluate the quality of choosing a_t for patients with history $\boldsymbol{h}_t$ assuming the best intervention is chosen at all future stages. In practice, these Q-functions are not known but must be estimated. For illustration, consider the linear form as

$$Q_t(\boldsymbol{h}_t, a_t) = \boldsymbol{h}_{t,1}^T \beta_{t,1} + a_t \boldsymbol{h}_{t,2}^T \beta_{t,2}.$$

In the two stage scenario, estimating the Q-functions, $\widehat{Q}_t(\boldsymbol{h}_t, a_t)$, is a three step procedure. First, using ordinary least squares regression, the estimates $\widehat{\beta}_{2,1}$ and $\widehat{\beta}_{2,2}$ are obtained by regressing the patient history on Y_2. Those estimates are used to estimate the fitted Q function at the second stage $\widehat{Q}_2(\boldsymbol{h}_2, a_2) = \boldsymbol{h}_{2,1}^T \widehat{\beta}_{2,1} + a_2 \boldsymbol{h}_{2,2}^T \widehat{\beta}_{2,2}$. The stage 1 pseudo outcome is $\tilde{Y}_1 = Y_1 + \max_{a_2} \widehat{Q}_2(\boldsymbol{h}_2, a_2)$. Note that if the outcome is only collected at the final stage (in other words, there is only one Y so there is no Y_2, Y_1), the stage 2 outcome is Y and the stage 1 psuedo outcome is $\tilde{Y} = \max_{a_2} \tilde{Q}_2(\boldsymbol{h}_2, a_2)$. The first stage patient history is regressed on $\tilde{Y}_1$ to obtain the estimates $\widehat{\beta}_{1,1}$ and $\widehat{\beta}_{1,2}$. The first stage fitted Q function is $\widehat{Q}_1(\boldsymbol{h}_1, a_1) = \boldsymbol{h}_{1,1}^T \widehat{\beta}_{1,1} + a_1 \boldsymbol{h}_{1,2}^T \widehat{\beta}_{1,2}$. Finally, the estimated optimal treatment decision is given by

$$D_t^*(\boldsymbol{h}_t) = argmax_{a_t} \widehat{Q}_t(\boldsymbol{h}_t, a_t).$$

Estimating the Q-functions is similar for three or more stage implementation where the predicted future outcome is used to create the estimates for the previous stage estimated Q-function.

While Q-learning is a very popular estimation technique, it is not without its limitations and suffers from some undesirable properties such as irregularity, non-smoothness and asymptotic bias [Robins, 2004]. The irregularity problem occurs in Q-learning when the last stage treatment is non-unique for some subjects in the population, which causes bias and inaccurate inference. To remedy this, Goldberg et al. [2013] uses special adaptive weights within the penalization. This corrects for the non-regularity condition by concentrating on the indifference hyperplane of patient covariates where two treatment have the same effect. The indifference hyperplane is the covariate region where there is no difference between treatments. Solving this irregularity condition involves correctly identifying the covariate values which lie on this hyperplane. This adaptive penalized Q-learning procedure can handle continuous covariates and performs better than regular penalized Q-learning method.

Instead of the typical first stage of Q-learning (which involves solving the minimization problem) the adaptive minimization problem involves solving

$$\Phi_{2n}(\theta_2) = \sum_{i=1}^{n} \{Y_{2i} - Q_2(\boldsymbol{h}_{2i}, a_{2i}; \beta_{2,1}, \beta_{2,2})\}^2 - \frac{\lambda_n}{n} \sum_{i=1}^{n} \widehat{\omega}_{ni} |\beta_{2,2}' \boldsymbol{H}_{2i(2)}|$$

where $\widehat{\omega}_{ni}$ are the data driven weights and λ_n is the tuning parameter. Then, $\tilde{\theta}_2$ (in traditional Q learning this is the set of parameters which minimizes the ordinary least squares regression function at the second treatment decision time point) is the minimizer of Φ_{2n} and the remaining steps of Q learning are the same after substituting in $\tilde{\theta}_2$ for the normal estimator. In order to obtain the oracle property (which means the estimator behaves asymptotically as if the indifference plane is already known) the selection of weights is critical. The goal is to find weights that penalize the observations that are close to or are on the indifference hyperplane and that provide weights that go asymptotically to zero for observations far from the hyperplane. This will help define where the indifference hyperplane is and resolve the irregularity problem.

As in most statistical research areas, after developing an estimator the next step is to assess its properties, oftentimes with the use of inference techniques such as confidence intervals. When estimating optimal DTR, common approaches such as Q-learning involve estimation and interference of parameters that are non-smooth functions of the underlying generative distribution. As was mentioned before, these estimates are irregular and asymptotically biased. Standard asymptotic approximations to the sampling distributions cannot be used to directly form reliable confidence intervals or carry out hypothesis testing [Laber et al., 2014b]. One method to construct confidence intervals is an m-out-of-n bootstrap procedure to correct the nonsmoothness. The confidence sets are constructed in a way to adapt to the irregularity present in the underlying generative model. The data driven adaptive choice of m produces asymptotically correct confidence sets under fixed alternatives. This method has the added benefit of conceptual and computational simplicity with a corresponding R package [Chakraborty et al., 2013].

The proposed adaptive scheme to select m is a class of resample sizes given by $m = n^{f(p)}$. The suggested simple form is proposed to be

$$\widehat{m} = n^{\frac{1+\alpha(1-\widehat{p})}{1+\alpha}}$$

where $\alpha > 0$ is a tuning parameter. α controls the smallest acceptable sample size and may be dictated by practical considerations or tuned using the data. A bootstrap algorithm is used for choosing α using data which appears to reduce conservatism. When the parameter of interest is a linear function of the parameters, $(c'\theta_{1,n})$, the algorithm first draws B_1 m-out-of-n first stage bootstrap samples and estimates $c^T\widehat{\theta}^{b_1}_{1,n}$. α is fixed at the smallest value in the grid and $\widehat{m}^{b_1}$ is then calculated using the equation above. This is repeated by drawing B_2 $\widehat{m}^{b_1}$-out-of-n second stage bootstrap samples and calculating $c^T\widehat{\theta}^{(b_1,b_2)}$, which is a double bootstrapped version of the estimate. For all b_1, compute $\left(\frac{\eta}{2}\right)$x100 and $\left(1-\frac{\eta}{2}\right)$x100 percentiles which are the lower bounds and upper bounds defined as $\widehat{l}^{b_1}_{DB}$ and $\widehat{u}^{b_1}_{DB}$ respectively. The coverage rate of the double bootstrap confidence interval from all first stage bootstrap data sets is

$$\frac{1}{B_1}\sum_{b_1=1}^{B_1} I\left(c^T\widehat{\theta}^{b_1}_{1,n} - \frac{\widehat{u}^{b_1}_{DB}}{\sqrt{\widehat{m}^{b_1}}} \leq c^T\theta_{1,n} \leq c^T\widehat{\theta}^{b_1}_{1,n} - \frac{\widehat{l}^{b_1}_{DB}}{\sqrt{\widehat{m}^{b_1}}}\right).$$

Increase α to the next highest value on the grid until the coverage rate is at or exceeds the nominal value and in that case pick the current value of α as the final value. The process is repeated until the coverage rate of the double bootstrap confidence interval attains a nominal coverage rate or all the options on the grid are exhausted.

Another methodology that can accommodate the irregularity from using Q-learning to estimate parameters is the locally consistent Adaptive Confidence Interval (ACI) [Laber et al., 2014b]. When construction of DTRs using Q-learning, there is particular interest in reducing bias of first stage coefficients. If the Q-function is near 0 with high probability there will be issues approximating the distribution of $\sqrt{n}\left(\widehat{\beta}_1 - \beta_1^*\right)$. Once the asymptotically biased parameters are identified, given the correct amount of shrinkage, a shrinkage estimator can reduce the bias. However, shrinking too aggressively leads to bad performance in finite samples. Constructing valid confidence intervals for irregular estimators is a difficult task because estimating the sampling distribution of the estimator cannot be done uniformly. The proposed solution is a locally consistent confidence interval for linear combinations of the first stage coefficients. The interest is not in construction of second stage confidence intervals because they can be estimated using standard methods for least square estimators. Since it is not possible to construct a uniformly convergent estimator of the limiting distribution of $\sqrt{n}\left(\widehat{\beta}_1 - \beta_1^*\right)$, for a given constant c the proposed method bounds $c^T\sqrt{n}\left(\widehat{\beta}_1 - \beta_1^*\right)$ between two regular uniformly convergent upper and lower bounds. These smooth bounds can be bootstrapped to form a confidence set for $c^T\beta_1^*$. The extension to more than two stages is straightforward as the last stage uses standard methods for least squares estimation, so the ACI would be used on all previous stages.

Similar to previously discussed analysis techniques, for medical research it is important to develop these techniques to accommodate censored data. Goldberg and Kosorok [2012] developed a Q-learning algorithm that allows for censored data when the outcome of interest is survival time and allows for a flexible number of stages in a randomized trial. Q-learning is expanded upon by using inverse probability censoring weighting to account for censored observations.

For each $t = 1, \ldots, T$, let the state S_t be the pair $S_t = (\boldsymbol{X}_t, Y_{t-1})$ where $\boldsymbol{X}_t$ is either a vector of covariates describing the condition of the patient before time t or it is null. If $\boldsymbol{X}_t$

is null then a failure happened during the t^{th} stage. Let Y_{t-1} be the length of time between decision points t and $t-1$. Hence, $\sum_{j=1}^{t} Y_j$ is the total survival time, or reward, up to and including stage t. Let $C \in [0, \tau]$ be the censoring variable. The goal is to find a policy that maximizes the expected rewards. Then, the optimal policy, D^*, is the one that approximately maximizes over all policies of $E_{0,\pi}\left[\left(\sum_{t=1}^{T} Y_t\right) \wedge \tau\right]$ where $\bar{T}$ is the random number of stage for the subject. This optimal policy is found using a three step algorithm. First the problem is mapped to an auxiliary problem. The auxiliary problem creates modified trajectories of a fixed length T and the modified sum of the rewards is less than or equal to τ to account for censoring. Next, the Q-functions are approximated $\{\widehat{Q}_1, \ldots, \widehat{Q}_T\}$ using the original Q-function framework. Last, the optimal treatment rule, D^*, is found by maximizing $\widehat{Q}_t$ over all possible a_t.

Recall the methodology introduced in the previous section for estimation of ITRs in a single stage. Two of those methods will be expanded on when estimating DTRs for multiple stages of treatment. In the single decision scenario presented by Zhang et al. [2012b] the estimation procedure was restructured into a classification problem. In this case the optimal classifier corresponds to the optimal treatment decision. The optimal classifier was found by estimating the Bayes classifier which is the one that minimizes the expected weighted misclassification error. This can be expanded upon for the two decision point scenario based on reassessing the problem as a monotone coarsening problem using an AIPWE to estimate the mean outcome [Zhang et al., 2013]. Assign Y^* to be the often unobserved potential outcome and Y^*_D to be the potential outcome associated with treatment regime D. The optimal treatment regime D^* is that which satisfies $E\left[Y^*(D^*)\right] \geq E\left[Y^*(D)\right]$, meaning it is that which maximizes the expected potential outcome. The problem is cast into a monotone coarsening problem where the coarsening happens at random if, for each t, the probability that the data are coarsened at level t given the full data depends only on the data observed at level t. Then, from Robins et al. [1994], under these coarsening assumptions if the coarsening mechanism is correctly defined then asymptotically linear consistent estimators for $E\left[Y^*(D_\eta)\right]$ for a fixed η have the form

$$\frac{\sum_{i=1}^{n} I(C_{\eta,i}=\infty)}{K_{\eta,k}\bar{\boldsymbol{X}}_{k,i}} Y_i + \frac{\sum_{i=1}^{n} \left\{ I(C_{\eta,i}=k) - \lambda_{\eta,k}(\bar{\boldsymbol{X}}_{k,i}) I(C_{\eta,i}>k) \right\}}{K_{\eta,k}(\bar{\boldsymbol{X}}_{k,i})} L_k(\bar{\boldsymbol{X}}_{k,i})$$

where $L_k(\bar{\boldsymbol{X}}_{k,i})$ are arbitrary functions, $C_{\eta,i}$ is the discrete coarsening variable, $K_{\eta,K}(\bar{\boldsymbol{X}}_K) = \prod_{k'=1}^{k} \left\{ 1 - \lambda_{\eta,k'}(\bar{\boldsymbol{X}}_{k'}) \right\}$ and $\lambda_{\eta,k'}(\boldsymbol{X}'_k)$ is the hazard function. The left side of the above estimator is on its own a consistent estimator if $\lambda_{\eta,k}(\bar{\boldsymbol{X}}_k)$ is correctly specified. Then the entire estimator is a doubly consistent robust estimator for $E\left[Y^*(D\eta)\right]$ if either $\lambda_{\eta,k}(\bar{\boldsymbol{X}}_k)$ are correctly specified or if $L_k(\bar{\boldsymbol{X}}_{k,i}) = Y^*(D_\eta)| \left\{ \bar{\boldsymbol{X}}^*_k(\bar{D}_{\eta_{k-1}}) = \bar{\boldsymbol{x}}_k \right\}$.

O-learning was presented in Section 3 as a machine learning approach which directly optimizes the value function $V(D)$ where each subject's weight is proportional to their clinical outcome. This is a weighted classification error problem since finding the D^* that maximizes $V(D)$ is equivalent to finding the D^* that minimizes $\bar{V}(D)$. O-learning can also be expanded to the two stage paradigm using a few strategies. One such way is backwards outcome weighted learning (BOWL) which modifies existing algorithms to solve a sequence of weighted classification problems [Zhao et al., 2014]. The algorithm is backwards fitting and at each time point T, the algorithm is as follows. The goal in the first stage is to minimize

$$\frac{n^{-1}\sum_{i=1}^{n} \left[Y_{iT}\phi\left\{ A_{iT} f_T(\boldsymbol{H}_{iT}) \right\} \right]}{\pi_T(A_{it}, \boldsymbol{H}_{it})} + \lambda_{T,n}||f_T||^2$$

with respect to f_T where $\widehat{f_T}$ is the minimizer. The optimal decision rule is

$$\widehat{D}_T(\boldsymbol{h}_T) = sign\left\{ \widehat{f_T}(\boldsymbol{h}_T) \right\},$$

and this stage is essentially equivalent to the single stage outcome weighted learning found in Zhao et al. [2012] and has a similar dual objective function as found in support vector machines. The second stage is, for $t = T-1, T-2, \ldots, 1$, to backward sequentially minimize

$$n^{-1}\sum_{i=1}^{n} \frac{\left(\sum_{j=1}^{T} Y_{ij}\right) \prod_{j=t+1}^{T} I\left\{ A_{ij} = \widehat{D}_j(\boldsymbol{H}_{ij}) \right\}}{\prod_{j=1}^{T} \pi_j(A_{ij}, \boldsymbol{H}_{ij})} x\phi\left\{ A_{ij}, f_t(\boldsymbol{H}_{it}) \right\} + \lambda_{t,n}||f_t||^2$$

where $\widehat{D}_{t+1}, \ldots, \widehat{D}_T$ are previously obtained.

A disadvantage of BOWL is the number of observations utilized by the algorithm decreases geometrically as t decreases. The authors explain this can be solved using iterative outcome weighted learning (IOWL) which involves re-estimating the optimal treatment rule at stage 2 after the stage 1 rule is estimated. This estimate is based on the subset of patients whose stage 1 treatment assignments are consistent with the optimal rule. The procedure would continue with a re-estimation of the stage 1 treatment rule based on the new optimal stage 2 rule. IOWL allows the exploration of different subjects through iterative re-estimation.

Zhao et al. [2014] also present simultaneous outcome weighted learning (SOWL) which frames estimation of DTRs as a single classification problem. This is an effective way of looking at the problem because a multiple stage treatment plan has not previously been estimated simultaneously using a single algorithm. The method directly optimizes the empirical counterpart of the value function in one step. Since this problem is computationally difficult (mostly because of the discontinuity of the indicator functions) a continuous and concave surrogate function is used in lieu of the product of indicators that would usually be required. In the two decision point scenario, the surrogate reward function is chosen to mimic hinge loss: $\psi(Z_1, Z_2) = min(Z_1 - 1, Z_2 - 1, 0) + 1$ where $Z_1 = A_1 f_1(H_1)$ and $Z_2 = A_2 f_2(H_2)$. Hence the SOWL estimator maximizes

$$n^{-1} \sum_{i=1}^{n} \frac{(\sum_{j=1}^{2} Y_{ij})\psi\left\{A_{i1} f_1(\boldsymbol{H}_{i1}), A_{i2} f_2(\boldsymbol{H}_{i2})\right\}}{\prod_{j=1}^{2} \pi_j(A_{ij}, \boldsymbol{H}_{ij})} - \lambda_n \Big(||f_1||^2 + \big||f_2|\big|^2\Big),$$

where the tuning parameter λ_n controls the amount of penalization. This can easily be extended to more than 2 stages.

Much exciting and significant work is being done in developing treatment rules for dynamic sequential decision making. To this effect, there have been promising advancements, but these methods often times need to be expanded upon or adapted to various specific settings. Science will forever be changing and research will be forever trying to keep up. While existing methodology can always be improved and generalized, there will always be a need for new and innovative mechanisms for estimating optimal DTRs.

5 Related Topics

5.1 Variable Selection

Variable selection is an important component of estimating optimal DTRs because tailoring variables are used to adapt the treatment plan to the individual. The goal is to avoid a priori hand picking tailoring variables, but instead use the data to select a subset of the tailoring variables that estimates a decision rule similar to the optimal rule chosen when using all variables. Including all possible variables as tailoring variables is inefficient and will often lead to over fitting. Once the tailoring variables are selected, they can be used when optimizing DTRs.

Biernot and Moodie [2010] discuss two computer science techniques that can be used for variable selection: the S-score criterion and the use of reducts. The S-score of a variable shows the expected increase in response that is observed by choosing the treatment based on the value of that variable. It combines the interaction of the covariate with the treatment and the proportion of the population exhibiting variability in that covariate. Higher values indicate stronger relationships between the variable and the treatment, and shows that a large proportion of patients would experience change in the optimal action if the variable was taken into consideration. This scoring is used to rank potential variables but each variable is evaluated separately, meaning correlation between variables is not taken into consideration. The S-score could also be used sequentiality such that the variable with the highest score is first selected, then the variable with the second highest score given the first variable is selected and so on.

The reducts approach was developed from rough set theory in computer science. The positive region is a set of all observations that can be uniquely classified into one equivalence class based on the non-decision variables. The reduct is the minimal set of tailoring variables that classifies individuals into unique decision equivalence classes as well as the complete set of variables does. Reducts help eliminate redundant variables while preserving information regarding the similarity of individuals in the sample. In the scenario with multiple reducts, one can select the variables most frequently seen in the reducts or can select amongst reducts by choosing the set of covariates with the highest S-score. This last hybrid method is believed to combine the strengths of these two methods. Unfortunately, reducts cannot be used on a continuous outcome.

Another way to approach variable selection is to simultaneously estimate optimal treatment regimes and identify significant variables. This is done with a penalized regression model that finds which variables interact with the treatment using a new loss based framework. Lu et al. [2013] introduces a method which does not require estimating the baseline mean function for the outcome of interest and is easily adaptable to shrinkage methods for variable selection based on their loss structure making it quickly implementable with current software. The authors suggest the loss function

$$L_{n,\phi}\left(\beta,\gamma\right)=\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\phi\left(\boldsymbol{X}_i;\gamma\right)-\beta^T\tilde{\boldsymbol{X}}_i\left\{A_i-\alpha\left(\boldsymbol{X}_i\right)\right\}\right]^2$$

where n is the number of observations, Y_i is the i^{th} patient's outcome, $\boldsymbol{X}_i$ is the i^{th} patient's prognostic variables, $\tilde{\boldsymbol{X}}=(1,\boldsymbol{X}^T)^T$, $A_i\in\{-1,1\}$ represents the dichotomous treatment choice, $\alpha(x)$ denotes the propensity score, and ϕ is an arbitrary function with a constant model for $\phi:\phi(x;\gamma)=\gamma$ and a linear model for $\phi:\phi(x;\gamma)=\gamma^T\tilde{x}$. This characterization of the loss function increases simplicity in adopting shrinkage penalties for variable selection. Employing the adaptive lasso penalty (or alternatively, the SCAD or minimax concavity penalty) the solution is the β which satisfies

$$\min_{\beta}L_{n,\phi}(\beta,\tilde{\gamma})+\lambda_n\sum_{j=1}^{p+1}w_j|\beta_j|$$

where λ_n is a tuning parameter and w_j are the weights such that $w_j^{-1}=|\tilde{\beta}_j|$ is used. Aside from estimating the optimal DTR, these β values are used to determine which variables are important in selecting the optimal DTR such that the important variables are those with nonzero coefficients.

Variable selection is an important part of estimating optimal DTRs because a parsimonious selection of tailoring variables will make the estimation faster and more reliable. Three methods have been presented here for these purposes, but more methodology has been published. It is imperative to use a selection technique that is relevant for the data set and can be effectively integrated into the analysis plan.

5.2 Multiple Outcomes

While development of these advanced estimation techniques is necessary to effectively personalize medical care, treatment of the entire person, not just one disease, should be considered as well. In the clinical setting, the patient or caregiver will likely be interested in balancing competing outcomes such as survival, quality of life and financial burden. While survival may be the ultimate goal for a cancer patient, a single mother of two may prefer a treatment that allows her to work (higher quality of life) which could lead to a longer treatment course, or patients may need to balance the financial burden with their treatment plan. This is a very new area of study inside the precision medicine umbrella, but it is important and quickly developing.

A crucial step in balancing competing outcomes for personalized patient care is eliciting the patient's or physician's preferences regarding the ideal tradeoff between the outcomes. Lizotte et al. [2012] produced an inverse preference elicitation approach which first considers all of the actions available at any given state. Then, for each action, asks what range of preferences makes that action a good choice. This provides a large amount of information about the potential actions at each state. The patients also have the ability to see if their preference is near the boundary or see if a small change in preference results in a change of recommended treatment. In this situation, the patient can feel confident that both treatments perform well and make the decision based on other potentially minor preferences. This method provides an efficient algorithm that computes the optimal policy for varying reward functions and provides insight into how the choice of reward influences the optimal treatment decision.

Alternatively, Laber et al. [2014a] developed a way to construct DTRs that does not require tradeoffs between outcomes by eliciting a clinically significant difference for each respective outcome. When the algorithm concludes that no single treatment is best, the patient's or doctor's preferences are able to be incorporated. They are free to choose the treatment arbitrarily based on other qualities that matter to them, such as cost. This method involves set-valued dynamic treatment regimes that take as input the current patient history and provide as output a set of recommended treatments.

Considering just the static set valued decision rules for the single decision time point, let Y and Z be the competing outcomes and Δ_Y, Δ_Z represent the predetermined clinically meaningful difference in the respective outcomes. In the ideal situation, the algorithm will produce one recommended treatment if that treatment provides significant benefit to one outcome without producing significant detriment to the other. However, in all other cases, the algorithm will produce a set of recommended treatments and the decision is left up to the clinician or patient. With

$$\tau_Y(h) = E\left[Y|H=h, A=1\right] - E\left[Y|H=h, A=-1\right]$$

$$\tau_Z(h) = E\left[Z|H=h, A=1\right] - E\left[Z|H=h, A=-1\right]$$

then the ideal decision rule $\pi_{\Delta}^{ideal}(h)$ is either

1. $sign\left\{\tau_Y(h)\right\}$ if $|\tau_Y(h)| > \Delta_Y$ and $sign\left\{\tau_Y(h)\right\}\tau_Z(h) > -\Delta_Z$
2. $sign\left\{\tau_Z(h)\right\}$ if $|\tau_Z(h)| > \Delta_Z$ and $sign\left\{\tau_Z(h)\right\}\tau_Y(h) > -\Delta_Y$
3. $\{-1, 1\}$ otherwise

Generalizing this procedure to dynamic set valued decision rules for two or more decision points, the algorithm is backwards regressive and Q-learning with linear working models is used to estimate $r_Y(h)$ and $r_Z(h)$. Using ordinary least squares, the optimal treatment set can be estimated from patient history. At the second stage, estimating $\pi_{2\Delta}^{ideal}$ is essentially the same as described for the single decision point situation. To find $\pi_{2\Delta}^{ideal}$, it is assumed that the best single treatment decision (not a set-valued decision) was made, τ_2. Then, $\pi_{1\Delta}^{ideal}(h_1, \tau_2)$, at the first stage is

1. $sign\left\{\tau_Y(h_1, \tau_2)\right\}$ if $|\tau_Y(h_1, \tau_2)| > \Delta_Y$ and $sign\left\{\tau_Y(h_1, \tau_2)\right\}\tau_Z(h_1, \tau_2) > -\Delta_Z$
2. $sign\left\{\tau_Z(h_1, \tau_2)\right\}$ if $|\tau_Z(h_1, \tau_2)| > \Delta_Z$ and $sign\left\{\tau_Z(h_1, \tau_2)\right\}\tau_Y(h_1, \tau_2) > -\Delta_Y$
3. $\{-1, 1\}$ otherwise

Hence, the optimal decision rule is the set valued rule

$$\pi_{1\Delta}^{ideal}(h_1) = \bigcup_{\tau_2 \in C\left(\pi_{2\Delta}^{ideal}\right)} \pi_{1\Delta}^{ideal}(h_1, \tau_2)$$

where $C\left(\pi_{2\Delta}^{ideal}\right)$ is the set of all treatment options compatible with $\pi_{2\Delta}^{ideal}$ and τ_2 is compatible with π_2 if and only if $\tau_2(h_2) \in \pi_2(h_2) \quad \forall \quad h_2$.

Progressing from estimating techniques for one outcome to multiple outcomes is the next step in the natural course of this specialization. Practically, patients and doctors will have more than one goal when developing a treatment plan and these complicated preferences should be taken into consideration if possible. As this is a new area, there is continuing progress.

5.3 DTRs for Observational Data

As clinical researchers were experimenting with new trial designs to improve patient treatment plans, epidemiologists were simultaneously investigating the relationships of time varying continuous exposures to various outcomes in observational data [Lavori and Dawson, 2014]. They developed a longitudinal generalization of Rubin's potential outcomes [Holland, 1986] for inference between exposure and outcomes for observational data which naturally led to an interest in developing DTRs for this data. These DTRs assume the exposures are assigned in a way that is conditionally independent of the potential future responses given the history of the patients and treatments up to the current state. This resembles the assumptions made when developing treatment plans using randomized prospective data. Sometimes situations arise where a randomized trial is impossible or impractical, so it is efficacious to perform an observational study instead or, quite simply, observational data may exist from a preexisting study. Using this resource can be more expedient and reduce significant financial burden because new patients are not needed and no treatments are given. The development of DTRs is often exploratory and hence it is potentially important to be able to estimate these treatment plans using large samples of observational data with the intention of validating the DTR in a confirmatory randomized trial. Furthermore, collecting observational data on time-varying outcomes, predictors and confounders can sometimes emulate a randomized trial that lacks baseline randomization.

An effective estimation technique when eliciting DTRs from observation data is the parametric g-formula. The parametric g-formula uses Robins' G-estimation to naturally estimate DTRs and can appropriately adjust for time-dependent cofounding variables [Young et al., 2011]. This formula is an alternative to inverse probability weighting that provides more efficient estimates but requires more parametric modeling assumptions. It can be computed for the potential outcome by non-parametrically estimating the value of each density function for all of the possible histories of patients. The formula takes the sum over the histories but requires that all possible covariates need to be categorical. For high dimensional data, the g-formula can only be carried out by estimating the density functions using parametric modeling assumptions then taking the sum over the histories via Monte Carlo simulation. Because of distributional a priori knowledge for certain histories, when estimating the g-formula parametric models are not needed to be imposed over all components of the densities and histories.

G-estimation in the context of estimating DTRs has advantages over traditional parametric approaches for producing consistent estimators. However, these estimators are asymptotically biased under a given structural nested mean model for certain data distributions (coined exceptional laws) and exhibit non-regular behavior. To combat this, Moodie and

Richardson [2009] presented a new approach called Zeroing Instead of Plugging In (ZIPI). ZIPI provides estimators nearly identical to those provided by g-estimation but with the benefit of reducing bias in those situations when decision rule parameters are not shared across intervals. More specifically in the context of constructing DTRs, the observed longitudinal distribution function is exceptional if at some interval there is a positive probability that the true optimal decision rule is not unique. For a distribution to be exceptional, the blip function must include at least one covariate (such as the previous treatment), and the probability that the true blip function has value 0 is positive. The proposed ZIPI method is considered a modification of g-estimation when there is no parameter sharing and detects and reduces bias in the presence of exceptional laws.

Moodie et al. [2012] extended one of the more frequently utilized methods of optimal DTR estimation, Q-learning, to accommodate observational data. A soft threshold approach is used which has a good performance in terms of bias and coverage in the non-regular settings. This approach shrinks the problematic term in the potential outcome towards zero. When using Q-learning for observational data, the basic approach requires the construction of a propensity score, $\pi(x) = P(A = 1|\boldsymbol{X} = \boldsymbol{x})$, or treatment model followed by some form of adjustment. It assumes the treatment received is independent of known covariates given the propensity score. This leads to unbiased estimates of the treatment effect based on the conditional expectation modeling the outcome given the propensity score. In inverse probability weighting analysis, the weights are used to create a pseudo-sample so that the treatment does not depend on the variables in the pseudo-sample. Including covariates into the models for the Q-functions can be implemented in four ways which perform well: including the covariates as linear terms in the Q-function, including the propensity score as a linear term in the Q-function, including quintiles of the interval-specific propensity score (which depends on a time varying confounding variable) as covariates in the j^{th} interval Q-function, and IPTW weighted with $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ defined as in traditional Q-learning.

Not all data can be collected using a randomized clinical trial, so it is imperative to develop methodologies that can estimate optimal DTRs using this observational data. This data is difficult because there is no randomization, but the assumption of independence between exposure and predicted future outcome simplifies the problem to an extent. While the parametric g-formula and Q-learning are great options for analysis, not all data will benefit from these techniques and more work would provide more resources and flexibility for scientists.

6 Conclusion

Data collected from SMARTs is an integral part of effectively developing DTRs. This has been a hot topic in clinical trial design in recent years and appears to be the future of clinical practice. They are flexible, informative studies which utilize all of the participants and can answer more questions than a traditional RCT. Pilot studies are essential for designing a SMART so that resources are optimized and the essential information is properly collected. While there has been progress made in properly designing SMARTs, there is still a lot of work to be done particularly in sample size estimation and handling missing data. A plethora of direct and indirect techniques have been presented here to highlight some of the most current methods available so the reader can make informed decisions when creating their analysis plan for a SMART design or even when working with observational data. One of many future directions in this area is practical implementation in clinical practice which involves estimating DTRs for competing outcomes, an area which is quickly expanding. The recent progress over the last ten years has been very exciting, but there are still many areas that could use further development and many topics that have not been explored yet. The future of implementing SMARTs for developing DTRs to personalize medicine is bright and promising.

References

Daniel Almirall, Scott N. Compton, Meredith Gunlicks-Stoessel, Naihua Duan, and Susan A. Murphy. Designing a Pilot Sequential Multiple Assignment Randomized Trial for Developing an Adaptive Treatment Strategy. *Statistics in Medicine*, 31(17):188–1902, 2012.

Peter Biernot and Erica E.M. Moodie. A Comparison of Variable Selection Approaches for

Dynamic Treatment Regimes. *The International Journal of Biostatistics*, 6(1):1557–4679, 2010.

Bibhas Chakraborty and Erica E.M. Moodie. *Statistical Methods for Dynamic Treatment Regimes*. Springer, 2013.

Bibhas Chakraborty, Eric B Laber, and Yingqi Zhao. Inference for Optimal Dynamic Treatment Regimes Using an Adaptive m-Out-of-n Bootstrap Scheme. *Biometrics*, 69(3):714–723, 2013.

Guanhua Chen, Donglin Zeng, and Michael R Kosorok. Personalized Dose Finding using Outcome Weighted Learning. Manuscript submitted for publication.

Linda M. Collins, Susan A. Murphy, Vijay N. Nair, and Victor J. Strecher. A Strategy for Optimizing and Evaluating Behavioral Interventions. *Annals of Behavioral Medicine*, 30 (1):65–73, 2005.

Linda M. Collins, Susan A. Murphy, and Victor Stretcher. The Multiphase Optimization Strategy (MOST) and the Sequential Multiple Assignment Randomized Trial (SMART): New Methods for More Potent eHealth Interventions. *American Journal of Preventative Medicine*, 32(5):112–118, 2007.

Linda M. Collins, Inbal Nahum-Shani, and Daniel Almirall. Optimization of behavioral dynamic treatment regimens based on the sequential, multiple assigment, randomization trial (SMART). *Clinical Trials*, 11(4):426–434, 2014.

Yair Goldberg and Michael R. Kosorok. Q-learning with Censored Data. *Annals of Statistics*, 40(1):529–560, 2012.

Yair Goldberg, Rui Song, and Michael R Kosorok. Adaptive Q-learning. *Institute of Mathematical Statistics Collections*, 9(1):150–162, 2013.

Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

Michael R Kosorok and Erica EM Moodie. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, volume 21. SIAM, 2016.

Eric B. Laber, Daniel J. Lizotte, and Bradley Ferguson. Set-valued Dynamic Treatment Regimes for Competing Outcomes. *Biometrics*, 70(1):53–61, 2014a.

Eric B. Laber, Daniel J. Lizotte, Min Qian, William E. Pelham, and Susan A. Murphy. Dynamic Treatment Regimes: Technical Challenges and Applications. *Electronic Journal of Statistics*, 8(1):1225–1272, 2014b.

Philip W. Lavori and Ree Dawson. Dynamic treatment regimes: Practical design considerations. *Clinical Trials*, 1(1):9–20, 2004.

Philip W. Lavori and Ree Dawson. Introduction to Dynamic Treatment Strategies and Sequential Multiple Assignment Randomization. *Clinical Trials*, 11(4):393–399, 2014.

Zhiguo Li and Susan A Murphy. Sample size formulae for two-stage randomized trials with survival outcomes. *Biometrika*, 98(3):503–518, 2011.

Daniel J. Lizotte, Michael Bowling, and Susan A. Murphy. Linear Fitted-Q Iteration with Multiple Reward Functions. *Journal of Machine Learning Research*, 13:3253–3295, 2012.

Wenbin Lu, Hao Helen Zhang, and Donglin Zeng. Variable Selection for Optimal Treatment Decision. *Statistical Methods in Medical Research*, 22(5):493–504, 2013.

Erica E. M. Moodie and Thomas S Richardson. Estimating Optimal Dynamic Regimes: Correcting Bias under the Null. *Scandinavian Journal of Statistics*, 37(1):126–146., 2009.

Erica E.M. Moodie, Thomas S. Richardson, and David A. Stephens. Demystifying optimal dynamic treatment regimes. *Biometrics*, 63(2):447–455, 2007.

Erica E.M. Moodie, Bibhas Chakraborty, and Michael S. Kramer. Q-learning for Estimating Optimal Dynamic Treatment Rules from Observational Data. *Candian Journal of Statistics*, 40(4):629–645, 2012.

Susan A. Murphy. An Experimental Design for the Development of Adaptive Treatment Strategies. *Statistics in Medicine*, 24(10), 2005a.

Susan A Murphy. An experimental design for the development of adaptive treatment strategies. 2005b.

Inbal Nahum-Shani. What is a JITAI? In *Proceedings of Workshop on Just In Time Adaptive Interventions (JITAIs)*, 2013.

Inbal Nahum-Shani, Min Qian, Daniel Almirall, William E. Pelham, Beth Gnagy, Greg Fabiano, Jim Waxmonsky, Jihnhee Yu, and Susan Murphy. Q-learning: A Data Analysis Method for Constructing Adaptive Interventions. *Psychological Methods*, 17(4):478–494, 2012.

Min Qian and Susan A. Murphy. Performance Guarantees for Individualized Treatment Rules. *Annals of Statistics*, 39(2):1180–1210, 2011.

Benjamin Rich, Erica E. M. Moodie, David A. Stephens, and Robert W. Platt. Model Checking with Residuals for g-estimation of Optimal Dynamic Treatment Regimes. *The International Journal of Biostatistics*, 6(2):12–22, 2010.

James M. Robins. *Proceedings of the Second Seattle Symposium in Biostatistics*, chapter Optimal Structural Nested Models for Optimal Sequential Decisions, pages 189–326. Springer, 2004.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

Susan M. Shortreed, Eric Laber, T. Scott Stroup, Joelle Pineau, and Susan A. Murphy. A multiple imputation strategy for sequential multiple assignment randomized trials. *Statistics in Medicine*, 33(24):4202–4214, 2014.

Pham Dinh Tao and Le Thi Hoai An. CONVEX ANALYSIS APPROACH TO D. C. PROGRAMMING: THEORY, ALGORITHMS AND APPLICATIONS. *ACTA Mathematica Vietnamica*, 22(1):289–355, 1997.

Lu Wang, Andrea Rotnitzk, Xihong Lin, Randall E. Millikan, and Peter F. Thall. Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer. *Journal of the American Statistical Association*, 107(498):493–508, 2012.

Christopher J.C.H Watkins and Peter Dayan. Q-Learning. *Machine Learning*, 8(3-4):279–292, 1992.

Jessica G. Young, Lauren E. Cain, James M. Robins, Eilis J. O'Reilly, and Miguel A. Hernan. Comparative Effectiveness of Dynamic Treatment Regimes: An Application of the Parametric G-Formula. *Statistics in Biosciences*, 3(1), 2011.

Baqun Zhang, Anastasios A. Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. Estimating Optimal Treatment Regimes from a Classification Perspective. *Stat*, 1(1):103–114, 2012a.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics*, 68(4):1010–1018, 2012b.

Baqun Zhang, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions. *Biometrika*, 100(3):681–694, 2013.

Yingqi Zhao, Donglin Zeng, A. John Rush, and Michael R. Kosorok. Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Yingqi Zhao, Donglin Zeng, Eric B. Laber, and Michael R. Kosorok. New Statistical Learning Methods for Estimating Opitmal Dynamic Treatment Regimes. *Journal of the American Statistical Association*, 2014.

Yingqi Zhao, Donglin Zeng, Eric B. Laber, Rui Song, Ming Yuan, and Michael R. Kosorok. Doubly Robust Learning for Estimating Individualized Treatment with Censored Data. *Biometrika*, 102(1), 2015.

Yufan Zhao, Donglin Zeng, Mark A. Socinski, and Michael R. Kosorok. Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer. *Biometrics*, 67(4): 1422 – 1433, 2011.