Published in final edited form as:

Biometrics. 2018 December ; 74(4): 1180–1192. doi:10.1111/biom.12894.

Optimal Two-Stage Dynamic Treatment Regimes from a Classification Perspective with Censored Survival Data

Rebecca Hager^{1,*}, Anastasios A. Tsiatis^{2,**}, and Marie Davidian^{2,***}

¹Division of Biometrics III, OB/CDER, Food and Drug Administration, Silver Spring, Maryland, U.S.A

²Department of Statistics, North Carolina State University, Raleigh, North Carolina, U.S.A

Summary

Clinicians often make multiple treatment decisions at key points over the course of a patient's disease. A dynamic treatment regime is a sequence of decision rules, each mapping a patient's observed history to the set of available, feasible treatment options at each decision point, and thus formalizes this process. An optimal regime is one leading to the most beneficial outcome on average if used to select treatment for the patient population. We propose a method for estimation of an optimal regime involving two decision points when the outcome of interest is a censored survival time, which is based on maximizing a locally efficient, doubly robust, augmented inverse probability weighted estimator for average outcome over a class of regimes. By casting this optimization as a classification problem, we exploit well-studied classification techniques such as support vector machines to characterize the class of regimes and facilitate implementation via a backward iterative algorithm. Simulation studies of performance and application of the method to data from a sequential, multiple assignment randomized clinical trial in acute leukemia are presented.

Keywords

Doubly robust; Inverse probability weighting; Optimal treatment regime; Sequential multiple assignment randomized trial; Support vector machine; Value search

1. Introduction

For patients with chronic diseases such as cancer or HIV infection, clinicians make multiple treatment decisions at key milestones over the course of the disease. A dynamic treatment regime (DTR) is a sequence of decision rules, each corresponding to a key decision point and each dictating which of the available treatment options to administer based on information on the patient to that point. DTRs, thus, formalize the process by which clinicians make treatment decisions in practice and are of obvious relevance to precision

* rebecca.hager@fda.hhs.gov. ** tsiatis@ncsu.edu. *** davidian@ncsu.edu.

Supplementary Materials

Web Appendices A–G, referenced in Sections 3–6, and R code implementing the methods are available with this article at the *Biometrics* website on Wiley Online Library.

medicine, which focuses on tailoring treatment decisions to individual patients using their characteristics. An important objective is to identify an optimal DTR, which, if used to select treatment for all patients in the population, would lead to an average health outcome that is more favorable than that achieved using any other treatment selection strategy.

Estimation of an optimal regime from data is thus of considerable interest. Using data from a longitudinal observational study can present challenges because these data are subject to time-dependent confounding, and they were not collected for the purpose of evaluating DTRs, so it is likely that key patient information was not recorded. The sequential multiple assignment randomized trial (SMART; Lavori and Dawson, 2004; Murphy, 2005), in which participants are randomized to the available, feasible treatment options at each of K pre-specified decision points and in which rich information is collected throughout, has been advocated as a gold standard data source for estimation of DTRs.

Extensive literature exists on estimation of optimal DTRs from data from SMARTs and observational studies when the outcome is observed on all participants; an abbreviated list includes Murphy (2003), Robins (2004), Zhang et al. (2012a, b, 2013), and Zhao et al. (2012, 2015a). For many chronic diseases, however, the relevant outcome is a possibly censored time to an event (survival time). For example, the primary outcome in North American Leukemia Intergroup Study C9710, coordinated by the Cancer and Leukemia Group B, now part of the Alliance for Clinical Trials in Oncology (Powell et al., 2010) in patients with acute promyelocytic leukemia was a possibly censored composite of time to failure to achieve complete remission (CR), relapse after achieving CR, or death, whichever came first. All subjects received the same induction chemotherapy and were randomized to have it followed by one of two consolidation therapies, two courses of all-trans-retinoic acid (ATRA) or two courses of ATRA combined with arsenic trioxide. Those who did not experience an event or censoring by completion of consolidation therapy were re-randomized to receive maintenance therapy of ATRA alone or ATRA plus oral methotrexate (Mtx) and mercaptopurine (MP). Thus, Study C9710 was a SMART with $K = 2$ decision points. Although the original goal was not to estimate an optimal DTR, the data are ideal for this purpose.

There is relatively little work on estimation of optimal regimes based on a censored time-to-event outcome. Bai et al. (2017) and Zhao et al. (2015b) propose methods for a single decision point ($K = 1$). Xu et al. (2016) propose a Bayesian framework for studying DTRs with $K \geq 2$ but do not consider estimation of optimal regimes. Goldberg and Kosorok (2012) adapt the method of Q-learning for multiple decisions to censored survival outcome. Q-learning can be non robust to misspecification of the models that must be posited at each decision point (Schulte, 2014); Goldberg and Kosorok also assume that censoring is independent of the event time, decision times, and patient characteristics. For $K = 1, 2$ Jiang et al. (2017) maximize an estimator of t -year survival probability for predetermined t ; however, for $K = 2$, they require the time of the second decision point to be pre-specified for all subjects, which typically it is not in a setting like that of Study C9710. Zhao et al. (2011) propose a Q-learning strategy where this and independent censoring are also assumed.

We propose a method for estimation of an optimal DTR based on a possibly censored survival outcome in the popular case $K = 2$. The estimator maximizes over all regimes in a specified class of regimes a locally efficient, doubly robust, augmented inverse probability weighted estimator for the expectation of a monotone function of the event time under a given regime in the class, referred to as the *value* of the regime. Censoring can take place at any time subsequent to the first decision and can depend on an individual's past information. The time of the second decision can vary across individuals. The method also allows the set of feasible treatment options at the second decision to vary depending on the history of the patient. By casting this optimization as a weighted classification problem (Zhang et al., 2012b; Zhao et al., 2012), we characterize the class of DTRs of interest through classification methods such as support vector machines. Implementation is via a backward recursive algorithm exploiting well-studied techniques for minimizing weighted classification error.

In Section 2, we propose a statistical framework in which the value of a given two-decision DTR is defined and present the locally efficient estimator for its value in Section 3. In Section 4, we describe the backward algorithm for estimation of an optimal regime. We present simulation studies in Section 5, and an analysis of Study C9710 in Section 6.

2. Statistical Framework

We begin by discussing the notion of a DTR for $K = 2$ decision points. Let x_1 denote baseline patient information (covariates) available at Decision 1, so that the patient's "history" at this decision is $h_1 = \{x_1\}$, and let $\mathcal{A}_1$ be set of possible treatment options. If option the $a_1 \in \mathcal{A}_1$ is administered at Decision 1 and the patient survives to Decision 2, the history h_2 available at Decision 2 includes h_1 , a_1 , and x_2 , additional covariates accruing between Decisions 1 and 2. Let $\mathcal{A}_2$ be the set of available treatment options at the second decision point. A DTR, $d = \{d_1(h_1), d_2(h_2)\}$, comprises rules corresponding to each decision point, where, for Decision k , rule d_k maps a patient's history h_k to the set of treatment options $\mathcal{A}_k$, $k = 1, 2$. Denote the class of all possible treatment regimes by $\mathcal{D}$. As in most of the literature on DTRs, we take $\mathcal{A}_k$, $k = 1, 2$, to comprise finite numbers of treatment options.

A useful statistical framework must accommodate the features noted in Section 1, which arise in many practical settings: patients may experience the event or be censored prior to Decision 2, so do not receive Decision 2 treatment; the time of the second decision is not prespecified and can vary across patients; and, depending on their histories, patients may be eligible for only a subset of the options in $\mathcal{A}_2$. Accordingly, the framework we now present is more complex than that in much of the DTR literature and requires additional notation.

We first establish a potential outcomes formulation (e.g., Zhang et al., 2012a, 2013) through which we define precisely the value of a fixed regime $d \in \mathcal{D}$; by "fixed," we mean a particular, given regime in $\mathcal{D}$. The value is the expected outcome if all patients in the population were to receive treatment at Decisions 1 and 2 according to the rules in d .

Consider a randomly chosen patient with history $H_1 = \{X_1\}$ at Decision 1, for baseline covariates X_1 . Define the first stage of treatment to be the period from administration of a treatment option at Decision 1 to occurrence of the event or to the time of the second decision, whichever comes first, and let $S^*(a_1)$ be the length of the first stage that would be achieved if the patient were to receive option $a_1 \in \mathcal{A}_1$ at Decision 1. Let $R^*(a_1)$ indicate his status at the end of stage 1: if $R^*(a_1) = 0$, $S^*(a_1)$ is the time to the event; if $R^*(a_1) > 0$, the event has not yet occurred, and $S^*(a_1)$ is the time to Decision 2 (start of stage 2 treatment). In the latter case, $R^*(a_1)$ takes a value r in $\{1, \dots, B\}$ indicating the subset of treatment options $\mathcal{A}_{2r} \subseteq \mathcal{A}_2$, say, he would be eligible to receive. When $R^*(a_1) > 0$, let $X_{2R^*(a_1)}^*(a_1)$

denote a vector of covariates that would arise between Decisions 1 and 2 for a patient who received a_1 at Decision 1, where the dimension and elements of $X_{2R^*(a_1)}^*(a_1)$ are determined by $R^*(a_1)$.

As an example, suppose that Decision 1 involves selection of induction therapy for leukemia patients, and Decision 2 involves selection of a maintenance treatment for those who show a specified response to induction therapy and of a salvage therapy for those who do not, so that $\mathcal{A}_2$ comprises both types of options. For a patient who receives induction option a_1 and would survive to have her response status ascertained, so for whom $R^*(a_1) > 0$, $R^*(a_1) = 1$ indicates response and thus eligibility to receive one of the maintenance options in the subset of such options $\mathcal{A}_{21}$, and $R^*(a_1) = 2$ indicates nonresponse and eligibility for one of the salvage therapies in the subset of such options $\mathcal{A}_{22}$. Here, $B = 2$. If $R^*(a_1) = 1$, $X_{2R^*(a_1)}^*(a_1)$

might include information collected on all patients receiving a_1 prior to ascertaining response status, some of which may be used to define response status, plus additional variables collected only on responders to inform selection of maintenance therapy, and similarly for $R^*(a_1) = 2$.

Define $T^*(a_1, a_2)$ to be the event time that the patient would achieve if he were to then to receive option $a_2 \in \mathcal{A}_2$ at Decision 2. For a patient for whom $R^*(a_1) = 0$, who would thus not receive Decision 2 treatment, $S^*(a_1)$ is his event time, so define $T^*(a_1, a_2) = S^*(a_1)$. Denote the set of all potential outcomes as

$$W^* = [S^*(a_1), R^*(a_1), I\{R^*(a_1) > 0\} X_{2R^*(a_1)}^*(a_1), T^*(a_1, a_2) \text{ for all } (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2], \text{ where } I(B)$$

$= 1$ if event B is true and 0 otherwise. For any $d \in \mathcal{D}$, the event time $T^*(d) = T^*(d_1, d_2)$ that would be achieved if a randomly chosen patient were to be treated by following the rules in d is then defined in terms of H_1 and W^* as

$$T^*(d) = \sum_{(a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2} (I\{d_1(H_1) = a_1, R^*(a_1) = 0\} S^*(a_1) + I\{d_1(H_1) = a_1, R^*(a_1) > 0, d_2\{H_2^*(a_1)\} = a_2\} T^*(a_1, a_2))$$

, where $H_2^*(a_1) = [H_1, a_1, S^*(a_1), R^*(a_1), I\{R^*(a_1) > 0\}X_{2R^*(a_1)}^*(a_1)]$. The potential outcomes framework is summarized in Table 1.

Often, a monotone function $g\{T^*(d)\}$ of event time is of interest, for example, restricted survival time, where $g(t) = \min(t, L)$ for a fixed cutoff time L ; survival status at a fixed time L , $g(t) = I(t \leq L)$; or event time itself, $g(t) = t$. It follows that the expected, possibly transformed event time if all patients in the population were to receive treatment according to the two-decision regime d is $\mathcal{V}_2(d) = E[g\{T^*(d)\}]$ for specified $g(\cdot)$; that is, the value of d .

This development formally characterizes, for specified $g(\cdot)$, the value $\mathcal{V}_2(d)$ of any fixed regime $d \in \mathcal{D}$ for $K = 2$ in terms of potential outcomes. In Section 3, we derive an estimator for $\mathcal{V}_2(d)$ based on the data that would be available from an observational study or clinical trial such as Study C9710, which is used to develop the proposed estimator for an optimal regime in Section 4. We first describe these data in the ideal case of no censoring.

For a randomly chosen patient, with H_1 as above, let A_1 be the option in $\mathcal{A}_1$ actually administered at Decision 1. Let S be the time to event or second decision, whichever occurs first; if $g(t) = \min(t, L)$, S is bounded above by L . Let R denote observed status, where $R = 0$ indicates that S is the event time, and $R > 0$ indicates that S is the time to the second decision, with $R = r$, $r = 1, \dots, B$, indicating the subset of options $\mathcal{A}_{2r} \subseteq \mathcal{A}_2$ the patient was eligible to receive. For $R > 0$, X_{2R} denotes covariates ascertained between Decisions 1 and 2, where the dimension of X_{2R} is determined by R ; A_2 denotes the second stage treatment in $\mathcal{A}_2$ actually received, and T is the event time. The observable data are then $\{H_1, A_1, S, R, I(R > 0)X_{2R}, I(R > 0)A_2, T\}$, where $T = S$ if $R = 0$, and are summarized in Table 1. The data observed up to time $u > 0$ are

$$H(u) = \{H_1, A_1, I(S \leq u), I(S \leq u)S, I(S \leq u)R, I(S \leq u, R > 0)X_{2R}, I(S \leq u, R > 0)A_2, I(T \leq u), I(T \leq u)T\}. \quad (1)$$

Now suppose that censoring can occur at any time $u > 0$, and let C denote time to censoring. We assume that the hazard function for censoring at time u , conditional on (1) and the potential outcomes, satisfies

$$\begin{aligned} \alpha\left\{u \middle| H(u), W^*\right\} &= \lim_{du \rightarrow 0} du^{-1} \Pr\left\{u \leq C < u + du \middle| C \geq u, H(u), W^*\right\} \quad (2) \\ &= [I(S > u)\alpha_1(u|H_1, A_1) + I(S \leq u, R > 0) \\ &\quad \alpha_2(u|H_1, A_1, S, R, X_{2R}, A_2)]I(T > u) \\ &= \alpha\left\{u \middle| H(u)\right\}, \end{aligned}$$

where $\alpha_1(u | h_1, \alpha_1)$ and $\alpha_2(u | h_1, \alpha_1, s, r, x_{2R}, \alpha_2)$ are positive functions of u and the indicated information. Under (2), censoring is noninformative in that the hazard at u depends only on information observed through time u and not on unobserved potential outcomes.

Let $V = \min(S, C)$ denote the observed length of the first stage, which now could be curtailed by censoring, and define $\Gamma = I(S < C)$, the indicator that censoring does not occur before S (the event time or time to Decision 2). If $g(t) = \min(t, L)$, redefine $V = \min(S, L, C)$ and $\Gamma = I(\min(S, L) < C)$. The observed status of the patient at the end of stage 1 is then ΓR , where R is as above. If $\Gamma R = 0$, V is either the time to event, $\Gamma = 1$ and $R = 0$, or time to censoring, $\Gamma = 0$; if $\Gamma R > 0$, $\Gamma = 1$ and $R > 0$, so that V is the time to second stage treatment, $\Gamma R = R$, and A_2 and X_{2R} are defined as above. Let $H_2 = \{H_1, A_1, V, \Gamma, \Gamma R, \Gamma I(R > 0) X_{2R}\}$ denote the observed history for a patient through the end of the first stage. Define $U = \min(T, C)$ and $\Delta = I(T < C)$; if $g(t) = \min(t, L)$, $U = \min(T, L, C)$, and $\Delta = I(\min(T, L) < C)$. If $\Gamma R = 0$, $V = U$, and either $\Gamma = 1$ or $\Gamma = 0$. If $\Gamma R > 0$, $U > V$, and $U - V$ is the possibly censored “additional life” after second stage treatment. These definitions are summarized in Table 1. The observed data on a sample of n individuals are then independent and identically distributed (i.i.d.)

$$\left\{ H_{1i}, A_{1i}, V_i, \Gamma_i, \Gamma_i R_i, \Gamma_i I(R_i > 0) X_{2R_i}, \Gamma_i I(R_i > 0) A_{2i}, U_i, \Delta_i \right\}, \quad (3)$$

$$i = 1, \dots, n.$$

3. Value Estimators for a Fixed Regime

We now present estimators for $\mathcal{V}_2(d) = E[g\{T^*(d)\}]$ for fixed $d \in \mathcal{D}$ based on the observed data (3). Several assumptions are required under which estimation of $\mathcal{V}_2(d)$ is possible from (3) (see Schulte, 2014, Section 2). The consistency assumption states that $R = R^*(A_1)$, $S = S^*(A_1)$, $X_{2R} = X_{2R^*(A_1)}^*(A_1)$, and $T = S^*(A_1)I(R = 0) + T^*(A_1, A_2)I(R > 0)$, so that possibly

censored, observable quantities are equal to the corresponding potential versions under the treatments actually received. The stable unit treatment value assumption (SUTVA) ensures that a patient’s outcome and covariates are not affected by how treatments are assigned to him and to others. Under the sequential randomization assumption (SRA), treatment at each stage is independent of all potential outcomes, conditional on history. Letting “ $\perp\!\!\!\perp$ ” denote “independent of,” the SRA is $A_1 \perp\!\!\!\perp W^* | H_1$ and $A_2 \perp\!\!\!\perp W^* / H_2, \Gamma R > 0$, where the latter condition is relevant only if stage two treatment is observed. The SRA holds in a SMART but is unverifiable from observational data. Validity of these assumptions must be evaluated critically based on the subject matter.

The positivity assumption ensures that all treatment options at each decision are represented in the data and in our setting involves several conditions. At Decision 1, we require $\pi_1(h_1, a_1) = \text{pr}(A_1 = a_1 | H_1 = h_1) > 0$ for all h_1 for which $\text{pr}(H_1 = h_1) > 0$ and $a_1 \in \mathcal{A}_1$. Using (2), we must have for all h_1, a_1, s for which $\text{pr}(H_1 = h_1, A_1 = a_1, S = s) > 0$ that

$$\begin{aligned} \text{pr}(C > s \mid H_1 = h_1, A_1 = a_1, S = s) & \quad (4) \\ = \exp\left\{-\int_0^s \alpha_1(u \mid h_1, a_1) du\right\} & > 0, \end{aligned}$$

so that there is positive probability of not being censored before the second decision. If a patient reaches the second stage without being censored, we need for $r = 1, \dots, B$ that $\text{pr}(A_2 = a_2 \mid H_1 = h_1, A_1 = a_1, S = s, R = r, \Gamma R > 0, X_{2R} = x_{2r}) > 0$ for $a_2 \in \mathcal{A}_{2r}$ for all h_1, a_1, s, r, x_{2r} for which $\text{pr}(H_1 = h_1, A_1 = a_1, S = s, R = r, \Gamma R > 0, X_{2R} = x_{2r}) > 0$. Finally, for patients reaching the second stage, we require a positive probability of not being censored before the event time, namely, using (2), for all $h_1, a_1, s, r, x_{2r}, a_2, t$ for which $\text{pr}(H_1 = h_1, A_1 = a_1, S = s, R = r, \Gamma R > 0, X_{2R} = x_{2r}, A_2 = a_2, T = t) > 0$,

$$\begin{aligned} \text{pr}(C > t \mid H_1 = h_1, A_1 = a_1, S = s, R = r, \Gamma R > 0, X_{2R} = x_{2r}, & \quad (5) \\ A_2 = a_2, T = t) = \exp\left\{-\int_s^t \alpha_2(u \mid h_1, a_1, s, r, x_{2r}, a_2) du\right\} & > 0, \end{aligned}$$

where in (4)–(5) the limits of integration are $\min(s, L)$ and $\min(t, L)$ when $g(t) = \min(t, L)$.

We now develop estimators for $\mathcal{V}_2(d) = E[g\{T^*(d)\}]$, $d = \{d_1(h_1), d_2(h_2)\}$, based on the data (3) under these assumptions. Let $\tilde{C}_{d,1} = I\{A_1 = d_1(H_1)\}$ be the indicator that a patient receives treatment consistent with d_1 ; under the assumptions,

$$\text{pr}(\tilde{C}_{d,1} = 1 \mid H_1, W^*) = \text{pr}(\tilde{C}_{d,1} = 1 \mid H_1) = \pi_{\tilde{C}_{d,1}}(H_1), \text{ say, where } \pi_{\tilde{C}_{d,1}}(H_1) = \pi_1\{H_1, d_1(H_1)\},$$

and the stage 1 propensity score $\pi_1(h_1, a_1)$ is defined above. Because the treatment options for which a patient with $R > 0$ is eligible at the second stage depend on R , for h_2 involving a particular $r = 1, \dots, B$, $d_2(h_2)$ reduces to a rule $d_{2r}(h_2)$ taking values in $\mathcal{A}_{2r}$. Thus, define

$$\pi_{d_{2r}}(h_2) = \text{pr}(A_2 = d_{2r}(h_2) \mid H_2, \Gamma = 1, R = r), \quad r = 1, \dots, B, \text{ the propensity for stage 2 treatment to be consistent with that selected by } d_2 \text{ when } R = r, \text{ and}$$

$$\tilde{C}_{d,2} = I\{A_1 = d_1(H_1), \Gamma = 1, R > 0, A_2 = d_2(H_2)\},$$

the indicator that a patient does not experience the event and is not censored before Decision 2 and receives treatment options consistent with those dictated by both rules in d . From (2), $\text{pr}\left\{C > u \mid H(u), W^*\right\} = \text{pr}\left\{C > u \mid H(u)\right\} = \mathcal{K}\left\{u \mid H(u)\right\} = \exp\left\{-\int_0^{\min(u, T)} \alpha\{w \mid H(w)\} dw\right\}$ is the survival function for censoring. Under the above assumptions, it can be shown that

$$\begin{aligned} \text{pr}(\tilde{C}_{d,2} = 1 \mid H_2, W^*) &= \text{pr}(\tilde{C}_{d,2} = 1 \mid H_2) = \pi_{\tilde{C}_{d,1}}(H_1) \mathcal{K}\left\{V \mid H(V)\right\} \\ &\times \left\{ \sum_{r=1}^B I(R=r) \pi_{d_{2r}}(H_2) \right\} I(\Gamma R > 0) = \pi_{\tilde{C}_{d,2}}(H_2), \text{ say.} \end{aligned} \quad (6)$$

For patients whose treatments received are consistent with those selected by d and whose event time is not censored, $T = T^*(d)$ is observed, but $T^*(d)$ is “missing” for all others. From semiparametric theory for monotone missing data (Tsiatis, 2006), it can be derived that the locally efficient estimator for $\mathcal{V}_2(d)$ is of the form

$$n^{-1} \sum_{i=1}^n \left\{ \Psi_{\text{IPW},i}^{(2)}(d) + \Psi_{\text{Aug1},i}^{(2)}(d) + \Psi_{\text{Aug2},i}^{(2)}(d) + \Psi_{\text{CAug1},i}^{(2)}(d) + \Psi_{\text{CAug2},i}^{(2)}(d) \right\}, \quad (7)$$

$$\begin{aligned} \Psi_{\text{IPW},i}^{(2)}(d) &= I(R_i = 0) \frac{\Gamma_i}{\mathcal{K}\{U_i \mid H_i(U_i)\}} \frac{\tilde{C}_{d,1,i}}{\pi_{\tilde{C}_{d,1}}(H_{1i})} g(U_i) \\ &+ \sum_{r=1}^B I(R_i = r) \frac{\Delta_i}{\mathcal{K}\{U_i \mid H_i(U_i)\} \mathcal{K}\{V_i \mid H_i(V_i)\}} \frac{\tilde{C}_{d,2,i}}{\pi_{\tilde{C}_{d,2}}(H_{2i})} g(U_i), \end{aligned} \quad (8)$$

$$\begin{aligned} \Psi_{\text{Aug1},i}^{(2)}(d) &= - \sum_{r=1}^B \left[I(R_i = r) \frac{\Gamma_i}{\mathcal{K}\{V_i \mid H_i(V_i)\}} \frac{\tilde{C}_{d,1,i}}{\pi_{\tilde{C}_{d,1}}(H_{1i})} \right. \\ &\times \frac{I\{A_{2i} = d_{2r}(H_{2i})\} - \pi_{d_{2r}}(H_{2i})}{\pi_{d_{2r}}(H_{2i})} \\ &\times E\{g(T_i) \mid H_{2i}, A_{2i} = d_{2r}(H_{2i}), R_i = r\}], \end{aligned} \quad (9)$$

$$\Psi_{\text{Aug2},i}^{(2)}(d) = - \frac{\tilde{C}_{d,1,i} - \pi_{\tilde{C}_{d,1}}(H_{1i})}{\pi_{\tilde{C}_{d,1}}(H_{1i})} E\left[g\left(T_i^*(d)\right) \mid H_{1i}\right], \quad (10)$$

$$\begin{aligned} \Psi_{\text{CAug1},i}^{(2)}(d) &= \sum_{r=1}^B \left[I(R_i = r) \frac{\tilde{C}_{d,2,i}}{\pi_{\tilde{C}_{d,2}}(H_{2i})} \right. \\ &\quad \times \int_{V_i}^{\infty} \frac{dM_i\{u|H_i(u)\}}{\mathcal{K}\{u|H_i(u)\} / \mathcal{K}\{V_i|H_i(V_i)\}} \\ &\quad \times E\{g(T_i) | T_i \geq u, H_{2i}, A_{2i} = d_{2r}(H_{2i}), R_i = r\} \Big], \end{aligned} \quad (11)$$

$$\Psi_{\text{CAug2},i}^{(2)}(d) = \frac{\tilde{C}_{d,1,i}}{\pi_{\tilde{C}_{d,1}}(H_{1i})} \int_0^{V_i} \frac{dM_i\{u|H_i(u)\}}{\mathcal{K}\{u|H_i(u)\}} E\left[g\left(T_i^*(d)\right) \middle| T_i^*(d) \geq u, H_{1i}\right]; \quad (12)$$

where the martingale increment $dM_i\{u|H_i(u)\} = dN_i(u) - \alpha\{u|H_i(u)\}Y_i(u)du$, with $N_i(u) = \mathbb{I}(U_i \leq u, \tau_i = 0)$; and $Y_i(u) = \mathbb{I}(U_i \leq u)$.

The first term $\Psi_{\text{IPW},i}^{(2)}(d)$, is nonzero only for patients whose event times are not censored and whose observed treatments coincide with those dictated by d . This condition is satisfied for patients who do not receive second stage treatment but are consistent with d_1 , for whom $S = V = U = T$ is observed. Their transformed event times in (7) are inversely weighted by the probability of not being censored and having stage 1 treatment consistent with d_1 given history. It is also satisfied for patients who receive second stage treatment and are consistent with both d_1 and d_2 , so that $T = U$ is observed. Their transformed event times are inversely weighted in (8) by the probability of this occurring, conditional on history. It can be shown under the above assumptions that $E\{\Psi_{\text{IPW}}^{(2)}(d)\} = \mathcal{T}_2(d)$.

The mean zero “augmentation” terms (9)–(12) recover information from other patients: $\Psi_{\text{Aug1},i}^{(2)}(d)$ and $\Psi_{\text{Aug2},i}^{(2)}(d)$ from patients whose treatments received are inconsistent with d , the former with first stage treatment consistent with d_1 but second stage treatment not consistent with d_2 , and the latter with both treatments inconsistent with d ; $\Psi_{\text{CAug1},i}^{(2)}(d)$ from patients whose first and second stage treatments are consistent with d but whose event time was then censored; and $\Psi_{\text{CAug2},i}^{(2)}(d)$ from patients whose first stage treatment is consistent with d_1 but whose event time was censored before second stage treatment.

To implement the estimator, one can posit, fit, and substitute models for the quantities in (7)–(12). The propensity score $\pi_1(h_1, a_1)$ is known in a clinical trial and can be estimated directly; in an observational study, $\pi_1(h_1, a_1)$ can be represented by a multinomial polytomous regression model (e.g., a logistic model when $\mathcal{A}_1$ comprises two options) and fitted via maximum likelihood. Substituting the fitted model $\hat{\pi}_1(h_1, a_1)$ yields the estimator

for $\pi_{\tilde{C}_{d,1}}(H_1)$ given by $\hat{\pi}_{\tilde{C}_{d,1}}(H_1) = \hat{\pi}_1\{H_1, d_1, (H_1)\}$. For each $r = 1, \dots, B$, the propensity $\pi_{d_{2r}}(h_2)$ is known in a SMART and can be estimated directly or modeled via multinomial regression and fitted by maximum likelihood, yielding $\hat{\pi}_{d_{2r}}(h_2)$. From (2), an estimator $\hat{\alpha}\{u|H(u)\}$ can be obtained by fitting proportional hazards models for $\alpha_1(u|H_1, A_1)$ and $\alpha_2(u|H_1, A_1, S, R, X_{2R}, A_2)$; substituting these in $\mathcal{K}\{u|H(u)\}$ and $dM\{u|H(u)\}$ yields estimators $\hat{\mathcal{K}}\{u|H(u)\}$ and $d\hat{M}\{u|H(u)\}$. An estimator $\hat{\pi}_{\tilde{C}_{d,2}}(H_2)$ can then be obtained by substitution of $\hat{\pi}_{\tilde{C}_{d,1}}(H_1)$, $\hat{\mathcal{K}}\{V|H(V)\}$ and $\hat{\pi}_{d_{2r}}(H_2)$, $r = 1, \dots, B$, in (6). The expectations in (9) and (11) can be estimated by fitted proportional hazards models for the “additional life” hazard functions $\lambda_{2r}(u|H_2, A_2) = \lim_{du \rightarrow 0} du^{-1} \Pr(u - V < u + du, R = 1 | U - V \leq u, H_2, A_2, R = r)$, $r = 1, \dots, B$, and substituting the fitted models $\hat{\lambda}_{2r}(u|H_2, A_2)$ to obtain the estimator

$$\hat{\mathcal{S}}_{2r}\left\{u \middle| H_2, A_2 = d_{2r}(H_2)\right\} = \exp\left[-\int_0^u \hat{\lambda}_{2r}\{w|H_2, A_2 = d_{2r}(H_2)\}dw\right] \text{ for}$$

$$\hat{\mathcal{S}}_{2r}(u|H_2, A_2) = P(T > u|H_2, A_2) \text{ and}$$

$$\begin{aligned} \hat{E}\left\{g(T) \middle| H_2, A_2 = d_{2r}(H_2), R = r\right\} &= \\ \int_0^\infty g(V+w) \left[-d\hat{\mathcal{S}}_{2r}\{w|H_2, A_2 = d_{2r}(H_2)\}\right], \\ \hat{E}\left\{g(T) \middle| T \geq u, H_2, A_2 = d_{2r}(H_2), R = r\right\} &= \\ \int_{u-V}^\infty g(V+w) \left[\frac{-d\hat{\mathcal{S}}_{2r}\{w|H_2, A_2 = d_{2r}(H_2)\}}{\hat{\mathcal{S}}_{2r}\{u|H_2, A_2 = d_{2r}(H_2)\}}\right]. \end{aligned}$$

Estimation of the expectations in (10) and (12) is more involved and is discussed in detail in Web Appendix A. As demonstrated there, this requires fitting models for the cause-specific hazard functions for V , $\lambda_{1r}(u|H_1, A_1) = \lim_{du \rightarrow 0} du^{-1} \Pr(u - V < u + du, \Gamma = 1, R = r | V \leq u, H_1, A_1)$, $r = 0, 1, \dots, B$, for example, proportional hazards models; and models for

$$\begin{aligned} Q_{1r}\{H_1, A_1 = d_1(H_1; \eta_1), S, R = r\} &= E[E\{g(T)|H_2, \\ A_2 = d_{2r}(H_2), R = r\} | H_1, A_1, S, R = r], & r = 1, \dots, B. \end{aligned} \quad (13)$$

Substitution of the foregoing fitted models in (7)–(12) leads to estimators $\hat{\Psi}_{\text{IPW}, i}^{(2)}(d)$, $\hat{\Psi}_{\text{Aug1}, i}^{(2)}(d)$, $\hat{\Psi}_{\text{Aug2}, i}^{(2)}(d)$, $\hat{\Psi}_{\text{CAug1}, i}^{(2)}(d)$, and $\hat{\Psi}_{\text{CAug2}, i}^{(2)}(d)$ and the estimator for $\mathcal{V}_2(d)$ given by

$$\hat{\mathcal{V}}_2^{\text{CAIPW}}(d) = n^{-1} \sum_{i=1}^n \left\{ \hat{\Psi}_{\text{IPW},i}^{(2)}(d) + \hat{\Psi}_{\text{Aug1},i}^{(2)}(d) + \hat{\Psi}_{\text{Aug2},i}^{(2)}(d) + \hat{\Psi}_{\text{CAug1},i}^{(2)}(d) + \hat{\Psi}_{\text{CAug2},i}^{(2)}(d) \right\}$$

(14)

This estimator is consistent for $\mathcal{V}_2(d)$ if either (i) the models for the propensities $\pi_1(h_1)$ and $\pi_{d_{2r}}(h_2)$, $r = 1, \dots, B$, and the censoring hazard model are correctly specified or (ii) the cause-specific and additional life hazard models and models in (13) are correctly specified. In a SMART, $\pi_1(h_1, a_1)$ and $\pi_{d_{2r}}(h_2)$ are known by design, so when the SMART involves only administrative (end of study) censoring, $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d)$ is a consistent estimator for $\mathcal{V}_2(d)$.

As above, $E\{\Psi_{\text{IPW}}^{(2)}(d)\} = \mathcal{V}_2(d)$, which suggests that, if the models for $\pi_1(h_1, a_1)$; $\pi_{d_{2r}}(h_2)$, $r = 1, \dots, B$; and $\alpha\{u|H(u)\}$ are correctly specified,

$$\hat{\mathcal{V}}_2^{\text{IPW}}(d) n^{-1} \sum_{i=1}^n \hat{\Psi}_{\text{Aug1},i}^{(2)}(d) \quad (15)$$

is a consistent estimator for $\mathcal{V}_2(d)$, but could be inconsistent if these models are misspecified. An estimator that does not involve the terms that recover information from censored patients,

$$\hat{\mathcal{V}}_2^{\text{AIPW}}(d) = n^{-1} \sum_{i=1}^n \left\{ \hat{\Psi}_{\text{IPW},i}^{(2)}(d) + \hat{\Psi}_{\text{Aug1},i}^{(2)}(d) + \hat{\Psi}_{\text{Aug2},i}^{(2)}(d) \right\}, \quad (16)$$

is also doubly robust in that, with the censoring hazard function correctly specified, it is a consistent for $\mathcal{V}_2(d)$ if (i) the propensity models or (ii) the remaining models are correctly specified. Relative to $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d)$ in (14), (15), and (16) sacrifice some efficiency and, in the case of (15), protection against inconsistency afforded by the double robustness property in favor of simplicity. In the case of no censoring, the estimator found by setting $\Gamma_i = 1$, $i = 1, \dots, n$, and $\mathcal{K}\{u|H(u)\} \equiv 1$ for all $u > 0$ in (16) is the locally efficient estimator.

A valid, approximate large sample variance estimator for $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d)$ is straightforward; consistent variance estimators for $\hat{\mathcal{V}}_2^{\text{IPW}}(d)$ and $\hat{\mathcal{V}}_2^{\text{AIPW}}(d)$ are considerably more complicated; see Web Appendix B. Alternatively, a nonparametric bootstrap can be used.

4. Estimation of an Optimal Regime

4.1. Value Search Estimation

An optimal two-decision regime, d^{opt} , say, is one that maximizes the value $\mathcal{V}_2(d)$ over $d \in \mathcal{D}$. In principle, given an estimator $\hat{\mathcal{V}}_2(d)$ for $\mathcal{V}_2(d)$, such as $\hat{\mathcal{V}}_2^{\text{IPW}}(d)$, $\hat{\mathcal{V}}_2^{\text{AIPW}}(d)$, or $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d)$, an estimator $\hat{d}^{\text{opt}}$ can be obtained by maximizing $\hat{\mathcal{V}}_2(d)$ over all $d \in \mathcal{D}$.

However, searching over all possible regimes in $\mathcal{D}$ is an almost certainly insurmountable challenge.

Accordingly, a popular approach (Zhang et al., 2012a, b, 2013; Zhao et al., 2012, 2015a, b) is to restrict attention to a class of regimes $\mathcal{D}_\eta \subset \mathcal{D}$, indexed by finite-dimensional parameter η , so comprising regimes of the form $d_\eta = \{d_1(h_1; \eta_1), d_2(h_2; \eta_2)\} = \{d_{1, \eta_1}(h_1), d_{2, \eta_2}(h_2)\}$.

Here, $d_2(h_2; \eta_2)$ reduces to a rule $d_{2r}(h_2; \eta_{2r}) = d_{2r, \eta_{2r}}(h_2)$, depending on r in $h_2 \eta = (\eta_1^T, \eta_2^T)^T$, $\eta_2 = (\eta_{21}^T, \dots, \eta_{2B}^T)^T$. We take η_{2r} to be nonoverlapping across $r = 1, \dots, B$. The restricted class $\mathcal{D}_\eta$ can be chosen on the basis of interpretability or ease of implementation or, as we discuss below, may be dictated by analogy to a classification problem. An optimal regime in $\mathcal{D}_\eta$ is denoted $d_{\eta^{\text{opt}}}$, where $\eta^{\text{opt}} = \arg \max_\eta \mathcal{V}_2(d_\eta)$. Although a globally optimal regime d^{opt} may or may not be contained in $\mathcal{D}_\eta$, $d_{\eta^{\text{opt}}}$ may be of independent interest.

It is natural to estimate η^{opt} by $\hat{\eta}^{\text{opt}} = \arg \max_\eta \hat{\mathcal{V}}_2(d_\eta)$ and thus an optimal restricted regime $d_{\eta^{\text{opt}}}$ by $\hat{d}^{\text{opt}}$, referred to as value search or policy search estimation. However, $\hat{\mathcal{V}}_2^{\text{IPW}}(d_\eta)$, $\hat{\mathcal{V}}_2^{\text{AIPW}}(d_\eta)$, and $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d_\eta)$ are nonsmooth, nonconvex functions in η , to which standard optimization techniques do not apply. Zhang et al. (2012a, 2013) use a genetic algorithm or grid search to maximize value estimators when η is low dimensional, but this is computationally prohibitive as the dimension increases. Thus, although conceptually straightforward, direct value search estimation may not be feasible in practice.

4.2. Estimation via Backward Iterative Algorithm

To circumvent this difficulty, we propose an backward iterative scheme for estimation of $d_{\eta^{\text{opt}}} \in \mathcal{D}_\eta$ that is in the same spirit as those given by Zhao et al. (2015a) and Tao and Wang (2017). The approach involves successive maximization of value estimators in first η_2 and then η_1 , thereby separating the maximization problem into two steps of lower dimension.

Consider Decision 2. Only patients who have not already experienced the event or been censored, so for whom $\Gamma R > 0$, are eligible to receive second stage treatment. Such a patient has history H_2 already established at Decision 2, so that selection of stage two treatment effectively is a single-decision problem in which a patient presents with “baseline” history H_2 at time V . This suggests that estimation of an optimal rule at Decision 2 for a two-decision regime can be likened to that of an optimal rule for a single-decision regime. Bai et al. (2017) present estimators for the value of a fixed, single-decision regime, reviewed in Web Appendix C. For a patient with $R > 0$, let H_2 play the role of “baseline” history, A_2 the role of single-decision treatment received, d_{2,η_2} the role of the rule at the single decision point, and V the role of time “0.” Then, by analogy to the single-decision, locally efficient estimator of Bai et al. (2017), an estimator for the “value” associated with d_{2,η_2} when first stage treatment is already determined and recorded in H_2 is

$$n^{-1} \sum_{i=1}^n \Gamma_i \left[\sum_{r=1}^B I(R_i = r) \left\{ \Psi_{\text{IPW},i}^{(1r)}(d_{2r}, \eta_{2r}) + \Psi_{\text{Aug},i}^{(1r)}(d_{2r}, \eta_{2r}) + \Psi_{\text{CAug},i}^{(1r)}(d_{2r}, \eta_{2r}) \right\} \right] \quad (17)$$

$$\Psi_{\text{IPW},i}^{(1r)}(d_{\eta, 2r}) = \frac{\Delta_i}{\mathcal{K}_i^* \{U_i | H_i(U_i)\}} \frac{I\{A_{2i} = d_{2r}(H_{2i}; \eta_{2r})\}}{\pi_{d_{2r}}(H_{2i}; \eta_{2r})} g(U_i), \quad (18)$$

$$\begin{aligned} \Psi_{\text{Aug},i}^{(1r)}(d_{\eta, 2r}) = & - \frac{I\{A_{2i} = d_{2r}(H_{2i}; \eta_{2r})\} - \pi_{d_{2r}}(H_{2i}; \eta_{2r})}{\pi_{d_{2r}}(H_{2i}; \eta_{2r})} \quad (19) \\ & \times E \left[g(T_i) \middle| H_{2i}, A_{2i} = d_{2r}(H_{2i}; \eta_{2r}), R_i = r \right], \end{aligned}$$

$$\begin{aligned} \Psi_{\text{CAug},i}^{(1r)}(d_{\eta, 2r}) = & \frac{I\{A_{2i} = d_{2r}(H_{2i}; \eta_{2r})\}}{\pi_{d_{2r}}(H_{2i}; \eta_{2r})} \int_{V_i}^{\infty} \frac{dM_i\{u | H_i(u)\}}{\mathcal{K}_i^* \{u | H_i(u)\}} \quad (20) \\ & \times E \left[g(T) \middle| T \geq u, H_{2i}, A_{2i} = d_{2r}(H_{2i}; \eta_{2r}), R_i = r \right], \end{aligned}$$

$\mathcal{K}_i^* \{u | H(u)\} = \mathcal{K}\{u | H(u)\} / \mathcal{K}\{V_i | H(V_i)\}$, and

$\pi_{d_{2r}}(h_2; \eta_{2r}) = \text{pr}\{A_2 = d_{2r}(h_2; \eta_{2r}) | H_2 = h_2, \Gamma = 1, R = r\}$. See Web Appendix D for the

rationale for (17). Estimators $\hat{\Psi}_{IPW,i}^{(1r)}(d_{2r}, \eta_{2r})$, $\hat{\Psi}_{Aug,i}^{(1r)}(d_{2r}, \eta_{2r})$, and $\hat{\Psi}_{CAug,i}^{(1r)}(d_{2r}, \eta_{2r})$ can be obtained by substituting fitted models in (18)–(20). With this substitution, (17) can be written as $\sum_{r=1}^B \hat{\mathcal{V}}_{1r}^{CAIPW}(d_{2r}, \eta_{2r})$, where

$$\hat{\mathcal{V}}_{1r}^{CAIPW}(d_{2r}, \eta_{2r}) = n^{-1} \sum_{i=1}^n \Gamma_i I(R_i = r) \left\{ \hat{\Psi}_{IPW,i}^{(1r)}(d_{2r}, \eta_{2r}) + \hat{\Psi}_{Aug,i}^{(1r)}(d_{2r}, \eta_{2r}) + \hat{\Psi}_{CAug,i}^{(1r)}(d_{2r}, \eta_{2r}) \right\}.$$

By analogy to (15) and (16), we can similarly define “value” estimators

$\sum_{r=1}^B \hat{\mathcal{V}}_{1r}^{IPW}(d_{2r}, \eta_{2r})$ and $\sum_{r=1}^B \hat{\mathcal{V}}_{1r}^{AIPW}(d_{2r}, \eta_{2r})$ where

$\hat{\mathcal{V}}_{1r}^{IPW}(d_{2r}, \eta_{2r}) = n^{-1} \sum_{i=1}^n \Gamma_i I(R_i = r) \hat{\Psi}_{IPW,i}^{(1r)}(d_{2r}, \eta_{2r})$ and

$\hat{\mathcal{V}}_{1r}^{AIPW}(d_{2r}, \eta_{2r}) = n^{-1} \sum_{i=1}^n \Gamma_i I(R_i = r) \left\{ \hat{\Psi}_{IPW,i}^{(1r)}(d_{2r}, \eta_{2r}) + \hat{\Psi}_{Aug,i}^{(1r)}(d_{2r}, \eta_{2r}) \right\}$. For each $r = 1, \dots,$

B , these are functions of η_{2r}

Step 1 of the backward algorithm involves maximizing any of these “value” estimators in η_2 , which, because η_{2r} , $r = 1, \dots, B$, are nonoverlapping, is equivalent to the following.

Step 1: Let $\hat{\mathcal{V}}_{1r}(d_{2r}, \eta_{2r})$ be one of $\hat{\mathcal{V}}_{1r}^{IPW}(d_{2r}, \eta_{2r})$, $\hat{\mathcal{V}}_{1r}^{AIPW}(d_{2r}, \eta_{2r})$ or $\hat{\mathcal{V}}_{1r}^{CAIPW}(d_{2r}, \eta_{2r})$ obtain

$\hat{\eta}_{2r}^{\text{opt}} = \arg \max_{\eta_{2r}} \hat{\mathcal{V}}_{1r}(d_{2r}, \eta_{2r})$; and estimate $d_{2r}(h_2; \eta_{2r}^{\text{opt}})$ by $d_{2r}(h_2; \hat{\eta}_{2r}^{\text{opt}}) = d_{2r, \hat{\eta}_{2r}^{\text{opt}}}(h_2)$, $r = 1,$

$\dots, B$. Write $\hat{\eta}_2^{\text{opt}} = (\hat{\eta}_{21}^{\text{opt}}, \dots, \hat{\eta}_{2B}^{\text{opt}})^T$, and let $d_2(h_2; \hat{\eta}_2^{\text{opt}}) = d_{2, \hat{\eta}_2^{\text{opt}}}(h_2)$ denote the overall

estimated optimal Decision 2 rule involving these estimated rules for each $r = 1, \dots, B$.

Now move backward to Decision 1. Our goal is to maximize $\mathcal{V}_2(d_\eta) = \mathcal{V}_2(d_{1, \eta_1}, d_{2, \eta_2})$ in η .

Step 1 yields $d_{2, \hat{\eta}_2^{\text{opt}}}$, an estimator for an optimal rule at Decision 2. We now seek an

estimator for η_1^{opt} acknowledging that an optimal rule is to be used at Decision 2. Intuitively, this suggests maximizing in η_1 an estimator for $\mathcal{V}_2(d_\eta)$ holding d_{2, η_2} fixed at $d_{2, \hat{\eta}_2^{\text{opt}}}$, as

follows.

Step 2. Let $\hat{\mathcal{V}}_2(d_\eta) = \hat{\mathcal{V}}_2(d_{1, \eta_1}, d_{2, \eta_2})$ be one of $\hat{\mathcal{V}}_2^{IPW}(d_\eta)$, $\hat{\mathcal{V}}_{1r}^{AIPW}(d_\eta)$ or $\hat{\mathcal{V}}_{1r}^{CAIPW}(d_\eta)$; obtain

$\hat{\eta}_1^{\text{opt}} = \arg \max_{\eta_1} \hat{\mathcal{V}}_2(d_{1, \eta_1}, d_{2, \hat{\eta}_2^{\text{opt}}})$; and estimate $d_1(h_1; \eta_1^{\text{opt}})$ by $d_1(h_1; \hat{\eta}_1^{\text{opt}}) = d_{1, \hat{\eta}_1^{\text{opt}}}(h_1)$.

Following steps 1 and 2, the estimator for an optimal regime $d_{\eta^{\text{opt}}}$ is $d_{\hat{\eta}^{\text{opt}}} = \left(d_{1, \hat{\eta}_1^{\text{opt}}}, d_{2, \hat{\eta}_2^{\text{opt}}} \right)$.

An associated estimator for the value $\mathcal{V}_2(d_{\eta^{\text{opt}}})$ achieved by $d_{\eta^{\text{opt}}}$ is $\hat{\mathcal{V}}_2(d_{\hat{\eta}^{\text{opt}}})$, where $\hat{\mathcal{V}}_2(\cdot)$ is any of $\hat{\mathcal{V}}_2^{\text{IPW}}(\cdot)$, $\hat{\mathcal{V}}_2^{\text{AIPW}}(\cdot)$, or $\hat{\mathcal{V}}_2^{\text{CAIPW}}(\cdot)$.

It is not clear that this two-step scheme yields a valid estimator for $d_{\eta^{\text{opt}}}$ maximizing $\mathcal{V}_2(d_{\eta})$.

In Web Appendix D, we argue that, if the restricted class $\mathcal{D}_{\eta}$ contains d^{opt} , so that d^{opt} and $d_{\eta^{\text{opt}}}$ are equivalent, then $d_{\hat{\eta}^{\text{opt}}}$ is a valid estimator for $d_{\eta^{\text{opt}}}$ but may not be if $d^{\text{opt}} \notin \mathcal{D}_{\eta}$.

4.3. Classification Perspective

For a restricted class of regimes $\mathcal{D}_{\eta}$ indexed by finite-dimensional η , the backward iterative algorithm separates maximization of $\mathcal{V}_2(d_{\eta})$ in η into two lower-dimensional steps. For general $\mathcal{D}_{\eta}$, these steps can still be computationally challenging. When there are two treatment options at Decision 1, $\mathcal{A}_1 = \{0, 1\}$, and two options at Decision 2 for each $r = 1, \dots, B$, coded as $\mathbf{0}$ and $\mathbf{1}$ for each r , $\mathcal{A}_{2r} = \{0, 1\}$, the rules d_{1, η_1} and $d_{2r, \eta_{2r}}$, $r = 1, \dots, B$, are binary, which suggests making an analogy of a decision rule to a classifier (Zhang et al., 2012b; Zhao et al., 2012, 2015a; Bai et al., 2017). As advocated by these authors, taking $\mathcal{D}_{\eta}$ to comprise regimes for which the form of rules is determined by a specific classification method allows well-developed computational strategies for classification problems to be exploited.

Consider step 1. For each $r = 1, \dots, B$, let $\hat{\mathcal{V}}_{1r}(d_{2r, \eta_{2r}})$ be one of $\hat{\mathcal{V}}_{1r}^{\text{IPW}}(d_{2r, \eta_{2r}})$, $\hat{\mathcal{V}}_{1r}^{\text{AIPW}}(d_{2r, \eta_{2r}})$, or $\hat{\mathcal{V}}_{1r}^{\text{CAIPW}}(d_{2r, \eta_{2r}})$, and let $\Gamma_i I(R_i = r) \hat{\Psi}_i^{(1r)}(d_{2r, \eta_{2r}})$ denote the relevant summand, which, although suppressed in the notation, depends on $(H_{2i}, A_{2i}, \Delta_i, U_i)$. We then have

$$\begin{aligned} \hat{\mathcal{V}}_{1r}(d_{2r, \eta_{2r}}) n^{-1} \sum_{i=1}^n \Gamma_i I(R_i = r) \hat{\Psi}_i^{(1r)}(d_{2r, \eta_{2r}}) &= n^{-1} \sum_{i=1}^n \Gamma_i I(R_i = r) \\ &\times \left[\hat{\Psi}_i^{(1r)}(1)_{d_{2r, \eta_{2r}}}(H_{2i}) + \hat{\Psi}_i^{(1r)}(0) \left\{ 1 - d_{2r, \eta_{2r}} \right\} \right] \\ &= n^{-1} \sum_{i=1}^n \Gamma_i I(R_i = r) \left\{ \hat{CF}^{(1r)}(H_{2i}, A_{2i}, \Delta_i, U_i) d_{2r, \eta_{2r}}(H_{2i}) + \hat{\Psi}_i^{(1r)}(0) \right\}. \end{aligned} \quad (21)$$

where $\hat{CF}_i^{(1r)} = \hat{CF}^{(1r)}(H_{2i}, A_{2i}, \Delta_i, U_i) = \hat{\Psi}_i^{(1r)}(1) - \hat{\Psi}_i^{(1r)}(0)$, and $\hat{\Psi}_i^{(1r)}(a)$, $a = 0, 1$, implies evaluation at $d_{2r, \eta_{2r}}(H_{2i}) = a$, a predictor for the contrast function

$CF^{(1r)}(H_2) = E[g\{T^*(A_1, 1)\} | H_2, R = r] - E[g\{T^*(A_1, 0)\} | H_2, R = r]$ for patient i ; see Web Appendix E. Because in maximizing (21) $\hat{\Psi}_i^{(1r)}(0)$ can be disregarded, as it does not depend on η_{2r} , it can be shown that (Zhang et al., 2012b) maximizing $\hat{\mathcal{V}}_{1r}(d_{2r}, \eta_{2r})$ in η_{2r} is equivalent to minimizing

$$\sum_{i=1}^n \Gamma_i I(R_i = r) \left| \hat{CF}_i^{(1r)} \right| I\left\{ I(\hat{CF}_i^{(1r)} \geq 0) \neq d_{2r, \eta_{2r}}(H_{2i}) \right\}. \quad (22)$$

For step 2, with $\hat{\mathcal{V}}_2(d_\eta) = \hat{\mathcal{V}}_2(d_{1, \eta_1}, d_{2, \eta_2})$ one of $\hat{\mathcal{V}}_2^{\text{IPW}}(d_\eta)$, $\hat{\mathcal{V}}_2^{\text{AIPW}}(d_\eta)$, or $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d_\eta)$, letting the relevant summand be $\hat{\Psi}_i^{(2)}(d_\eta)$, by similar manipulations and definitions,

$$\begin{aligned} \hat{\mathcal{V}}_2(d_{1, \eta_1}, d_{2, \eta_2^{\text{opt}}}) &= n^{-1} \sum_{i=1}^n \hat{\Psi}_i^{(2)}(d_{1, \eta_1}, d_{2, \eta_2^{\text{opt}}}) \quad (23) \\ &= n^{-1} \sum_{i=1}^n \left\{ \hat{CF}_i^{(2)} d_{1, \eta_1}(H_{1i}) + \hat{\Psi}_i^{(2)}(0, d_{2, \eta_2^{\text{opt}}}) \right\} \end{aligned}$$

where $\hat{CF}_i^{(2)} = \hat{CF}^{(2)}(H_{2i}, A_{2i}, \Delta_i, U_i) = \hat{\Psi}_i^{(2)}(1, d_{2, \eta_2^{\text{opt}}}) - \hat{\Psi}_i^{(2)}(0, d_{2, \eta_2^{\text{opt}}})$, a predictor of the

contrast function $CF^{(2)}\left(d_{2, \eta_2^{\text{opt}}}, H_1\right)$ for patient i , where

$CF^{(2)}(d_2, H_1) = E[g\{T^*(1, d_2)\} | H_1] - E[g\{T^*(0, d_2)\} | H_1]$. Then maximizing $\hat{\mathcal{V}}_2(d_{1, \eta_1}, d_{2, \eta_2^{\text{opt}}})$

in d_{1, η_1} is equivalent to minimizing

$$\sum_{i=1}^n \left| \hat{CF}_i^{(2)} \right| I\left\{ I(\hat{CF}_i^{(2)} \geq 0) \neq d_{1, \eta_1}(H_{1i}) \right\}. \quad (24)$$

Each of the minimization criteria in (22) for $r = 1, \dots, B$ and (24) has the form of a weighted classification error, with $\left| \hat{CF}_i^{(1r)} \right|$ and $\left| \hat{CF}_i^{(2)} \right|$ playing the role of “weight,” $I(\hat{CF}_i^{(1r)} \geq 0)$ and $I(\hat{CF}_i^{(2)} \geq 0)$ playing the role of “label,” and $d_{2r, \eta_{2r}}(H_{2i})$ and $d_{1, \eta_1}(H_{1i})$ playing the role of “classifier.” Thus, if $\mathcal{D}_\eta$ is chosen to comprise rules in the form of classifiers corresponding to any popular classification method, such as support vector machines (SVM; Cortes and Vapnik, 1995) or classification and regression trees (CART; Breiman et al., 1984), well-developed computational strategies for minimizing weighted classification error can be used

to implement steps 1 and 2. For illustration here, we focus on $\mathcal{D}_\eta$ characterized by linear SVM, so that its elements $d_\eta = (d_{1,\eta_1}, d_{2,\eta_2})$ are of the form $d_{1,\eta_1}(h_1) = I\left\{\eta_1^T f_1(h_1) \geq 0\right\}$ and $d_{2,\eta_{2r}}(h_2) = I\left\{\eta_{2r}^T f_{2r}(h_2) \geq 0\right\}$, $r = 1, \dots, B$, where $f_1(h_1)$ and $f_2(h_2)$ are specified, vector-valued functions of h_1 and h_2 with first element equal to one.

It is straightforward that, under this choice of $\mathcal{D}_\eta$, (22) can be rewritten as

$$\sum_{i=1}^n \Gamma_i I(R_i = r) \left| \hat{CF}_i^{(1r)} \right| \phi_{0-1} \left[\left\{ 2I(\hat{CF}_i^{(1r)} \geq 0) - 1 \right\} \eta_{2r}^T f_{2r}(H_{2i}) \right], \quad (25)$$

where $\phi_{0-1}(x) = I(x < 0)$ is the so-called 0–1 loss function; and similarly for (24),

$$\sum_{i=1}^n \left| \hat{CF}_i^{(2)} \right| \phi_{0-1} \left[\left\{ 2I(\hat{CF}_i^{(2)} \geq 0) - 1 \right\} \eta_1^T f_1(H_{1i}) \right]. \quad (26)$$

The 0–1 loss function is nonconvex. Optimization of objective functions involving nonconvex loss functions is challenging and cannot be carried out using conventional techniques; thus, (25) and (26) make explicit the difficulty with direct maximization in steps 1 and 2.

A standard strategy for minimizing weighted classification error as in (22) and (24) is to replace the nonconvex 0–1 loss function by a convex “surrogate”; a popular choice used with SVM is the “hinge loss” function $\phi_{hinge}(x) = (1 - x)^+ = \max(1 - x, 0)$ (Zhao et al., 2012). Making this substitution in (25) leads to implementation of step 1 for $r = 1, \dots, B$ as

$$\begin{aligned} \hat{\eta}_{2r}^{\text{opt}} = \arg \min_{\eta_{2r}} & \left(\sum_{i=1}^n \Gamma_i I(R_i = r) \left| \hat{CF}_i^{(1r)} \right| \phi_{hinge} \left[\left\{ 2I(\hat{CF}_i^{(1r)} \geq 0) - 1 \right\} \times \eta_{2r}^T f_{2r}(H_{2i}) \right] \right. \\ & \left. + k_{2r} \sum_{j=1}^{p_{2r}} \left| \eta_{2r,j} \right| \right), \end{aligned} \quad (27)$$

where $p_{2r} = \dim(r_{2r})$, and the second ℓ_1 penalty term prevents overfitting and supports variable selection; k_{2r} can be chosen by cross-validation; see Section 6. It can be shown that using the hinge loss function results in an estimated rule that is Fisher consistent for a true optimal rule as in Zhao et al. (2012); see Web Appendix E. Similarly, step 2 becomes

$$\hat{\eta}_1^{\text{opt}} = \arg \min_{\eta_1} \left(\sum_{i=1}^n \left| \hat{CF}_i^{(2)} \right| \phi_{hinge} \left[\left\{ 2I(\hat{CF}_i^{(2)} \geq 0) - 1 \right\} \eta_1^T f_1(H_{1i}) \right] + k_1 \sum_{j=1}^{p_1} \left| \eta_{1,j} \right| \right), \quad (28)$$

where $p_1 = \dim(\eta_1)$. As in Bai et al. (2017), (27) and (28) can be converted to linear convex optimization problems and solved easily using linear programming software.

Under regularity conditions and a nonexceptional law, Zhang et al. (2012) proposed and Luedtke and van der Laan (2016) demonstrated that approximate standard errors for $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d_{\hat{\eta}^{\text{opt}}})$ can be obtained by substituting $d_{\hat{\eta}^{\text{opt}}}$ in the expression for fixed d .

Variance estimators for $\hat{\mathcal{V}}_2^{\text{IPW}}(d_{\hat{\eta}^{\text{opt}}})$ and $\hat{\mathcal{V}}_2^{\text{AIPW}}(d_{\hat{\eta}^{\text{opt}}})$ are more problematic. See Web Appendix B.

5. Simulation Studies

To evaluate performance of the methods, we carried out simulations under (i) a SMART scenario and two observational study scenarios. Although theoretically, the CAIPW and AIPW value estimators at steps 1 and 2 are doubly robust and thus consistent if either the propensity models or event models and models for (13) are correctly specified, as noted in Web Appendix A, specifying correct models for (13) is almost impossible. Thus, for observational studies, where the propensities are unknown, double robustness may not always be feasible. As, intuitively, consistency of the value estimators should impact directly the quality of estimation of an optimal regime, this suggests that special attention should be paid to propensity modeling. Accordingly, the observational study scenarios incorporate two cases: (ii) “gross” misspecification of the propensity models, and (iii) “mild” misspecification reflecting an analyst’s failure to include a covariate not strongly associated with treatment.

In all scenarios, $n = 500$; $B = 1$; $\mathcal{A}_1 = \{0, 1\}$; $\mathcal{A}_{21} = \{0, 1\}$; $g(t) = \min(t, L)$, $L = 150$; and $H_1 = X_1 = (X_{11}, X_{12})^T$, with $X_{11} \sim \mathcal{N}(0, 1)$, $X_{12} \sim \text{uniform}(0, 1)$. We took $A_1 \sim \text{Bernoulli}(b_1)$, where $b_1 = 0.5$ in scenario (i), $b_1 = \text{expit}(-0.1 + 0.75X_{11})$ in (ii), and $b_1 = \text{expit}(0.1 + 0.2X_{11} - 0.4X_{12})$ in (iii); and $\text{expit}(u) = (1 + e^{-u})^{-1}$. We then generated $S_d \sim \text{exponential}\{-5.5 + 0.15X_{11} + 0.2X_{12} - A_1(1.3 + 1.55X_{11} - 2.6X_{12})\}$, the potential event time in stage 1; and $S_{ss} \sim \text{exponential}\{-4.2 + 0.15X_{11} + 0.2X_{12} - A_1(-1.2 + 1.1X_{11} + 2.2X_{12})\}$, the potential time to stage 2, and took $S = \min(S_d, S_{ss})$. We generated C as described below. If $\Gamma = I\{\min(S, L) < C\} = 1$, if $S = S_d$, $R = 0$ and $T = S$; if $S = S_{ss}$, $R = 1$; and $V = \min(S, L, C)$. If $R = 1$, we generated $X_{21} \sim \text{Bernoulli}\{\text{expit}(0.2 + 0.5X_{11} + 0.4X_{12} + 0.6A_1)\}$, and took $A_2 \sim \text{Bernoulli}(b_2)$, where $b_2 = 0.5$ in (i), $b_2 = \text{expit}(-1.0 + 2.0X_{21})$ in (ii), and $b_2 = \text{expit}(-0.3 - 0.6X_{12} + X_{21})$ in (iii). With $H_2 = \{H_1, A_1, V, \Gamma, R, \Gamma I(R > 0)X_{21}\}$, “additional life” beyond Decision 2 was generated as $AL \sim \text{exponential}\{-3.5 + 0.12X_{11} - 0.42X_{12} - 0.52X_{21} - A_2(0.55 + 0.8X_{11} - 0.6X_{12} - 0.2A_1 - 0.65X_{21})\}$, and $T = S + AL$. We then found $U = \min(T, L, C)$, and $\mathcal{C} = I\{\min(T, L) < C\}$.

We carried out simulations for all scenarios under a variety of distributions for C and report results for three representative censoring mechanisms: (1) exponential(0.003), (2) uniform(50, 250), and (3) exponential(0.008), resulting in 19%, 18%, and 41% censoring overall, respectively. Qualitative results are insensitive to the censoring mechanism. We took the classification perspective in Section 4.3 with linear SVM and $f_1(u) = f_2(u) = (1, u^T)^T$, so

that $\mathcal{D}_\eta$ has elements $d_\eta = (d_{1,\eta_1}, d_{2,\eta_2})$ with $d_{1,\eta_1}(h_1) = I(\eta_1^T \tilde{h}_1 \geq 0)$, $d_{21,\eta_{21}}(h_2) = I(\eta_{21}^T \tilde{h}_2 \geq 0)$, $\tilde{h}_1 = (1, h_1^T)^T$, $\tilde{h}_2 = (1, h_2^T)^T$. Under these choices, $\mathcal{V}_2(d_{\eta \text{opt}}) = 93.9$ for all scenarios. See Web Appendix F for details on how the true optimal restricted regime was obtained.

Under each scenario, censoring distribution, and configuration of correctly or incorrectly specified models described below, we generated 1000 Monte Carlo data sets and implemented steps 1 and 2 using each of the IPW, AIPW, and CAIPW versions of $\hat{\mathcal{V}}_{1r}(\cdot)$ and $\hat{\mathcal{V}}_2(\cdot)$, obtaining $d_{\hat{\eta} \text{opt}}$ and associated estimates $\hat{\mathcal{V}}_2^{\text{IPW}}(d_{\hat{\eta} \text{opt}})$, $\hat{\mathcal{V}}_2^{\text{AIPW}}(d_{\hat{\eta} \text{opt}})$, and $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d_{\hat{\eta} \text{opt}})$ of $\hat{\mathcal{V}}_2(d_{\hat{\eta} \text{opt}})$ for each. We also considered no censoring, where $\hat{\mathcal{V}}_2^{\text{AIPW}}(d_{\hat{\eta} \text{opt}})$ is the locally efficient estimator. Standard errors were obtained as described in Web Appendix B. For each estimator $d_{\hat{\eta} \text{opt}}$, as described in Web Appendix F, we also obtained $\hat{\mathcal{V}}_2(d_{\hat{\eta} \text{opt}})$, the expected transformed outcome that would be achieved if all patients in the population were to follow regime $d_{\hat{\eta} \text{opt}}$. Finally, we obtained based on separate simulated test data the proportions of patients on average for whom the treatment option(s) selected by $d_{\hat{\eta} \text{opt}}$ agree with those selected by the true optimal rule at the first decision and overall; see Web Appendix F for details.

In all cases, we represented the “additional life” hazard function $\lambda_{21}(u | H_2, A_2)$ by a proportional hazards model $\lambda_{20}(u) \exp\{\beta_{21}X_{11} + \beta_{22}X_{12} + \beta_{23}A_1 + \beta_{24}A_{21} + A_2(\beta_{25} + \beta_{26}X_{11} + \beta_{27}X_{12} + \beta_{28}A_1 + \beta_{29}X_{21})\}$ for unspecified baseline hazard function $\lambda_{20}(u)$, which is correctly specified; an incorrect model was obtained by deleting X_{11} in (i) and (ii) and X_{12} in (iii). The cause-specific event hazard functions for $r = 0, 1$ were represented correctly by the proportional hazards model $\lambda_{10r}(u) \exp\{\beta_{11r}X_{11} + \beta_{12r}X_{12} + A_1(\beta_{13r} + \beta_{14r}X_{11} + \beta_{15r}X_{12})\}$ and incorrectly by deleting X_{11} in (i) and (ii) and X_{12} in (iii). We represented (13) as the linear model $\xi_{10} + \xi_{11}X_{11} + \xi_{12}X_{12} + \xi_{13}A_1$. As in Web Appendix A, this model is almost certainly incorrect so was misspecified in all cases and made “even more incorrect” by deleting X_{11} in (i) and (ii) and X_{12} in (iii). Finally, we represented $\alpha_1(u | H_1, A_1)$ and $\alpha_2(u | H_1, A_1, S, R, X_{2r}, A_2)$ in (2) by proportional hazards models depending on X_{11} , X_{12} , A_1 , and X_{12} as described in Web Appendix F, which contain the true censoring distributions.

In (i), $\pi_1(H_1, 1)$ and $\pi_2(H_2, 1)$ are known and were represented as sample proportions as in Web Appendix F, where

$\pi_{d_{21}}(H_2; \eta_2) = \pi_2(H_2, 1)d_{21}(H_2; \eta_{21}) + \left\{1 - \pi_2(H_2, 1)\right\} \left\{1 - d_{21}(H_2; \eta_{21})\right\}$. In (ii) and (iii), these were modeled correctly as $\pi_1(H_1, 1; \psi_1) = \text{expit}(\psi_{10} + \psi_{11}X_{11} + \psi_{12}X_{12})$ and $\pi_2(H_2, 1; \psi_2) = \text{expit}\{\psi_{20} + \psi_{21}X_{11} + \psi_{22}X_{12} + \psi_{23}X_{21} + A_1(\psi_{24} + \psi_{25}X_{11} + \psi_{26}X_{12} + \psi_{27}X_{21})\}$, which contain the true propensities. These were modeled incorrectly in (ii) by deleting all

covariates; that is, taking $\pi_1(H_1, 1; \psi_1) = \text{expit}(\psi_{10})$ and $\pi_2(H_2, 1; \psi_2) = \text{expit}(\psi_{20})$. In (iii), these were modeled incorrectly by deleting X_{12} .

Tables 2–4 present the results. In all scenarios, estimation based on CAIPW performs best, even when the hazard models or, in the case of (ii) and (iii), propensity models, are misspecified. The Monte Carlo average of values achieved by the estimated optimal regime is close to the true optimum (93.9), and that for the estimated value of an optimal regime is close to the true optimum in most cases, but see below. The average of standard errors is close to the Monte Carlo standard deviation for $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d_{\hat{\eta}^{\text{opt}}})$, and confidence intervals achieve the nominal level when all models are correct, as discussed in Web Appendix B; this persists even if the hazard models are misspecified. The estimated optimal regime selects treatment consistent with the true optimal regime a greater proportion of the time than those based on IPW and AIPW in almost all cases. In (i) and (iii), $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d_{\hat{\eta}^{\text{opt}}})$ appears doubly robust even though the models for (13) are always incorrect, but shows upward bias under “gross” propensity misspecification in (ii), which may be too severe to overcome the incorrect models for (13). Performance is degraded when all models are misspecified.

In all scenarios, performance of the estimators based on IPW and AIPW is worse. The averages of values achieved by the estimated regimes are smaller than the true optimum, and $\hat{\mathcal{V}}_2^{\text{IPW}}(d_{\hat{\eta}^{\text{opt}}})$ and $\hat{\mathcal{V}}_2^{\text{AIPW}}(d_{\hat{\eta}^{\text{opt}}})$ are biased upward in most cases, as observed in other contexts (Zhang et al., 2012a, 2013). It is not surprising that maximizing consistent (under correct models) value estimators yields overestimates, but this is much more pronounced than for $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d_{\hat{\eta}^{\text{opt}}})$. As discussed in Web Appendix B, standard errors and confidence intervals for these estimators are expected to be conservative, as borne out in Tables 2–4.

6. Analysis of Study C9710

We apply the proposed methods to North American Leukemia Intergroup Study C9710 with composite time to event outcome defined in Section 1, referred to as event-free survival (EFS) by Powell et al. (2010). Here, $\mathcal{A}_1$ comprises consolidation therapies ATRA (coded as 0) and ATRA in combination with arsenic trioxide (1), and $B = 1$. Patients event-free and uncensored at the end of consolidation, so with $R = 1$, were randomized to maintenance therapy of ATRA (coded as 0) or ATRA + Mtx + MP (1), and $\mathcal{A}_{21} = \{0, 1\}$.

Baseline covariates in X_1 include age, gender, race, ethnicity, Eastern Cooperative Oncology Group (ECOG) performance status, risk group, white blood cell count, platelet count, serum creatinine, and hemoglobin. For patients with $R = 1$, X_{2R} includes adverse events and toxicities occurring during induction or consolidation therapy. We consider only adverse events experienced by at least 5% of patients. See Web Appendix G for further details on the data used here, which involve $n = 468$ patients at stage 1312 of whom progressed to Decision 2 (maintenance therapy). We consider $g(t) = \min(t, L)$ with $L = 2500$ days, which is censored overall for 19% of patients, 6% occurring in stage 1.

We used the methods in Section 4.3 with CAIPW estimators at steps 1 and 2, with $\mathcal{D}_\eta$ induced by linear SVM and k_{21} and k_1 chosen by cross-validation to estimate $d_{\eta^{\text{opt}}}$. Details of development of the required models, including the use of stepwise variable selection, are presented in Web Appendix G. The estimated optimal rule at Decision 2 is $d_{2, \hat{\eta}_2^{\text{opt}}}(h_2) = d_{21, \hat{\eta}_{21}^{\text{opt}}}(h_{21}) = I(0.45 - 0.89TB \geq 0)$, where TB is the indicator that the adverse event of thrombosis/thrombus/embolism occurred during induction or consolidation therapy. That at Decision 1 is $d_{1, \hat{\eta}_1^{\text{opt}}}(h_1) \equiv 1$, which dictates that all patients should receive ATRA with arsenic trioxide, regardless of baseline characteristics. The associated estimate of the value using $\hat{\mathcal{V}}_2^{\text{CAIPW}}(d_{\hat{\eta}^{\text{opt}}})$ is 2165 days (standard error 55 days).

Powell et al. (2010) considered separate, conventional analyses comparing consolidation and maintenance therapies, respectively. The estimated optimal rule at Decision 1 is consistent with the strong evidence found by Powell et al. favoring ATRA with arsenic trioxide to ATRA on the basis of EFS. For maintenance therapy, the authors found borderline evidence favoring ATRA+Mtx+MP over ATRA alone. The estimated optimal rule at Decision 2 offers insight into this finding. This rule selects maintenance therapy ATRA+Mtx+MP for all patients except those experiencing the adverse event ($TB = 1$), which was the case for 10% of study subjects who reached the second stage. Viewing consolidation and maintenance therapy as components of a regime thus highlights for further scrutiny this subpopulation of patients.

7. Discussion

We have proposed an approach to estimation of a two-stage optimal dynamic treatment regime and the associated value when the outcome is a censored time to an event. By an analogy to a weighted classification problem, the class of regimes is restricted by choice of a classification method, and well-established computational strategies can be used to implement the approach via a backward iterative scheme. Simulation studies and an application demonstrate good practical performance of the methods. We are working on extension to more than two decisions; this work will be reported elsewhere.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants P01CA142538, R01HL118336, and U10CA180821. The authors thank the Alliance for Clinical Trials in Oncology for the Study C9710 data. This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

References

- Bai X, Tsiatis AA, Lu W, Song R. 2017; Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime Data Analysis*. 23:585–604. [PubMed: 27480339]
- Breiman, L, Friedman, J, Stone, CJ, Olshen, RA. *Classification and Regression Trees*. Boca Raton: Chapman and Hall; 1984.
- Cortes C, Vapnik V. 1995; Support-vector networks. *Machine Learning*. 20:273–297.
- Goldberg Y, Kosorok MR. 2012; Q-learning with censored data. *The Annals of Statistics*. 40:529–560. [PubMed: 22754029]
- Jiang R, Lu W, Song R, Davidian M. 2017; On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society, Series B*. 79:1165–1185.
- Lavori PW, Dawson R. 2004; Dynamic treatment regimes: Practical design considerations. *Clinical Trials*. 1:9–20. [PubMed: 16281458]
- Luedtke AR, van der Laan MJ. 2016; Statistical inference for mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics*. 44:713–742. [PubMed: 30662101]
- Murphy SA. 2003; Optimal dynamic treatment regimes (with discussion). *Journal of the Royal Statistical Society, Series B*. 65:331–366.
- Murphy SA. 2005; An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*. 24:1455–1481. [PubMed: 15586395]
- Powell BL, Moser B, Stock W, Gallagher RE, Willman CL, Stone RM, et al. 2010; Arsenic trioxide improves event-free and overall survival for adults with acute promyelocytic leukemia: North American Leukemia Inter-group Study C9710. *Blood*. 116:3751–3757. [PubMed: 20705755]
- Robins, JM. Optimal structural nested models for optimal sequential decisions. In: Lin, DY, Heagerty, PJ, editors *Proceedings of the Second Seattle Symposium in Biostatistics*. New York: Springer; 2004. 189–326.
- Schulte PJ, Tsiatis AA, Laber EB, Davidian M. 2014; Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science*. 29:640–661. [PubMed: 25620840]
- Tao Y, Wang L. 2017; Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics*. 73:145–155. [PubMed: 27213913]
- Tsiatis, AA. *Semiparametric Theory and Missing Data*. New York: Springer; 2006.
- Xu Y, Müller P, Wahed AS, Thall PF. 2016; Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *Journal of the American Statistical Association*. 111:921–950. [PubMed: 28018015]
- Zhang B, Tsiatis AA, Laber EB, Davidian M. 2012a; A robust method for estimating optimal treatment regimes. *Biometrics*. 68:1010–1018. [PubMed: 22550953]
- Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber EB. 2012b; Estimating optimal treatment regimes from a classification perspective. *Stat*. 1:103–114. [PubMed: 23645940]
- Zhang B, Tsiatis AA, Laber EB, Davidian M. 2013; Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*. 100:681–694.
- Zhao YQ, Zeng D, Laber EB, Kosorok MR. 2015a; New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*. 110:583–598. [PubMed: 26236062]
- Zhao YQ, Zeng D, Laber EB, Song R, Yuan M, Kosorok MR. 2015b; Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*. 102:151–168. [PubMed: 25937641]
- Zhao YQ, Zeng D, Rush AJ, Kosorok MR. 2012; Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*. 107:1106–1118. [PubMed: 23630406]
- Zhao Y, Zeng D, Socinski MA, Kosorok MR. 2011; Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*. 67:1422–1433. [PubMed: 21385164]

Table 1

Summary of notation in temporal order. Some variables are redefined as in the text if $g(t) = \min(t, L)$.

Variable	Definition
$H_1 = \{X_1\}$	Baseline information
	<i>Potential outcomes</i>
$\{S^*(a_1), R^*(a_1)\}$	Length of first stage and status under $a_1 \in \mathcal{A}_1$ $R^*(a_1) = 0$: $S^*(a_1)$ = time to event $R^*(a_1) > 0$: $S^*(a_1)$ = time to second stage treatment $R^*(a_1) = r = 1, \dots, B$ determines set of feasible options $\mathcal{A}_{2r} \subseteq \mathcal{A}_2$
$X_{2R}^*(a_1)$	Information collected between Decisions 1 and 2 if $R^*(a_1) > 0$
$T^*(a_1, a_2)$	Time to event under $a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2 (= S^*(a_1)$ if $R^*(a_1) = 0$)
	<i>Observed data, no censoring</i>
A_1	Treatment in $\mathcal{A}_1$ received at Decision 1
(S, R)	Length of first stage and status under A_1 $R = 0$: S = time to event $R > 0$: S = time to second stage treatment $R = r = 1, \dots, B$ determines set of feasible options $\mathcal{A}_{2r} \subseteq \mathcal{A}_2$
X_{2R}	Information collected between Decisions 1 and 2 if $R > 0$
H_2	History at Decision 2 = $\{H_1, A_1, S, R, \mathbb{I}(R > 0)X_{2R}\}$
A_2	Treatment in $\mathcal{A}_2$ received at Decision 2 if $R > 0$
T	Time to event (= S if $R = 0$)
	<i>Observed data, censoring</i>
A_1	Treatment $\mathcal{A}_1$ received at Decision 1
C	Potential censoring time
(V, Γ)	$V = \min(S, C)$ = length of first stage $\Gamma = I(S < C)$ = indicator of no censoring in stage 1
ΓR	Status at end of stage 1 $\Gamma R = 0$: V = time to event or censoring, $R = 0$ or $\Gamma = 0$ $\Gamma R > 0$: V = time to second stage treatment, $R > 0, \Gamma = 1, \Gamma R = R$ $R = r = 1, \dots, B$ determines set of feasible options $\mathcal{A}_{2r} \subseteq \mathcal{A}_2$
X_{2R}	Information collected between Decisions 1 and 2 if $\Gamma R > 0$
H_2	History at Decision 2 = $\{H_1, A_1, V, \Gamma, \Gamma R, \mathbb{I}(\Gamma R > 0)X_{2R}\}$
A_2	Treatment in $\mathcal{A}_2$ received at Decision 2 if $R > 0$

Variable	Definition
(U, δ)	$U = \min(T, C) = \text{time to event or censoring (} = V \text{ if } \Gamma R = 0, \Gamma = 1 \text{)},$ $\delta = I(T \leq C) = \text{indicator of no censoring}$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Simulation results using 1000 Monte Carlo data sets, $n = 500$, SMART scenario (i), for different censoring schemes described in the text. Misspec indicates if the “additional life” and cause-specific event hazard functions were correctly (Correct) or incorrectly (Event) misspecified; for Event, the model for (13) was “even more incorrectly” specified as described in the text. $\hat{\mathcal{V}}_2(d_{\eta}^{\text{opt}})$ is the average over the 1000 data sets of Monte Carlo estimates of the true value of the estimated restricted regime for each data set, obtained for each data set using g-computation as described in Web Appendix F based on 10,000 Monte Carlo replications. $\hat{\mathcal{V}}_2(d_{\eta}^{\text{opt}})$ is Monte Carlo average (standard deviation) over the 1000 data sets of the estimated value of the optimal regime. SE is the Monte Carlo average (standard deviation) of estimated standard errors for $\hat{\mathcal{V}}_2(d_{\eta}^{\text{opt}})$. Coverage is based on 95% Wald-type confidence intervals for the true value $\mathcal{V}_2(d_{\eta}^{\text{opt}}) = 93.9$. Agree Stage 1 is the average proportion of patients over 10,000 independent data sets for which the treatment option selected by the estimated Decision 1 rule agrees with that dictated by the true optimal Decision 1 rule. Agree Overall is the average proportion for which the options selected by the estimated regime are consistent with those dictated by the true optimal regime; see Web Appendix F.

Table 2

Misspec	Method	$\hat{\mathcal{V}}_2(d_{\eta}^{\text{opt}})$	$\hat{\mathcal{V}}_2(d_{\eta}^{\text{opt}})$	SE	Coverage	Agree	
						Stage 1	Overall
Correct	AIPW	92.5 (1.7)	94.7 (3.6)	3.7 (0.2)	0.96	0.90	0.80
	IPW	89.6 (3.5)	96.6 (7.2)	7.8 (0.4)	0.96	0.79	0.66
Correct	<i>Exp(0.003) censoring</i>						
	CAIPW	92.1 (2.1)	94.3 (4.0)	4.0 (0.2)	0.96	0.88	0.78
	AIPW	90.5 (3.0)	95.0 (5.4)	6.8 (0.6)	0.98	0.82	0.69
	IPW	88.4 (4.0)	97.6 (8.4)	10.2 (0.8)	0.98	0.76	0.61
Event	CAIPW	91.7 (2.4)	93.9 (4.2)	4.3 (0.2)	0.95	0.86	0.75
	AIPW	90.1 (3.2)	94.7 (5.4)	6.9 (0.6)	0.98	0.81	0.67
Correct	<i>Unif(50,250) censoring</i>						
	CAIPW	92.3 (1.9)	94.0 (3.9)	4.0 (0.2)	0.95	0.88	0.78
	AIPW	89.6 (3.5)	95.8 (6.3)	8.6 (1.0)	0.99	0.79	0.65
	IPW	88.3 (4.0)	99.0 (9.0)	11.5 (1.1)	0.98	0.75	0.61
Event	CAIPW	91.8 (2.2)	93.9 (4.0)	4.2 (0.2)	0.96	0.86	0.76
	AIPW	89.4 (3.5)	95.8 (6.3)	8.7 (1.0)	0.99	0.79	0.65

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Misspec	Method	$\hat{\mathcal{V}}_2(d_{\eta}^{\text{opt}})$	$\hat{\mathcal{V}}_2(d_{\eta}^{\text{opt}})$	SE	Coverage	Agree	
						Stage 1	Overall
Correct		<i>Exp(0.008) censoring</i>					
	CAIPW	91.6 (2.4)	93.7 (4.5)	4.9 (0.4)	0.96	0.86	0.74
	AIPW	86.9 (4.4)	97.1 (8.3)	12.6 (2)	1.00	0.72	0.55
	IPW	86.2 (4.6)	100.9 (11.3)	15.6 (2.1)	0.99	0.70	0.54
	CAIPW	91.0 (2.8)	93.4 (4.6)	5.1 (0.4)	0.97	0.84	0.71
	AIPW	86.7 (4.5)	97.3 (8.5)	12.7 (2)	1.00	0.72	0.54

Table 3

Simulation results using 1000 Monte Carlo data sets, $n = 500$, observational study scenario (ii), for different censoring schemes described in the text; the case of no censoring is excluded for brevity. Misspec indicates which models were misspecified; Propensity indicates misspecification of the propensity models only, Event indicates misspecification of the “additional life” and cause-specific hazard functions and, “even more incorrect” misspecification of the model for (13), and, Both indicates that all of these models were misspecified. All other entries are as in Table 2.

Misspec	Method	$\mathcal{V}_2(d_{\eta}^{\text{opt}})$	$\hat{\mathcal{V}}_2(d_{\eta}^{\text{opt}})$	SE	Agree			
					Coverage	Stage 1	Overall	
Correct	<i>Exp(0.003) censoring</i>							
	CAIPW	91.4 (2.6)	94.7 (4.1)	4.3 (0.8)	0.95	0.85	0.74	
	AIPW	89.0 (3.9)	96.1 (5.9)	7.6 (1.9)	0.98	0.78	0.63	
	IPW	87.5 (4.5)	99.0 (8.5)	12.4 (3.4)	0.99	0.76	0.57	
	CAIPW	92.7 (1.1)	98.6 (4.1)	4.4 (0.2)	0.85	0.91	0.80	
Propensity	AIPW	91.3 (2.3)	99.2 (5.4)	7.4 (0.7)	0.97	0.86	0.71	
	IPW	86.3 (0.5)	144.4 (9.4)	11.2 (0.7)	0.00	0.94	0.52	
Event	CAIPW	91.2 (2.7)	94.5 (4.2)	4.5 (0.9)	0.96	0.84	0.73	
	AIPW	88.8 (4.0)	96.0 (5.8)	7.8 (2.0)	0.99	0.78	0.63	
Both	CAIPW	92.3 (1.4)	101.3 (4.3)	4.7 (0.3)	0.66	0.90	0.78	
	AIPW	91.1 (2.7)	102.1 (5.6)	7.6 (0.7)	0.90	0.87	0.71	
Correct	<i>Unif(50,250) censoring</i>							
	CAIPW	91.5 (2.6)	94.2 (4.0)	4.2 (0.6)	0.96	0.85	0.74	
	AIPW	87.9 (4.4)	97.4 (7.2)	9.9 (3.3)	0.99	0.76	0.59	
	IPW	86.9 (4.7)	100.1 (9.4)	14.0 (4.4)	0.99	0.75	0.55	
	CAIPW	92.7 (1.1)	97.9 (4.0)	4.4 (0.3)	0.88	0.91	0.80	
Propensity	AIPW	90.4 (3.0)	100.2 (6.7)	9.2 (1.2)	0.98	0.84	0.67	
	IPW	86.2 (0.6)	145.3 (10.3)	12.6 (1)	0.00	0.94	0.51	
Event	CAIPW	91.3 (2.7)	94.0 (4.0)	4.4 (0.9)	0.97	0.84	0.74	
	AIPW	87.8 (4.4)	97.5 (7.1)	10.2 (3.5)	0.99	0.76	0.59	
Both	CAIPW	92.4 (1.4)	100.8 (4.2)	4.6 (0.3)	0.70	0.90	0.78	
	AIPW	90.4 (3.1)	103.1 (6.9)	9.4 (1.2)	0.94	0.85	0.68	
<i>Exp(0.008) censoring</i>								

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Misspec	Method	$\hat{\mathcal{H}}_2(d_{\text{opt}}^h)$	$\hat{\mathcal{H}}_2(d_{\text{opt}}^h)$	SE	Coverage	Stage 1	Overall
							Agree
Correct	CAIPW	90.8 (3.0)	94.4 (4.7)	5.2 (0.9)	0.96	0.84	0.71
	AIPW	85.7 (5.0)	101.3 (10.7)	15.1 (5.3)	1.00	0.71	0.52
	IPW	85.4 (5.1)	105.7 (13.8)	19.6 (6.5)	1.00	0.71	0.51
Propensity	CAIPW	92.2 (1.6)	98.4 (4.8)	5.4 (0.4)	0.90	0.90	0.77
	AIPW	88.2 (4.2)	102.2 (9.2)	13.3 (2)	0.99	0.78	0.59
	IPW	86.3 (1.3)	145.3 (13.1)	17.2 (2)	0.05	0.92	0.52
Event	CAIPW	90.5 (3.2)	94.2 (4.8)	5.4 (1.1)	0.96	0.82	0.69
	AIPW	85.6 (5.0)	101.2 (10.9)	15.3 (5.4)	0.99	0.71	0.52
	CAIPW	91.9 (1.7)	101.2 (5.0)	5.7 (0.4)	0.79	0.89	0.75
Both	AIPW	88.2 (4.6)	104.9 (9.4)	13.6 (1.9)	0.98	0.79	0.60

Table 4

Simulation results using 1000 Monte Carlo data sets, $n = 500$, observational study scenario (iii), for different censoring schemes described in the text; the case of no censoring is excluded for brevity. All entries are as in Table 3.

Misspec	Method	$\mathcal{V}_2(d_{\text{opt}})$	$\hat{\mathcal{V}}_2(d_{\text{opt}})$	SE	Coverage	Agree	
						Stage 1	Overall
Correct	CAIPW	92.0 (2.1)	94.4 (3.9)	4.1 (0.3)	0.95	0.87	0.77
	AIPW	89.9 (3.3)	95.2 (5.3)	6.9 (0.7)	0.99	0.80	0.66
	IPW	88.8 (3.7)	96.4 (7.2)	10.6 (1.2)	0.99	0.78	0.61
Propensity	CAIPW	92.0 (2.1)	94.3 (4.0)	4.1 (0.3)	0.96	0.87	0.76
	AIPW	89.9 (3.2)	95.0 (5.5)	6.9 (0.8)	0.99	0.80	0.66
	IPW	88.1 (3.6)	99.5 (7.7)	10.7 (1.1)	0.98	0.76	0.59
Event	CAIPW	92.0 (2.1)	94.4 (3.9)	4.1 (0.3)	0.96	0.87	0.77
	AIPW	89.8 (3.4)	95.1 (5.3)	6.8 (0.7)	0.99	0.80	0.66
	CAIPW	91.9 (2.2)	94.6 (4.0)	4.1 (0.3)	0.94	0.87	0.76
Both	AIPW	89.8 (3.3)	95.3 (5.4)	6.9 (0.8)	0.99	0.80	0.65
	<i>Unif(50,250) censoring</i>						
Correct	CAIPW	92.2 (1.9)	94.3 (3.9)	4.0 (0.3)	0.96	0.88	0.78
	AIPW	88.7 (3.9)	96.3 (6.6)	8.8 (1.4)	0.99	0.78	0.61
	IPW	88.1 (4.0)	98.3 (8.4)	12.1 (1.8)	0.99	0.77	0.58
Propensity	CAIPW	92.2 (1.9)	94.1 (3.9)	4.0 (0.3)	0.96	0.87	0.78
	AIPW	88.6 (3.9)	96.1 (6.4)	8.9 (1.4)	0.99	0.77	0.61
	IPW	87.5 (3.8)	101.2 (8.5)	12.2 (1.6)	0.99	0.74	0.56
Event	CAIPW	92.2 (1.9)	94.3 (3.9)	4.0 (0.3)	0.96	0.88	0.78
	AIPW	88.7 (3.9)	96.3 (6.5)	8.8 (1.4)	0.99	0.78	0.61
	CAIPW	92.2 (1.8)	94.4 (3.9)	4.0 (0.3)	0.96	0.88	0.78
Both	AIPW	88.6 (4.0)	96.5 (6.4)	8.9 (1.4)	0.99	0.77	0.61
	<i>Exp(0.008) censoring</i>						
Correct	CAIPW	91.4 (2.4)	94.0 (4.7)	4.9 (0.5)	0.96	0.85	0.73
	AIPW	86.7 (4.7)	98.6 (9.0)	13.0 (2.5)	1.00	0.73	0.54

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Misspec	Method	$\hat{\mathcal{H}}_2(d_h^{\text{opt}})$	$\hat{\mathcal{H}}_2(d_h^{\text{opt}})$	SE	Coverage	Stage 1	Overall
Propensity	IPW	86.5 (4.6)	100.9 (10.7)	16.5 (3.0)	1.00	0.73	0.53
	CAIPW	91.4 (2.4)	93.8 (4.7)	4.9 (0.4)	0.95	0.85	0.73
	AIPW	86.7 (4.7)	98.6 (8.9)	13.1 (2.4)	1.00	0.73	0.54
	IPW	86.2 (4.5)	103.7 (11.1)	16.6 (2.8)	0.99	0.71	0.53
Event	CAIPW	91.4 (2.4)	94.0 (4.7)	4.9 (0.5)	0.96	0.85	0.73
	AIPW	86.6 (4.7)	98.6 (9.0)	13.0 (2.5)	0.99	0.73	0.54
Both	CAIPW	91.4 (2.5)	94.2 (4.7)	5.0 (0.4)	0.96	0.85	0.73
	AIPW	86.7 (4.7)	98.8 (8.9)	13.1 (2.4)	1.00	0.73	0.54