Title
Super-Learning of an Optimal Dynamic Treatment Rule

Journal
The International Journal of Biostatistics, 12(1)

ISSN
2194-573X

Authors
Luedtke, Alexander R
van der Laan, Mark J

Publication Date
2016-05-01

DOI
10.1515/ijb-2015-0052

Peer reviewed

Published in final edited form as:

Int J Biostat. 2016 May 01; 12(1): 305–332. doi:10.1515/ijb-2015-0052.

Super-Learning of an Optimal Dynamic Treatment Rule

Alexander R. Luedtke[*] and
Division of Biostatistics, University of California, Berkeley, CA, USA

Mark J. van der Laan
Division of Biostatistics, University of California, Berkeley, CA, USA

Mark J. van der Laan: laan@berkeley.edu

Abstract

We consider the estimation of an optimal dynamic two time-point treatment rule defined as the rule that maximizes the mean outcome under the dynamic treatment, where the candidate rules are restricted to depend only on a user-supplied subset of the baseline and intermediate covariates. This estimation problem is addressed in a statistical model for the data distribution that is nonparametric, beyond possible knowledge about the treatment and censoring mechanisms. We propose data adaptive estimators of this optimal dynamic regime which are defined by sequential loss-based learning under both the blip function and weighted classification frameworks. Rather than *a priori* selecting an estimation framework and algorithm, we propose combining estimators from both frameworks using a super-learning based cross-validation selector that seeks to minimize an appropriate cross-validated risk. The resulting selector is guaranteed to asymptotically perform as well as the best convex combination of candidate algorithms in terms of loss-based dissimilarity under conditions. We offer simulation results to support our theoretical findings.

Keywords

causal inference; cross-validation; dynamic treatment; loss function; oracle inequality

1 Introduction

Consider a time-dependent random variable consisting of baseline covariates, initial treatment and censoring indicator, intermediate covariates, subsequent treatment and censoring indicator, and a final outcome. A dynamic treatment rule is a rule that deterministically assigns treatment as a function of the available history. If treatment is assigned at two time points, then this dynamic treatment rule consists of two rules, one for each time point [1–3]. The mean outcome under a dynamic treatment is a counterfactual quantity of interest representing what the mean outcome would have been if everybody would have received treatment according to the dynamic treatment rule [1, 4–7]. We define the optimal dynamic multiple time-point treatment rule as the rule that maximizes the mean

[*]**Corresponding author**: Division of Biostatistics, University of California, Berkeley, CA, USA, aluedtke@berkeley.edu.

outcome under the dynamic treatment, where the candidate rules are restricted to only respond to a user-supplied subset of the baseline and intermediate covariates.

In this article, we aim to use an ensemble method known as super-learning [8–10] to learn the optimal dynamic treatment rule. This method will allow the user to data adaptively select from many state of the art methods for estimating the optimal treatment rule, yielding an estimation scheme which we prove performs at least as well as the best of these methods.

1.1 State of the art

Researchers have aimed to learn optimal rules from data generated by (sequential) multiple assignment randomized trials (SMART) [11]. Researchers have also aimed to learn dynamic treatments from observational studies: Cotton and Heagerty [12], Orellana et al. [13], Robins et al. [14], Petersen et al. [15, 16], Moodie et al. [17]. These observational and sequentially randomized studies provide an opportunity to learn an optimal multiple time-point dynamic treatment defined as the treatment rule that maximizes the mean dynamic-regime specific counterfactual outcome over a user supplied class of dynamic regimes.

The literature on Q-learning defines the optimal dynamic treatment among *all* dynamic treatments in a sequential manner [11, 18–21]. The optimal treatment rule for the second line treatment is defined as the maximizer of the conditional counterfactual mean outcome, given the observed past, over the possible second line treatments. The optimal treatment rule for the first line treatment is defined as the maximizer of the conditional mean counterfactual outcome, given baseline covariates, over the possible values for the initial treatment, under the assumption that the second line treatment is assigned according to the optimal rule for the second line treatment. The optimal rule can be learned through fitting the sequential regressions. Ernst et al. [23] and Ormoneit and Sen [24] use regression trees and kernel regression estimators, respectively.

Murphy [18] and Robins [19, 21] develop structural nested mean models tailored to optimal dynamic treatments. These models assume a parametric model for the “blip function” defined as the additive effect of a blip in current treatment on a counterfactual outcome, conditional on the observed past, in the counterfactual world in which future treatment is assigned optimally. Each blip function defines the optimal treatment rule for that time point by maximizing it over the treatment, so that knowing the blip functions allows one to calculate the optimal dynamic treatment by starting with maximizing the last blip function and iterating backwards in time. These models are semi-parametric since they rely on a parametric model of the blip function (at least in a SMART), but they leave the nuisance parameters unspecified. Structural nested mean models have been generalized to learning optimal rules that are restricted to only using a (counterfactual) subset of the past [21]; Section 6.5 in [25].

Qian and Murphy [26] and Zhao et al. [27] show that the estimation of the optimal dynamic treatment can be reduced to a classification problem. Rubin and van der Laan [28] and Zhang et al. [29] independently identify entire families of such reductions to classification. Zhao et al. [30] extend these results to the multiple time point setting.

1.2 Super-learning of an optimal dynamic treatment rule

Our proposed estimators of the V-optimal rule are based on sequential loss-based super-learning which involves the application of an ensemble method known as super-learning to fit each rule given an estimate of the optimal rule at future time points. The super-learner is defined by a family of candidate estimators, a risk for each candidate estimator, and selection among all candidate estimators based on a cross-validation based estimator of this risk. Some of these candidate estimators could be based on parametric models of the blip functions, while others are based on regression or classification machine learning algorithms. By previously established oracle inequality results on the cross-validation selector established in the above references, our results guarantee that in a SMART the super-learner will be asymptotically equivalent with the estimator selected by the oracle selector and thereby outperform any of the parametric model based estimators and any of the other estimators in the family of candidate estimators, under the assumption that none of the parametric models are correctly specified. If one of the parametric models is correctly specified, the proposed method achieves an almost parametric log n/n rate. In this manner, our sequential super-learner is at each stage doing an asymptotically optimal job in fitting the blip function relative to its user supplied class of candidate estimators. Past findings strongly suggest that this will also result in superior performance in most practical situations relative to *a priori* selecting one particular estimation procedure [31, 32]. We also outline how to develop a cross-validated targeted minimum loss-based estimator of the cross-validated risk.

This work is related to that appearing in Section 9 of Robins [21], which discusses a cross-validated selector to choose between different parametric working models for the blip function. Robins04 cites oracle inequalities in an earlier version of Dudoit and van der Laan (2005) [22], for choosing between working models for the blip function using cross-validation. He also introduces a double robust loss function that can be used to estimate the blip functions. We expand this work in several directions. First, we formally give conditions under which the oracle inequality will hold, including for several of our losses. This makes clear what is needed in order to interpret one's cross-validation selector as (nearly) optimal. Second, our selector considers weighted combinations of candidates rather than simply choosing the best of the candidate algorithms. Additionally, we provide several families of loss functions, thereby allowing the user more options for candidate algorithms for the cross-validation selector, including machine learning algorithms available in prepackaged software. We encourage the reader to review Section 9 of Robins [21], as it contains thought-provoking discussion that (for the most part) applies to the results in this paper.

For the sake of presentation, we focus on two-time point treatments in this article. In the appendix of an accompanying technical report (van der Laan and Luedtke [33]) we generalize these results to general multiple time point treatments.

1.3 Organization of article

Section 2 formulates the estimation problem. Section 2.1 defines the optimal rule as a causal parameter, and gives identifiability assumptions under which the causal parameter is

identified with a statistical parameter of the observed data distribution. Section 2.2 outlines how one can sequentially learn the optimal rule.

Section 3 describes a cross-validation selector known as a super-learner that can combine multiple estimation algorithms. Section 3.1 gives the oracle inequality for the second time point treatment. A finite sample oracle inequality is given to support the proposed methodology and the asymptotic implications of this inequality are described. Section 3.2 describes the super-learner for estimating the treatment rule at the first time point.

The cross-validation selector relies on a loss function for the optimal rule. We give examples of such loss functions in Section 4. Section 4.1 describes sequential estimation of blip functions based on any loss function that provides a valid estimate of a conditional mean (e.g. squared error loss), where the sign of the estimated conditional mean is used to estimate the optimal rule. Section 4.2 aims to directly estimate the optimal treatment by maximizing the sequential mean outcomes under the fitted rules, where the treatment at future time points is set according to the previously fit rule. Section 4.3 shows that maximizing an estimate of the mean outcome can be written as a weighted classification problem that includes a rich class of previously described classification loss functions. All loss functions presented in Section 4 rely on correct specification of the treatment/censoring mechanism, which is trivially true in an RCT without missingness. Double robust generalizations of the loss functions in Section 4 are presented in Appendix A. Section 5 gives conditions under which some of the loss functions presented in Section 4 satisfy the conditions of the oracle inequality for the cross-validation selector.

Section 6 outlines a cross-validated targeted minimum loss-based estimator (CV-TMLE) for the sequential mean outcome losses presented in Section 4.2. The CV-TMLE is a substitution estimator, and thus naturally respects the bounded nature of the data. Appendix D describes a non-sequential super-learner which directly uses the estimated mean outcome under the optimal rule to combine candidate estimators.

Section 7 presents the simulation methods. The simulations compare our proposed super-learner to single choices of machine learning algorithms and misspecified parametric models. Section 8 presents the simulation results. Section 9 closes with a discussion and directions for future work.

All proofs can be found in Appendix A.

2 Formulation of optimal dynamic treatment estimation problem

2.1 Parameter of interest

We use the same formulation for the parameter of interest as is given in Section 2 of van der Laan and Luedtke [34]. We restate important notation here, but refer to the other paper for a more thorough discussion of the context and assumptions which identify our statistical parameter with a causal parameter.

For a discrete-time process $X(\cdot)$, we will use the notation $\bar{X}(t) = (X(s): 0 \leq s \leq t)$, where $X(-1) = \emptyset$. Suppose we observe n i.i.d. copies $O_1, \ldots, O_n \in \mathcal{O}$ of $O = (\bar{L}(1), \bar{A}(1), Y) \sim P_0$, where $A(j) = (A_1(j), A_2(j))$, $A_1(j)$ is a binary treatment and $A_2(j)$ is an indicator of not being right-censored at "time" j, $j = 0, 1$. Each time point j has covariates $L(j)$ that precede treatment, $j = 0, 1$, and the outcome of interest is given by Y and occurs after time point 1. Let $\mathcal{M}$ be a statistical model that makes no assumptions on the marginal distribution $Q_{0,\, L(0)}$ of $L(0)$ and the conditional distribution $Q_{0,\, L(1)}$ of $L(1)$, given $A(0)$, $L(0)$, but might make assumptions on the conditional distributions $g_{0,\, A(j)}$ of $A(j)$, given $\bar{A}(j-1)$, $\bar{L}(j)$, $j = 0, 1$. We will refer to g_0 as the treatment/censoring mechanism, which can be factorized in a treatment mechanism g_{01} and censoring mechanism g_{02} as follows:

$$g_0(O) = \prod_{j=1}^{2} g_{0,1,A(j-1)}(A_1(j) \mid \bar{A}(j-1), \bar{L}(j)) g_{0,2,A(j-1)}(A_2(j) \mid A_1(j), \bar{A}(j-1), \bar{L}(j)).$$

Throughout this article we will automatically assume the positivity assumption:

$$P_0\left(0 < \min_{a_1 \in \{0,1\}} g_{0,A(0)}(a_1, 1 \mid L(0))\right) = 1, \qquad (1)$$
$$P_0\left(0 < \min_{a_1 \in \{0,1\}} g_{0,A(1)}(a_1, 1 \mid \bar{L}(1), A(0))\right) = 1.$$

The strong positivity assumption will be defined as this assumption (1), but where the 0 is replaced by a $\delta > 0$.

Let $(A(0), V(1))$ be a function of $(L(0), A(0), L(1))$, and let $V(0)$ be a function of $L(0)$. Let $V = (V(0), V(1))$. Consider dynamic treatment rules $V(0) \mapsto d_{A(0)}(V(0)) \in \{0, 1\} \times \{1\}$ and $(A(0), V(1)) \mapsto d_{A(1)}(A(0), V(1)) \in \{0, 1\} \times \{1\}$ for assigning treatment $A(0)$ and $A(1)$, respectively. Note that the rules for $A(0)$ and $A(1)$ are only a functions of $V(0)$ and $(A(0), V(1))$, respectively, and are restricted to set the observations to uncensored. Let $\mathcal{D}$ be the set of all such rules. We assume that $V(0)$ is a function of $V(1)$, but in the theorem below we indicate an alternative assumption. At times we abuse notation and let $a(0) \in \{0, 1\} \times \{1\}$ and $a(1) \in \{0, 1\} \times \{1\}$ represent the static rules at the first and second time points in which everyone receives treatment $a(0)$ or $a(1)$.

Define the distribution $P_{0,\, d}$ as the distribution with density

$$p_{0,d}(L(0), A(0), L(1), A(1), Y)$$
$$\equiv I(A = d(V)) q_{0,L(0)}(L(0)) q_{0,L(1)}(L(1) \mid L(0), A(0)) q_{0,Y}(Y \mid \bar{L}(1), \bar{A}(1)),$$

where $q_{0,\, L(0)}$, $q_{0,\, L(1)}$, and $q_{0,\, Y}$ are the densities for $Q_{0,\, L(0)}$, $Q_{0,\, L(1)}$, and $Q_{0,\, Y}$ and all densities are absolutely continuous with respect to some dominating measure μ. This probability distribution $P_{0,\, d}$ is the G-computation formula [1, 3, 35], and can be identified with the counterfactual distribution in which the treatment rule d is, possibly contrary to fact,

implemented for the entire population. We use the notation L_d (or Y_d, O_d) to mean the random variable with distribution $P_{0,d}$.

In this article we are concerned with estimation of the V-optimal rule defined as

$$d_0 = \arg\max_{d \in \mathcal{D}} E_{P_{0,d}} Y_d.$$

We note that d_0 is not necessarily unique, but that there exists a closed-form expression for a maximizer as we will show in the upcoming theorem. In this work our objective is to obtain an estimated rule d_n with the property that $E_{P_{0,d_n}} Y_{d_n}$ (approximately) maximizes the mean outcome under a treatment rule. Thus we only care about the optimal mean outcome under a treatment rule, and the non-uniqueness of the maximizer is irrelevant. Given positivity, $P_{0,d}$ is a well-defined mapping of the observed data distribution, and thus our statistical problem of interest is well-defined even if $P_{0,d}$ is not identified with the counterfactual distribution of interest.

The next theorem states an explicit form of the V-optimal individualized treatment rule d_0 as a function of P_0. We proved the theorem in van der Laan and Luedtke [34].

Theorem 1—Suppose V(0) is a function of V(1). A V-optimal rule d_0 can be represented as the following explicit parameter of P_0:

$$\bar{Q}_{20}(a(0), v(1)) = E_{P_0}(Y_{a(0),A(1)=(1,1)} \mid V_{a(0)} = v(1)) - E_{P_0}(Y_{a(0),A(1)=(0,1)} \mid V_{a(0)}(1) = v(1)),$$

$$d_{0,A(1)}(A(0), V(1)) = (I(\bar{Q}_{20}(A(0), V(1)) > 0), 1),$$

$$\bar{Q}_{10}(v(0)) = E_{P_0}(Y_{(1,1),d_{0,A(1)}} \mid V(0)) - E_{P_0}(Y_{(0,1),d_{0,A(1)}} \mid V(0)),$$

$$d_{0,A(0)}(V(0)) = (I(\bar{Q}_{10}(V(0)) > 0), 1),$$

where $a(0) \in \{0, 1\} \times \{1\}$. *If* $V(1)$ *does not include* $V(0)$, *but, for all* $(a(0), a(1)) \in (\{0, 1\} \times \{1\})^2$,

$$E_{P_0}(Y_{a(0),a(1)} \mid V(0), V_{a(0)}(1)) = E_{P_0}(Y_{a(0),a(1)} \mid V_{a(0)}(1)), \quad (2)$$

then the above expression for the V-optimal rule d_0 is still true.

The non-uniqueness of the first and second time point optimal rules occurs precisely on the sets where $\bar{Q}_{10}(v(0)) = 0$ and $\bar{Q}_{20}(a(0), v(1)) = 0$, respectively. One can vary treatment arbitrarily on these sets without affecting the mean outcome of the rule.

We refer to $\bar{Q}_{20}$ and $\bar{Q}_{10}$ as the blip functions. These functions are sometimes referred to as optimal blip-to- reference functions, where the reference treatment level is set to $A = 0$

(Section 4.1 of Chakraborty and Moodie, 2013) [36]. The blip functions can be easily interpreted under causal assumptions given in, e.g., Robins [1]. In particular, $\bar{Q}_{20}(a(0), v(1))$ represents the additive treatment effect for second time point treatment within the strata in which $V_{a(0)}(1) = v(1)$, in a world where everyone is assigned $a(0)$ at the first time point. The first time point blip function $\bar{Q}_{10}(v(0))$ represents the $V(0) = v(0)$ strata specific average treatment effect for the first time point treatment in a world where everyone receives the optimal treatment $d_{0,A(1)}$ at the second time point.

2.2 Sequential learning of the optimal rule

This section assumes the existence of loss functions which can be used to estimate the optimal rule. We show that such loss functions exist and give examples in Section 4.

Let g be a treatment/censoring mechanism and $d_{A(1)}$ be a second time point rule. For a function f_2 mapping $(A(0), V(1))$ to $\mathbb{R}$, let $L_{2,g}(f_2)$ map from $\mathcal{O}$ to $\mathbb{R}$. For a function f_1 mapping from $V(0)$ to $\mathbb{R}$, let $L_{1,d_{A(1)},g}(f_1)$ map from $\mathcal{O}$ to $\mathbb{R}$. We refer to $L_{2,g}$ and $L_{1,d_{A(1)},g}$ as loss functions. Define the risk minimizers

$$f_{20} = \arg\min_{f_2} P_0 L_{2,g_0}(f_2), \qquad (3)$$
$$f_{10} = \arg\min_{f_1} P_0 L_{1,d_{0,A(1)},g_0}(f_1),$$

where the minimimums are over all measurable functions f_2 of $(A(0), V(1))$ and f_1 of $V(0)$ and we note that $g = g_0$ in the above expression. Above we use the notation $Pt = E_P[t(O)]$ for a distribution P and a function t. These loss functions should have the property that the risk minimizers are latent functions which return the optimal rule:

$$d_{0,A(1)}(A(0), V(1)) = I(f_{20}(A(0), V(1)) > 0),$$
$$d_{0,A(0)}(V(0)) = I(f_{10}(V(0)) > 0).$$

By Theorem 1, one valid choice of f_{20} and f_{10} is given by $\bar{Q}_{20}$ and $\bar{Q}_{10}$. The equation above suggests that one can estimate the optimal rule using empirical risk minimization. First, one estimates the treatment/censor- ing mechanism g_0 with g_n. Next, one estimates f_{20} with the minimizer of an estimate of $P_n L_{2,g_n}(f_2)$ over f_2 in some class $\mathcal{F}_2$, where P_n is the empirical distribution of the n observations. An estimate $d_{n,A(1)}$ of $d_{0,A(1)}$ is given by plugging the estimate of f_{20} into the above expression. Now, taking this estimate of $d_{0,A(1)}$ as fixed, one can iterate this same procedure to estimate $d_{0,A(0)}$ by minimizing $P_n L_{1,d_{n,A(1)},g_n}(f_1)$ over f_1 in some class $\mathcal{F}_1$.

While appealing for its simplicity, empirical risk minimization often yields estimates which overfit f_{20} and f_{10} when the classes $\mathcal{F}_2$ and $\mathcal{F}_1$ are too large. At the same time, choosing $\mathcal{F}_2$ and $\mathcal{F}_1$ large is exactly what one wants given that the minima in are over all measurable

functions. Typically correctly specifying a small, e.g. parametric, $\mathcal{F}_2$ or $\mathcal{F}_1$ which contains f_{20} or f_{10} (or even contains an approximation thereof) will be challenging due to the complexity of real world problems.

In the next section we will present a cross-validated procedure known as super-learning which implements the general sequential procedure we have described in this section, but uses sample splitting to allow one to both allow one to estimate f_{20} and f_{10} over a large class while also avoiding overfitting. We support these claims with an oracle inequality.

3 Sequential super-learning of the optimal rule

We now present an ensemble method called super-learning that combines candidate estimators of the optimal rule at a particular time point into a single estimated rule for that time point. At each time point, the final estimator satisfies an oracle inequality stating that it will asymptotically perform at least as well as the best convex combination of candidates in the library in terms of loss-based dissimilarity under mild conditions. The super-learner methodology allows for data adaptive candidate estimators, by which we mean estimators that are consistent over a large semi-parametric model. Our super-learner can select the best resolution for a data set based on data adaptive and parametric candidate estimators.

3.1 Second time point

For the sake of presentation we will present these results for IPCW loss functions in an RCT without missingness, but the oracle inequalities for the double robust losses are straightforward extensions of the results in this section. At the end of this section we give examples of loss functions that satisfy the conditions for the oracle inequality derived from the blip function, mean performance, and weighted classification approaches.

We start by introducing the notation used in this section. Let $B_n \in \{0, 1\}^n$ denote a random split of the data into a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$ so that np_n of the elements in each realization of B_n have value 1 for some $p_n \in (0, 1)$. Let P^0_{n, B_n} and P^1_{n, B_n} denote the corresponding empirical distributions of these two complementary subsamples.

Let $\hat{f}_{2, j}$, $j = 1, \ldots, J$, denote an estimator of a latent function that takes a distribution P as input and outputs an estimate $\hat{f}_{2, j}(P)$ for which the indicator that $\hat{f}_{2, j}(P)(A(0), V(1))$ is positive gives an estimate of the optimal rule $d_{0, A(1)}$ evaluated at $(A(0), V(1))$. For example, $\hat{f}_{2, j}$ may return an estimate of $\bar{Q}_{20}$, though we will show that other latent functions satisfy this property.

We first give a general oracle inequality as presented in van der Laan, Polley, and Hubbard [10] for estimating $d_{0, A(1)}$. Let α_n fall in a grid G_n of $K(n)$ points on Δ_{J-1}, where Δ_{J-1} represents the $(J-1)$-simplex is the set of all $\alpha \in [0, 1]^J$ such that $\sum_j \alpha_j = 1$. We have the following finite sample result.

Theorem 2—Let L_{2,g_0} be some loss function that relies on g_0 which takes as input a function $f_2: \mathscr{A}(0) \times \mathscr{V}(1) \to \mathbb{R}$ and yields a function of O. Let $f_{20} = \arg\min_{f_2} P_0 L_{2,g_0}(f_2)$. Suppose that:

$$\sup_{f_2} \sup_{o \in O} \left| L_{2,g_0}(f_2)(o) - L_{2,g_0}(f_{20})(o) \right| < \infty, \quad (4)$$

$$\sup_{f_2} \frac{Var_{P_0}(L_{2,g_0}(f_2)(O) - L_{2,g_0}(f_{20})(O))}{E_{P_0}[L_{2,g_0}(f_2)(O) - L_{2,g_0}(f_{20})(O)]} < \infty. \quad (5)$$

where the supremums are over all measurable functions $f_2: \mathscr{A}(0) \times \mathscr{V}(1) \to \mathbb{R}$ and we take 0/0 = 0. For all $\alpha \in \Delta_{J-1}$, define $\hat{f}_{2,\alpha}(P) = \sum_{j=1}^{J} \alpha_j \hat{f}_{2,j}(P)$. For a fixed sample of size n, define:

$$\alpha_n = \arg\min_{\alpha \in G_n} E_{B_n} P^1_{n,B_n} L_{2,g_0}(\hat{f}_{2,\alpha}(P^0_{n,B_n})).$$

Then for all $n \in \mathbb{N}$ and $\lambda > 0$,

$$E_{P_0^n} E_{B_n} P_0 \left\{ L_{2,g_0}(\hat{f}_{2,\alpha_n}(P^0_{n,B_n}) - L_{2,g_0}(f_{20}) \right\} \leq (1+\lambda) E_{P_0^n} \min_{\alpha \in G_n} E_{B_n} P_0 \left\{ L_{2,g_0}(\hat{f}_{2,\alpha}(P^0_{n,B_n})) - L_{2,g_0}(f_{20}) \right\} + C(\lambda) \frac{\log K(n)}{n p_n},$$

where $C(\lambda) \geq 0$ is a constant that may rely on P_0 and P_0^n represents the distribution of the observed n i.i.d. draws from P_0.

The above theorem is a special case of Corollary 3.2 in van der Laan, Dudoit, and van der Vaart [9] so the proof is omitted. In this article we focus on *U*-fold cross-validation. In *U*-fold cross-validation, the data is split into *U* mutually exclusive and exhaustive sets of size approximately *n/U* uniformly at random. Each set is then used as the validation set once, with the union of all other sets serving as the training set. The fact that *n* may not be divisible by *U* so that the validation sets are not all exactly the same size will not matter asymptotically and will make little difference in finite samples.

We can choose G_n so that any point on the simplex can be arbitrarily well approximated by a point on the grid of polynomial size $K(n)$ asymptotically. Given a Lipschitz condition on the loss-based dissimilarity, the approximation error by using points on $K(n)$ instead of the entire simplex is asymptotically negligible. Such a Lipschitz condition will hold under

bounding conditions for all loss functions to be discussed in this paper except the mean outcome and weighted 0–1 losses discussed in Sections 4.2 and 4.3. We posit that the finite sample result on the grid G_n still gives a useful asymptotic result over the entire simplex under reasonable conditions even when this Lipschitz condition does not hold.

The limiting result is referred to as an oracle inequality because we asymptotically do as well as the oracle in selecting α (up to an almost parametric $O(\log n/n)$ term) in terms cross-validated loss-based dissimilarity averaged across training samples. The methodology shows that there is no need to *a priori* decide on a single loss function or algorithm to fit the optimal rule – simply including all candidate methods of interest in the super-learner library guarantees that we asymptotically do at least as well as the best of the algorithms in terms of the cross-validated loss-based dissimilarity resulting from the chosen L_{2,g_0}.

In Algorithm 1 we describe how to implement the super-learner algorithm using U-fold cross-validation for a given data set a collection of prediction algorithms $\hat{f}_{2,1},\ldots,\hat{f}_{2,J}$. Let L_{2,g_0} be a loss function satisfying the conditions of Theorem 2. For simplicity we assume that n is divisible by U. Rather than optimize for α_n over G_n, we recommend (approximately) optimizing over the entire simplex.

Algorithm 1

Super-learner estimation of $d_{0,A(1)}$

function SuperLearner($o_1,\ldots,O_n,\hat{f}_{2,1},\ldots,\hat{f}_{2,J}$)

Let F be a randomly ordered vector of length n containing n/U 1s, n/U 2s,…, n/U U's

Initialize an empty matrix X of dimension $n\times J$

for $u=1$ to U **do**

for $j=1$ to J **do**

Fit the estimate $f_{2,u,j}$ by running $\hat{f}_{2,j}$ on the set $\{O_i: F_i\neq u\}$

For all i such that $F_i=u$, let $X_{i,j}=f_{2,u,j}(A(0)_i, V(1)_i)$

Run an optimization routine to solve: $\alpha_n=\arg\min_{\alpha\in\Delta_{J-1}}\sum_{u=1}^{U}\sum_{i=F_i=u}L_{2,g_0}\left(\sum_{j=1}^{J}\alpha_j X_{i,j}\right)$

for $j=1$ to J **do**

Fit the estimate f_j by running $\hat{f}_{2,j}$ on the $\{O_i: i=1,\ldots,n\}$

Let $f_{\alpha_n}\equiv\sum_{j=1}^{J}\alpha_{n,j}f_j$

return $d_{n,A(1)}\equiv(a(0),v(1)\mapsto I(f_{\alpha_n}(a(0),v(1))>0)$..

For an observational study or an RCT with an unknown censoring mechanism, an estimate $g_{n,u}$ of g_0 can be estimated each training sample $u=1,\ldots,U$. We then let:

$$\alpha_n = \arg \min_{\alpha \in \Delta_{J-1}} \sum_{u=1}^{U} \sum_{i=F_i=u} L_{2,g_n,u}\left(\sum_{j=1}^{J} \alpha_j f_{2,u,j}\right)(O_i). \quad (6)$$

3.2 First time point

The approach for estimating the optimal rule at the first time point is analogous to the second time point, with the caveat that it takes an estimate of $d_{0,A(1)}$ as a nuisance function. To incorporate the estimate of the nuisance function, we suggest using the same approach used to incorporate an estimate of g_0 when it is unknown as we do in (6). In particular, this means estimating the nuisance function $d_{0,A(1)}$ on training set u and using this estimate of the nuisance function to obtain an estimate of a latent function at the first time point based on training set u for each algorithm j. One can then learn the convex combination similarly to as in (6), and apply this convex combination to the candidates learned on the full data set, which take an estimate of $d_{0,A(1)}$ based on the entire data set as nuisance function. The rate of convergence of the estimated first time point rule to $d_{0,A(0)}$ will be upper bounded by the rate of convergence of the estimated second time point rule to $d_{n,A(1)}$, see Theorem 1 of van der Laan and Dudoit [8] for a detailed exposition.

To estimate the nuisance function $d_{0,A(1)}$, we suggest using the super-learner procedure presented in Algorithm 1, leading to a nested cross-validation procedure. In terms of runtime, this can cost up to a factor of U. If this is a concern, one can simply use the estimate of $d_{n,A(1)}$ resulting from the entire data set as nuisance function for all folds. Such a practice is not generally advisable because it invalidates the oracle inequality and necessitates empirical process conditions on the candidates. In general we look to avoid such conditions since they limit the data adaptivity of the estimators. Thus we believe that an honest cross-validation scheme in which the candidate estimators are only functions of the training samples is valuable for estimating rules in practice. In a forthcoming work we will present an honest cross-validation procedure which does not require an additional runtime factor of U.

4 Sequential loss functions for the V-optimal rule

We will derive three classes of loss functions for the V-optimal rule. Having a collection of possible loss functions will allow us to provide our super-learner with a variety of candidate algorithms. The first is based on estimating the blip function. The second aims to directly maximize an estimate of the mean outcome under the optimal rule. The third is based on previously described weighted classification approaches [27–29]. In the next section we will give conditions under which these loss functions satisfy the conditions for the oracle inequality given in Theorem 2.

We will generally assume that $d_{A(1)} = d_{0,A(1)}$ when stating results in this section. Nonetheless, it is straightforward to show that the first time point loss functions in this section are valid for estimating the optimal fitted rule given correct specification of the

treatment/censoring mechanism under the constraint that the second time point treatment must follow the possibly suboptimal rule $d_{A(1)}$.

For presentation purposes, all of the loss functions described in this section are inverse probability of censoring weighted (IPCW) loss functions. That is, these loss functions are correct if the treatment/censoring mechanism is specified correctly, which is trivially true in an RCT without missingness. We denote a (possibly misspecified) treatment/censoring mechanism estimate with g. We take $g_{A(0)}$ and $g_{A(1)}$ to be the resulting first and second time point treatment/censoring mechanisms.

In the appendix we present double robust versions all of the loss functions and theorems given in this section so that the loss functions will be correct if either the treatment/censoring mechanism is correctly specified or if particular conditional expectations of the outcome are correctly specified. Because the IPCW versions of the theorems are special cases of the double robust versions, we only give proofs for the double robust case in the appendix.

The simplicity of the IPCW formulations comes at the expense of robustness and efficiency. In an observational study or an RCT with missingness, one must also estimate the treatment and/or censoring mechanism g_0. The rate of convergence of the final estimate when g_0 is estimated will be upper bounded by the rate at which the estimate g converges to g_0. For this reason we suggest using the more efficient double robust inverse probability of censoring weighted (DR-IPCW) loss function presented in the appendix. For double robust loss functions the rate of convergence of the estimate will be upper bound by a product of the rate of convergence of the treatment/censoring mechanism estimate and the outcome regression estimate.

4.1 Blip functions

We first give a formulation which aims to sequentially learn the blip functions at each time point. That is, we aim to sequentially learn the V-strata-specific average treatment effect at each time point. For the second time point, we find this strata-specific average treatment effect under the counterfactual distribution in which the first time point treatment is fixed at $a(0) \in \{0, 1\} \times \{1\}$. For the first time point, we find this under the counterfactual distribution in which the second time point follows the estimated second time point rule.

The blip function was the target of estimation in Robins [21], though the blip was estimated using structural nested models instead of using our IPCW or augmented IPCW loss functions. Robins gives an alternative loss function for estimating the blip functions in a sequential decision problem in his Corollary 9.2. He subsequently discusses challenges for using this loss function sequentially due to the fact that one cannot estimate the two decision rules simultaneously. Nonetheless, the sequential procedure can still yields valid losses, as we show below and he shows in his Corollary 9.2.

Define

$$D_2(g)(O) = A_2(1)\frac{2A_1(1)-1}{g_{A(1)}(O)}Y. \quad (7)$$

The choice of $D2(g)$ is motivated by the fact that $E_{P_{0,a(0)}}[D_2(g_0)(O_{a(0)})\Big|V_{a(0)}(1)=v(1)] = \bar{Q}_{20}(a(0),v(1))$ Let $P_{0,\,a(0)}$ denote the static-intervention specific G-computation distribution $P_{0,\,a(0)}$ and $O_{a(0)}$ represents a counterfactual observation under this distribution. Let $L^F_{2,D_2(g)}(\bar{Q}_2)(O)$ denote a valid loss function for estimating $E_{P_{0,a(0)}}[D_2(g_0)\Big|V_{a(0)}(1)=v_{a(0)}(1))]$, in the sense that

$$(a(0),v(1)) \mapsto E_{P_{0,a(0)}}[D_2(g)(O_{a(0)})\Big|V_{a(0)}(1)=v(1)]$$

minimizes

$$\sum_{\tilde{a}(0)\in\{0,1\}\times\{1\}} E_{P_{0,\tilde{a}(0)}}\left[L^F_{2,D_2(g)}(\bar{Q}_2)(O_{\tilde{a}(0)})\right] \quad (8)$$

over all measurable functions $\bar{Q}_2$ of $a(0)$ and $v(1)$. Because the minimum is over all measurable functions, one can split the above sum and minimize the expected loss (risk) first for $\tilde{a}=(0,1)$, and then for $\tilde{a}=(1,1)$. At the end of this section we provide two examples of loss functions satisfying this property. In fact, one can construct a valid $L^F_{2,D_2(g)}$ from any loss that can be used to fit a conditional mean. To identify the resulting risk function with the observed data distribution, we apply the IPCW mapping [8]:

$$L_{2,g}(\bar{Q}_2)(O) = \frac{A_2(0)}{g_{A(0)}(O)}L^F_{2,D_2(g)}(\bar{Q}_2)(O), \quad (9)$$

Note that we inverse weight by the entire first time point treatment/censoring mechanism, not just the censoring mechanism at the first time point. We will use the sign of the $\bar{Q}_2$ which minimizes $L_{2,\,g}$ to estimate $d_{0,\,A(1)}$. The use of the IPCW mapping is motivated by the fact that $E_{P_0}[L_{2,g}(\bar{Q}_2)(O)]$ is equal to the expression in (8).

For a given rule at the second time point $d_{A(1)}$, define

$$D_1(g)(O) = A_2(0)\frac{2A_1(0)-1}{g_{A(0)}(O)}Y. \quad (10)$$

The choice of $D_1(g)$ is motivated by the fact that $E_{P_0, d_{0,A(1)}}[D_1(g_0)(O) \mid V(0) = v(0)] = \bar{Q}_{20}(v(0))$. Let $L^F_{1,D_1(g)}$ be some loss that satisfies:

$$E_{P_0, d_{A(1)}}[D_1(g)(O_{d_{A(1)}}) \mid V(0) = \cdot] = \arg\min_{Q_1} P_{0,d_{A(1)}} L^F_{1,D_1(g)}(\bar{Q}_1), \quad (11)$$

where $P_{0,d_{A(1)}}$ represents the post-intervention distribution corresponding with the dynamic intervention $d_{A(1)}$ and $O_{d_{A(1)}}$ represents a counterfactual observation under this distribution. Our proposed loss function is obtained by applying the IPCW mapping to the above loss function:

$$L_{1,d_{A(1)},g}(\bar{Q}_1)(O) = \frac{I(A(1) = d_{A(1)}(A(0), V(1)))}{g_{A(1)}(O)} L^F_{1,D_1(g)}(\bar{Q}_1). \quad (12)$$

We now state a theorem that gives conditions under which the above loss functions allow us to learn the optimal rule d_0.

Theorem 3—Suppose the positivity assumption holds at g_0. Then:

$$P_0\left\{L_{2,g_0}(\bar{Q}_2) - L_{2,g_0}(\bar{Q}_{20})\right\} = \sum_{a(0)} P_{0,a(0)}\left(L^F_{2,D_2(g_0)}(\bar{Q}_2) - L^F_{2,D_2(g_0)}(\bar{Q}_{20})\right),$$
$$P_0\left\{L_{1,d_{0,A(1)},g_0}(\bar{Q}_1) - L_{1,d_{0,A(1)},g_0}(\bar{Q}_{10})\right\} = P_{0,d_{0,A(1)}}\left(L^F_{1,D_1(g_0)}(\bar{Q}_1) - L^F_{1,D_1(g_0)}(\bar{Q}_{10})\right), \quad (13)$$

where the sum is over $a(0) \in \{0, 1\} \times \{1\}$. As a consequence:

$$\bar{Q}_{20} = \arg\min_{Q_2} P_0 L_{2,g_0}(\bar{Q}_2), \quad (14)$$
$$\bar{Q}_{10} = \arg\min_{Q_1} P_0 L_{1,d_{0,A(1)},g_0}(\bar{Q}_1).$$

A double robust generalization of the above theorem appears with proof in the appendix. We will refer to the quantities in (13) as loss-based dissimilarities for L_{2,g_0} and $L_{2,d_{0,A(1)},g_0}$, which represent the difference between the P_0-expected loss (risk) at a candidate function and the P_0-expected loss (risk) at the true parameter value. The loss-based dissimilarity is defined analogously for general losses.

The expressions in (14) make L_{2,g_0} and $L_{2,d_{0,A(1)},g_0}$ valid losses. Even if the estimated rule is not the optimal rule, one can show that the blip function at the first time point will maximize the mean outcome under the constraint of the suboptimal second time point rule. In an observational study, we have access to an empirical rather than the true observed data distribution. Hence it may be important to consider the smoothness of $L^F_{2,D_2(g)}$ and $L^F_{1,D_1(g)}$ in the neighborhood of the minimizers in (8) and (11) so that reasonable estimation of the sequential risk functions is possible. We close this section with an examples of a loss function which satisfies the conditions of Theorem 3.

Example 1—Squared error loss.

$$L^F_{2,D_2(g),MSE(\bar{Q}_2)}(o) = h_2(a(0),v(1))[D_2(g)(o) - \bar{Q}_2(a(0),v(1))]^2,$$
$$L^F_{1,D_1(g),MSE(\bar{Q}_1)}(o) = h_1(a(0))(D_1(g)(o) - \bar{Q}_1 v(0)))^2,$$

where h_2 and h_1 represent positive user-supplied weight functions of $(a(0), v(1))$ and $v(0)$, respectively. By Theorem 3

$$P_0\left\{L_{2,g_0,MSE(\bar{Q}_2)} - L_{2,g_0,MSE(\bar{Q}_{20})}\right\} = \sum_{a(0)} P_0\left\{h_2(\bar{Q}_2 - \bar{Q}_{20})^2(a(0),V_{a(0)}(1))\right\},$$
$$P_0\left\{L_{1,d_{0,A(1)},g_0,MSE(\bar{Q}_1)} - L_{1,d_{A(1)},g_0,MSE(\bar{Q}_{10})}\right\} = P_0\left\{h_1(\bar{Q}_1 - \bar{Q}_{10})^2(V(0))\right\}.$$

□

4.2 Performance of rule

We now describe a risk function which sequentially targets the performance of the fitted rule in terms of mean outcome. By definition, $d_0 = \arg\max_{d \in \mathcal{D}} E_{P_0,d}Y_d$. It follows immediately that $-E_{P_0}Y_d$ is a valid risk function for a candidate rule d. In van der Laan and Luedtke [34] we discuss two estimates of $-E_{P_0}Y_d$. Rather than restate these results, we state a single theorem which summarizes these findings and refer the reader to van der Laan and Luedtke [34] for a thorough discussion of the proposed methods.

Define:

$$\widetilde{L}_{2,g}(d_{A(1)})(O) = -\frac{A_2(0)}{g_{A(0)}(O)}\frac{I(A(1) = d_{A(1)}(A(0),V(1)))}{g_{A(1)}(O)}Y. \quad (15)$$

Let $d_{A(1)}$ be a treatment rule for the second time point. Define:

$$\widetilde{L}_{1,d_{A(1)},g}(d_{A(0)})(O) = -\frac{I(A(1)=d_{A(1)}(A(0),V(1)))}{g_{A(1)}(O)}\frac{I(A(0)=d_{A(0)}(V(0)))}{g_{A(0)}(O)}Y.$$

Theorem 4—Suppose the positivity assumption holds at g_0. Then:

$$P_0\left\{\widetilde{L}_{2,g_0}(d_{A(1)})-\widetilde{L}_{2,g_0}(d_{0,A(1)})\right\}=\sum_{a(0)}P_0 I(d_{A(1)}\neq d_{0,A(1)})\left|\bar{Q}_{20}\right|(a(0),V_{a(0)}),$$
$$P_0\left\{\widetilde{L}_{1,d_{0,A(1)},g_0}(d_{A(0)})-\widetilde{L}_{1,d_{0,A(1)},g_0}(d_{0,A(0)})\right\}=P_0 I(d_{A(0)}\neq d_{0,A(0)})\left|\bar{Q}_{10}\right|(V(0)),$$

where the sum is over $a(0)\in\{0,1\}\times\{1\}$. It follows that the minimizer of $P_0\widetilde{L}_{2,g_0}(d_{A(1)})$ over rules $d_{A(1)}$ is an optimal second time point rule, and the minimizer of $P_0\widetilde{L}_{1,d_{0,A(1)},g_0}(d_{A(0)})$ over rules $d_{A(0)}$ is an optimal first time point rule.

A double robust generalization of the above theorem appears with proof in the appendix.

4.3 Weighted classification

We now show that maximizing $E_{P_0}Y_d$ can be viewed as a risk minimization problem resulting from using a weighted 0-1 loss function. This result is a longitudinal extension to that of Zhang et al. [29]. We then show that a rich class of smooth surrogate loss functions can be used to improve computational tractability, a result which is slightly more general than an earlier result in Zhao et al. [30]. We will use the definitions of D_1 and D_2 from the Section 4.1.

Let $\mathbb{R}\rightarrow\mathbb{R}$ represent the function $Z(x)=I(x\geq 0)$. Define:

$$K_{2,g}(O)=\frac{A_2(0)}{g_{A(0)}(O)}D_2(g)(O),$$
$$\hat{L}_{2,g}(d_{A(1)})(O)=\left|K_{2,g}(O)\right|I(d_{A(1)}(A(0),V(0))\neq(Z\circ K_{2,g}(O),1)),$$

where $\circ$ denotes function composition. For some fixed $d_{A(1)}$, define:

$$K_{1,d_{A(1)},g}(O)=\frac{I(A(1)=d_{A(1)}(A(0),V(1)))}{g_{A(1)}(O)}D_1(d_{A(1)},g))(O),$$
$$\hat{L}_{1,d_{A(1)},g}(d_{A(0)})(O)=\left|K_{1,d_{A(1)},g}(O)\right|I\left(d_{A(0)}(V(0)\neq(z\circ K_{1,d_{A(1)},g}(O),1)\right).$$

The following theorem shows that the optimal rule can be learned through a sequential classification problem using $Z\bigcirc\ K_{2,g}(O)$ and $Z\circ K_{1,d_{A(1)},g}(O)$ as outcomes and weighted

0–1 loss functions with weights $|K_{2,g}|$ and $\left|K_{1,d_{A(1)},g}\right|$, where the weights respectively do not rely on $d_{A(1)}$ or $d_{A(0)}$, i.e. the current rule the routine aims to learn.

Theorem 5—Suppose the positivity assumption holds at g and g_0. Then for any $(d_{A(0)}, d_{A(1)}) \in \mathcal{D}_2$:

$$\hat{L}_{2,g}(d_{A(1)}) = \tilde{L}_{2,g}(d_{A(1)}) + C_{2,g},$$
$$\hat{L}_{1,d_{A(1)},g}(d_{A(0)}) = \tilde{L}_{1,d_{A(1)},g}(d_{A(0)}) + C_{1,d_{A(1)},g},$$

where $C_{2,g}(O)$ and $C_{1,d_{A(1)},g}(O)$ do not rely on $d_{A(1)}$ or $d_{A(0)}$, respectively. It follows that $\hat{L}_{2,g}$ and $\hat{L}_{1,d_{0,A(1)},g}$ are valid loss functions for sequentially estimatini $d_{0,A(1)}$ and $d_{0,A(0)}$ if $g = g_0$.

A double robust generalization of the above theorem appears with proof in the appendix. The above theorem shows that the weighted classification losses yield the same loss-based dissimilarities as the corresponding mean performance based losses.

It is well known that risk functions from the 0-1 loss can be difficult to optimize in practice. For this reason people often use convex surrogate loss functions which yield an easier optimization problem. A detailed examination of the theoretical properties of these loss functions is presented in Bartlett et al. [37]. We will use the same results to establish that a convex surrogate of the indicator function used in the definition of $\hat{L}_{2,g}$ yields a valid loss function. Let be some convex function that is differentiable at 0 with $\phi'(0) < 0$. Define the losses

$$L_{2,\phi,g}(f_2)(O) = \left|K_{2,g}(O)\right| \phi\left(f_2(A(0), V(1))(2Z \circ K_{2,d_{A(1)},g}(O) - 1)\right)$$
$$L_{1,\phi,d_{A(1)},g}(f_1)(O) = \left|K_{1,d_{A(1)},g}(O)\right| \phi\left(f_1(V(0))(2Z \circ K_{1,d_{A(1)},g}(O) - 1)\right).$$

We show in Appendix B that the f_2^* which minimizes $P_0 L_{2,\phi,g_0}(f_2)$ over f_2 is a latent function for an optimal rule, i.e. $I(f_2^* > 0)$ is an optimal rule, provided $0 < E_{P_0}\left|K_{2,g_0}(O)\right| < \infty$. A sufficient condition for $E_{P_0}\left|K_{2,g_0}(O)\right| < \infty$ is that g_0 satisfies the strong positivity assumption and Y is uniformly bounded. We also give the double robust generalization of this result. Because nonnegatively weighted linear combinations of convex functions are convex, $L_{2,\phi,g}$ is necessarily convex. Thus the proposed procedure yields an empirical risk that is easy to minimize via convex optimization techniques. A similar proof shows the validity of the loss for the first time point rule.

The results in this section are closely related to those of Zhao et al. [30], namely their backward outcome weighted learning (BOWL) estimation procedure. Unlike the procedure Zhao et al. [30], the procedure in the appendix allow for the use of a double robust type rather than an inverse probability weighted type loss function. Nonetheless, the objective of the current section is simply to obtain more candidate estimators or possible loss functions to use with the super-learner algorithm. Thus we refer the reader to Zhao et al. [30] for an overview of some of the theoretical properties of the inverse weighted version of this loss function and give no further details here.

We close this section with two examples of valid weighted surrogate loss functions, and refer the reader to Bartlett et al. [37] for more examples. One can verify that ϕ is convex and differentiable at 0 in both of these examples.

Example 2—Weighted log loss: $\phi(x) = \log(1 + e^{-x})$. Then:

$$L_{2,\phi,g}(f_2)(O) = \left|K_{k,g}(O)\right| \log\left(1 + e^{-f_2(A(0),V(1))(2Z \circ K_{2,g}(O) - 1)}\right),$$
$$L_{1,\phi,d_{A(1)},g}(f_1)(O) = \left|K_{1,d_{A(1)},g}(O)\right| \log\left(1 + e^{-f_1(V(0))(2Z \circ K_{1,d_{A(1)},g}(O) - 1)}\right).$$

□

Example 3—Weighted hinge loss: $\phi(x) = \max(1 - x, 0)$. Then:

$$L_{2,\phi,g}(f_2)(O) = \left|K_{2,g}(O)\right| \max\left(1 - f_2(A(0),V(1))[2Z \circ K_{2,g}(O) - 1], 0\right),$$
$$L_{2,\phi,g}(f_1)(O) = \left|K_{1,d_{A(1)},g}(O)\right| \max\left(1 - f_1(V(0))[2Z \circ K_{1,d_{A(1)},g}(O) - 1], 0\right).$$

Zhao et al. [30] focus on this loss function in their paper, though note that the method extends to other surrogates ϕ for the indicator function. □

5 Examples of loss functions satisfying the conditions of the oracle inequality

Theorem 2 gave an oracle inequality showing that we asymptotically estimate the best candidate rule at each time point given our sample, subject to the implementation of possibly suboptimal rules at future time points. The oracle inequality relies (4), i.e. that the loss is uniformly bounded, and (5), i.e. that the variance of the loss can be upper bounded by a constant times its expectation. We now give examples of loss functions which satisfy these conditions. Each of the below examples makes use of a subset of following assumptions (each example specifies the subset it uses). The fourth assumption is only used in Example 4 and is discussed there.

A1 $Pr_{P_0}(|Y| < M) = 1$ for some $M < \infty$.

A2 The strong positivity assumption holds at g_0 for some $\delta > 0$.

A3 Each of the estimators in the candidate library produces estimates of uniformly bounded range, where the uniformity is over input distributions P.

A4 There exists some constant $c > 0$ that may rely on P_0 such that $\left|\bar{Q}_{20}(A(0), V_{a(0)}(1))\right| \geq cE_{P_0}[Y^2_{a(0),a(1)} | V_{a(0)}]$ almost surely with respect to the distribution in which the first treatment is set to $a(0)$ for $a(0), a(1) \in \{0, 1\} \times \{1\}$.

Example 1 (continued)

Squared error loss. We consider the unweighted case so that $h_2 = 1$. If 1, 2, and 3 then $\left|L_{2,g_0,MSE}\right|$ is uniformly bounded and thus satisfies (4). For all f_2, it can be shown that:

$$Var_{P_0}(L_{2,g_0,MSE}(f_2) - L_{2,g_0,MSE}(\bar{Q}_{20})) \leq E_{P_0}\left(L_{2,g_0,MSE}(f_2) - L_{2,g_0,MSE}(\bar{Q}_{20})\right)^2$$
$$\leq M_1 E_{P_0}\left[L_{2,g_0,MSE}(f_2) - L_{2,g_0,MSE}(\bar{Q}_{20})\right],$$

where $M_1 = \sup_{f_2} \sup_{o \in \mathcal{O}} \left(2D_2(g_0)(o) - (f_2 + \bar{Q}_{20})(a(0), v(1))\right)^2 \geq 0$ is bounded by the stated assumptions. Thus the condition in (5) holds. The alternative squared error blip loss developed in Section 9 of Robins has a certain appeal because it is bounded even if A2 does not hold. Future simulation studies will help shed light on how these two losses perform in practice. □

Example 2 (continued)

Weighted log loss. Suppose 1, 2, and 3. It follows that $\left|K_{2,g_0}\right|$ is almost surely bounded.

These conditions immediately show that (4) holds. The result is obvious if $E_{P_0}\left|K_{2,g_0}\right| = 0$, so suppose $E_{P_0}\left|K_{2,g_0}\right| > 0$. To show that (5) holds, one can use a similar change of measure argument as the one applied in the proof of our Theorem 9 in the appendix to account for the weighting and apply Corollary 5.4 in van der Laan et al. [9] to:

$$\phi(x) = \log(1 + e^{-x}) = -\log\left(\frac{1}{1 + e^{-x}}\right).$$

□

Example 4

Mean performance. Define:

$$L_{2,g_0}(f_2)(O) = -\frac{A_2(0)}{g_{A(0)}(O)}\frac{I(A(1)=d_{A(1)}(A(0),V(1)))}{g_{A(1)}(O)}Y,$$

where $d_{A(1)}(A(0), V(1))) = I(f_2(A(0), V(1)) > 0)$ and we have modified the definition in (15) so that L_{2,g_0} depends directly on the latent function. Suppose A1, A2, and A4. The loss-based dissimilarity representation in Theorem 4 shows that $\tilde{L}_{2,g_0}$ satisfies (4). We show that the stated conditions suffice for (5) in Appendix C.

Assumption A4 can be viewed as a margin condition that ensures that the classification problem is not too difficult. In particular, it requires that the strata-specific treatment effect be larger than both of the expected squared outcomes under the counterfactual distributions where $a(1)$ is fixed without censoring. For binary Y, this means that the absolute average treatment effect in each strata of $V_{a(0)}$ be larger than some fixed proportion of the counterfactual prevalence of the outcome in strata of $V_{a(0)}$ when we set $a(0)$ and $a(1)$. □

6 CV-TMLE of risk

The empirical risk estimates resulting from the loss functions provided in Section 4 are valid in the sense that they are minimized at the true optimal treatment regime. Nonetheless, the empirical risk resulting from the given loss functions (and the double robust losses presented in the appendix) are not substitution estimators and thus can fail to respect a key constraint of the model: the fact that the risk is bounded. To improve finite sample performance, we propose using a CV-TMLE based estimate of risk. The CV-TMLE is a substitution estimator and thus naturally respects the bounded nature of our data. The CV-TMLE was originally proposed in Zheng and van der Laan [38]. Diaz et al. [39] use a CV-TMLE to estimate the risk of the causal dose response curve. In van der Laan and Luedtke [34] we presented a CV-TMLE for the cross-validated mean outcome under a fitted rule. Here we present a sequential CV-TMLE that estimates the risks resulting from Theorem 4. In Appendix D we present a nonsequential CV-TMLE that aims to directly maximize the mean outcome under the fitted two time point rule. Robins [21] presents a non-sequential IPCW loss for sequential decisions at the end of his Section 9, though does not provide a double robust version for the multiple time point case.

To distinguish between the convex combinations for super-learner at the first and second time points in this section, we will use the notation $a_{A(k)}$ for the convex combination used at time k, $k = 0, 1$.

Suppose we use the sequential negative mean performance risk function from Section 4.2 and conditions hold so that the risk is bounded. Consider selecting the convex combination $a_{A(1)}$ and $a_{A(0)}$ for the super-learner presented in the previous section when g_0 is known (e.g., in an RCT without missingness). Suppose the outcome Y is bounded. While the

empirical risk is root-n consistent under conditions (and the double robust empirical risk is even asymptotically efficient under conditions), the given risk estimates may not respect the bounded nature of the data in finite samples.

Given an $\alpha_{A(1)}$, we can estimate the risk for the second time point rule indexed by this $\alpha_{A(1)}$. The CV- TMLE for the second time point is identical to the CV-TMLE presented in Appendix B.2 of van der Laan and Luedtke [34], with the exception that the covariate for ε_2 is replaced by

$$\frac{A_2(0)I(A(1) = I\left(\sum_j \alpha_{A(1),j} f_{2,u,j}(A(0), V(1)) > 0\right)}{g_0(O)},$$

and the covariate for ε_1 is replaced by $A_2(0)/g_{0,A(0)}(O)$. The conditions for the validity of the resulting risk estimate are not presented here, but are analogous to those presented in Diaz et al. [39]. The CV-TMLE has the same double robustness and asymptotic efficiency properties as the cross-validated empirical mean of the double robust loss. For more details, we refer the reader to van der Laan and Luedtke [34].

Fitting the rule at the first time point is similar, with the covariate for ε_2 replaced by

$$\frac{I\left(A(0) = I(\sum_j \alpha_{A(0),j} f_{1,u,j}(V(0)) > 0)\right) I\left(A(1) = d_{n,u,A(1)}(A(0), V(1))\right)}{g_0(O)},$$

where $d_{n,u,A(1)}$ is a nuisance parameter for the second time point rule learned only on training sample u. The covariate for ε_1 is then given by $I(A(0) = I(\sum_j \alpha_{A(0),j} f_{1,u,j}(V(0)) > 0))/g_{0,A(0)}(O)$.

One could in fact derive CV-TMLEs of the risks resulting from any of the losses presented in this paper.

7 Simulation methods

Section 7.1 and Section 7.2 respectively introduce the data and methods for estimating the optimal rule d_0 in the one and two time point cases.

7.1 Single time point

We start by presenting two single time point simulations. In an accompanying technical report we directly describe the single time point problem. Here, we instead note that a single time point optimal treatment is a special case of a two time point treatment when only the second treatment is of interest. In particular, we can see this by taking $L(0) = V(0) = \varnothing$, estimating $\bar{Q}_{2,0}$ without any dependence on $a(0)$, and correctly estimating $\bar{Q}_{1,0}$ with the constant function zero. We can then let $I(A(0) = d_{n,A(0)}(V(0))) = 1$ for all $A(0), V(0)$ wherever the indicator appears in our calculations. Because the first time point is not of interest, we only describe $\bar{Q}_{2,0}$ and the second time point treatment/censoring mechanism for

this simulation. We refer the interested reader to our accompanying technical report for a thorough discussion of the single time point case.

7.1.1 Data—This is the same simulation as that presented in Section 8.1.1 of van der Laan and Luedtke [34]. We simulate 1,000 data sets of 1,000 observations from a randomized controlled trial without missingness. We let $L(1) = (W_1,..., W_4)$. The data is generated as follows:

$$W_1, W_2, W_3, W_4 \mid A(0) \overset{iid}{\sim} N(0, 1),$$
$$A_1(1) \mid \bar{L}(1), A(0) \sim Bern(1/2),$$
$$A_2(1) \mid A_1(1), A(0), \bar{L}(1), \sim Bern(1),$$
$$\text{logit} E_{P_0}\left[Y \mid \bar{A}(1), \bar{L}(1), H = 0\right] = 1 - W_1^2 + 3W_2(1) + A_1(1)(5W_3^2 - 4.45),$$
$$\text{logit} E_{P_0}\left[Y \mid \bar{A}(1), \bar{L}(1), H = 1\right] = -0.5 - W_3 + 2W_1W_2 + A_1(1)(3\left|W_2\right| - 1.5),$$

where Y is a Bernoulli random variable and H is an unobserved $Bern(1/2)$ variable independent of $\bar{A}(1)$, $\bar{L}(1)$.

Static treatments (treating everyone or no one at the second time point) have approximately the same mean outcome of 0.464. The optimal rule has mean outcome $E_{P_0} Y_{d_0} \approx 0.536$ when $V(1) = W_3$ and the optimal rule has mean outcome $E_{P_0} Y_{d_0} \approx 0.563$ when $V(1) = (W_1, W_2, W_3, W_4)$.

7.1.2 Estimation methods—We assume that the treatment/censoring mechanism is known. For ease of interpretation, we consider two estimates of $E_{P_0}[Y \mid \bar{A}(1), W]$: (i) a naive estimate of 1/2 for all $A(1), W$, and (ii) the true conditional expectation $E_{P_0}[Y \mid \bar{A}(1), W]$. We note that (i) is slightly different from an IPCW estimator in that it contains a term which stabilizes the inverse weighted outcome term in the (cross-validated) empirical or CV-TMLE estimate of risk. This stabilized approach should do slightly better in our simulation since the conditional mean of Y given $\bar{A}(1), W$ is approximately centered around 1/2. In practice we always recommend using a double robust approach, even if just an intercept-only best guess of the conditional mean as we do here. One can always (approximately) center the outcome by subtracting 1/2. This turns out to be equivalent to misspecifying $E_{P_0}[Y \mid \bar{A}(1), W]$ to be the constant function 1/2.

We also use super-learner to estimate $\bar{Q}_{2,0}$. Table 1 shows the methods used from the SuperLearner package in R [40] and the corresponding estimating methodology with which they were estimated. The multivariate adaptive regression splines algorithm was only used for $V = W_1,\ldots, W_4$. We separately consider the candidates generated according to the

squared error and surrogate log loss functions, and also consider a candidate library that includes both the squared error and surrogate log loss function methods.

To generate convex combinations of predictors we maximize the CV-TMLE or CV-DR-IPCW estimates of mean outcome (see [34] for a description of the estimating equation based CV-DR-IPCW estimator). We approximate solutions to the resulting optimization problems using the Subplex routine in the nloptr package in R [41]. We use thirty starting values selected randomly from the simplex to avoid sensitivity to intitial conditions, and also include the selection of α based on the weighted log loss criterion as an initial value. We also consider minimizing the cross-validated empirical risk functions derived from the squared error and weighted log loss functions. We do not truncate the latent functions, though we note only the empirical MSE blip function estimates can be unbounded, and this should not cause problems in our data set because the outcome is bounded. We compare the mean outcome under the rules generated by several combinations of candidate libraries and criteria for choosing the convex combination.

To evaluate the performance of the described methods we will use the mean performance of the estimated rule as a criterion, which is given by $E_{P_0} Y_{d_n}$ for a given rule d_n. We estimate E_{P_0} using 10^6 Monte Carlo simulations.

7.2 Two time points

We now show that our proposed method can sequentially learn a rule with good performance.

7.2.1 Data—This is the same simulation as that presented in Section 8.1.2 of van der Laan and Luedtke [34]. We again simulate 1,000 data sets of 1,000 observations from a randomized controlled trial without missingness. The observed variables have the following distribution:

$$
\begin{aligned}
&L_1(0), L_2(0) \overset{iid}{\sim} Unif(-1,1),\\
&A_1(0) \,\big|\, L(0) \sim Bern(1/2),\\
&A_2(0) \,\big|\, A_1(0), L(0) \sim Bern(1),\\
&U_1, U_2 \,\big|\, A(0), L(0) \overset{iid}{\sim} Unif(-1,1),\\
&L_1(1) \,\big|\, A(0), L(0), U_1, U_2 \sim U_1(1.25A_1(0) + 0.25),\\
&L_1(1) \,\big|\, A(0), L(0), L_1(1), U_1, U_2 \sim U_2(1.25A_1(0) + 0.25),\\
&A_1(1) \,\big|\, A(0), \bar{L}(1) \sim Bern(1/2),\\
&A_1(1) \,\big|\, A(0), A_1(1), \bar{L}(1) \sim Bern(1),\\
&Y \,\big|\, \bar{A}(1), \bar{L}(1) \sim Bern(0.4 + 0.069 b(\bar{A}(1), \bar{L}(1))),
\end{aligned}
$$

where

$$b(\bar{A}(1), \bar{L}(1)) \equiv 0.5A_1(0)\left(-0.8 - 3(\text{sgn}(L_1(0)) + L_1(0)) - L_2(0)^2\right) + A_1(1)\left(-0.35 + (L_1(1) - 0.5)^2\right) + 0.08A_1(0)A_1(1).$$

Static treatments yield mean outcomes $E_{P_0}Y_{(0,1),(0,1)} = 0.400$, $E_{P_0}Y_{(0,1),(1,1)} \approx 0.395$, $E_{P_0}Y_{(1,1),(0,1)} \approx 0.361$, and $E_{P_0}Y_{(1,1),(1,1)} \approx 0.411$. The true optimal treatment has mean outcome $E_{P_0}Y_{d_0} \approx 0.485$ when $V(0) = L(0)$ and $V(1) = (A(0), \bar{L}(1))$.

7.2.2 Estimation methods—As in the single time point case, we treat the treatment/censoring mechanism as known. Rather than estimate $E_{P_0}[Y \mid \bar{A}(1), \bar{L}(1)]$ when estimating $d_{0, A(1)}$, we consider two extreme cases, namely plugging in either the truth or the constant function 1/2 for the desired expectation. Once the rule $d_{n, A(1)}$ at the second time point has been estimated, we estimate $E_{P_0}[Y_{d_{n,A(1)}} \mid A(0), L(0)]$ by either plugging in the truth, which can be computed analytically using the G-computation formula, or the constant function 1/2. In our simulations we only consider the cases where either both or neither of the sequential regressions is estimated correctly. All simulations use the IPCW mapping to relate the full data loss function to the observed data distribution (see Appendix A).

We use the candidate library in Table 1, with the exception that the Bayes GLM algorithm was excluded from these runs due to an occasional error from the software and the multivariate adaptive regression spline model was also excluded. The convex combinations for the sequential super-learners are selected using the cross-validated empirical risk resulting from the surrogate log loss function and the CV-TMLE estimate of the negative mean outcome risk. The weights $1/g_{0, A(0)}(O)$ and $I(A(1) = d_{n, A(1)}(O))/g_{0, A(1)}(O)$ were incorporated into the procedures for estimating $d_{0, A(1)}$ and $d_{0, A(0)}$ by weighting the candidate algorithms and the empirical risk optimization problem. The fitted rule $d_{n, A(1)}$ used to weight the losses for estimating $d_{0, A(0)}$ was not fitted on the training samples as we recommended in Section 3.2 due to computational constraints.

8 Simulation results

8.1 Single time point

Figures 1(a) and 1(b) respectively give performance results of the super-learner based methods when $V(1) = W_3$ and $V(1) = W_1, \ldots, W_4$. In this simulation, combining both the weighted classification and the regression libraries performs well in both cases. The regression methods with the MSE risk criterion also performs well for all settings of our simulation. The CV-TMLE and CV-DR-IPCW are outperformed by all other methods for selecting α regardless of the specification of $E_{P_0}[Y \mid \bar{A}(1), W]$ in the single time point simulation, but as we note below they perform well compared to many of the individual

algorithms. Correctly specifying the estimate of $E_{P_0}[Y \mid \bar{A}(1), W]$ improves performance for all candidate libraries and choices of the convex combination vector a. Comparing the weighted classification and blip function approaches is difficult given the different candidate library sizes, but both perform well overall.

Multivariate adaptive regression splines appear do the best of all algorithms in the super-learner library when $V(1) = W_1, \ldots, W_4$, though only slightly better than the super-learner fits which do not require *a priori* specification of a single algorithm. The super-learner which used both blip and weighted classification based candidates outperformed all other algorithms in the candidate library. This super-learner performs similarly to the neural network algorithm when $V(1) = W_3$ and outperforms all other algorithms in the candidate library.

All generalized linear model (GLM) methods performed poorly for all settings. For example, when a stepwise regression which includes interaction was used to estimate the blip function and $E_{P_0}[Y \mid \bar{A}(1), W]$ was correctly specified, the mean performance was respectively 0.465 and 0.483 when $V = W_3$ and $V = W_1, \ldots, W_4$. Thus here we see a setting where using data adaptive methods is important for good estimation of the optimal rule.

8.2 Two time points

Figure 2 shows that the the performance of several estimation methods in the two time point case. It appears that the optimal rule for our simulation can be well described by a generalized linear model. In particular, we see a stepwise regression with only main terms outperform all other methods under consideration, including our super-learners. Though the weighted classification based stepwise regression was not included in our model, we ran this algorithm alone to compare to the blip function based stepwise regression. The results were similar, with mean performance of approximately 0.470 for both settings considered.

Although the stepwise regression algorithm performed better for the given data generating distribution at this sample size, the super-learners which aim to maximize an estimate of the mean performance perform well overall. Note that some of the data adaptive methods, such as blip function based neural networks and classification based recursive partitioning perform poorly compared to the other methods. On average across the thousand runs, the super-learner which seeks to maximize the CV-TMLE estimate of the mean outcome and has the conditional mean correctly specified gave the most weight at the first time point to the following algorithms: blip stepwise regression, 0.12; blip stepwise regression with interactions, 0.10; and blip forward stepwise regression, 0.09. Thus our super-learner appears to have learned to select a linear as opposed to a more data adaptive estimator for the latent function.

The mean outcome based super-learners slightly outperformed the weighted log based super-learners in terms of mean performance for both settings.

9 Discussion

This article investigated nonparametric estimation of a V-optimal dynamic treatment. We proposed sequential loss-based super-learning to construct such a nonparametric estimator of the V-optimal rule. When applied in sequentially randomized controlled trials, this method is guaranteed to asymptotically outperform any competitor (with respect to loss-based dissimilarity) at each stage by simply including it in the library of candidate estimators. Some of the proposed sequential super-learners aim to minimize risks associated with learning some latent function which gives the optimal rule. One of these super-learners aims to optimize the performance of the fitted rule itself by maximizing the mean outcome. This seems to be more targeted towards our goal, but our theoretical claim suggests that stronger conditions are needed for the oracle inequality for this selector to hold.

Our simulation results support our theoretical findings. The super-learners always performed comparably to the best candidate in the library, and our theoretical results suggest that increasing sample size will improve their relative performance further.

In the current article we defined the treatment as binary at each time point. Consider now a treatment that has k possible values. We can then define a vector of binary indicators that identify the treatment. We can now apply the results for the multiple time-point treatment case in the appendix of van der Laan and Luedtke [33], since this represents a special case in which at some time-point there are no intermediate covariates between binary treatments. Because the rate of convergence at each time point is upper bounded by the convergence rates at previously fitted time points, there may be better approaches when $\log_2 k \gg 1$.

The sophistication of estimation and inference strategies for optimal treatment regimes has progressed dramatically in recent years thanks to the innovative work of many researchers. We look forward to continued statistical and computational advancements in this field, and to the eventual implementation of these treatment strategies on a large scale.

Acknowledgments

This research was supported by an NIH grant R01 AI074345-06. AL was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. The authors would like to thank Erin LeDell and Sam Lendle for their insights into developing a computationally efficient super-learner using stochastic gradient descent. The authors would also like to thank the anonymous reviewers and Erica Moodie for their invaluable comments and suggestions to improve the quality of the paper.

Appendix

A Double robust loss functions

Below Q represents a parameter value, where the parameter maps from a distribution P to a collection of conditional distributions. Alternatively, we can set these estimates equal to 0 for IPCW-like risk estimates. We use Q_0 to denote the parameter mapping applied to P_0, i.e. the collection of conditional distributions under the observed data distribution P_0. All of the mappings used in this section only require expectations under the conditional distributions in Q. Thus in practice standard regression algorithms can be used to estimate the needed portions of Q_0. When we write conditional expectations under Q as E_Q, it will always be

clear from context what parameter mapping (conditional distribution) of P_0 the appropriate part of Q is supposed to estimate.

Estimates for the optimal rule can be obtained using any regression or classification software, including data adaptive techniques. Because products of differences of Q and Q_0 and g and g_0 will serve as remainder terms for the final risk estimates, it is important to consistently estimate as many of these quantities of interest as possible, ideally at a reasonable rate. Note that the desire for consistent estimates of Q_0 likely precludes the use of parametric regressions for fitting Q, though parametric regressions can be taken as candidates in a cross-validation based algorithm such as SuperLearner. If known, any knowledge of Q_0 or g_0 may be incorporated into the estimates.

Throughout this section we introduce double robust versions of functions defined in the main text. Rather than introduce new notation to account for this, we simply add a Q next to the g in the notation, e.g. $D_2(g)$ becomes $D_2(Q, g)$ and $L_{2,g}$ becomes $L_{2,Q,g}$.

A.1 Blip functions

Define

$$D_2(Q,g)(O) = A_2(1)\frac{2A_1(1)-1}{g_{0,A(1)}(O)}(Y - E_Q[Y \mid \bar{L}(1), \bar{A}(1))]$$
$$+ E_Q[Y \mid \bar{L}(1), A(0), A(1) = (1,1)] - E_Q[Y \mid \bar{L}(1), A(0), A(1) = (0,1)],$$

Let $L^F_{2,D_2(g)}(\bar{Q}_2)(O)$ denote a valid loss function for estimating $E_{P_{0,a(0)}}[D_2(Q,g) \mid V_{a(0)}(1) = v_{a(0)}(1))]$, in the sense that

$$(a(0), v(1)) \mapsto E_{P_{0,a(0)}}[D_2(Q,g)(O_{a(0)}) \mid V_{a(0)}(1) = v(1)]$$

minimizes

$$\sum_{\tilde{a}(0) \in \{0,1\} \times \{1\}} E_{P_{0,\tilde{a}(0)}}\left[L^F_{2,D_2(g)}(\bar{Q}_2)(O_{\tilde{a}(0)})\right]$$

over all measurable functions $\bar{Q}_2$ of $a(0)$ and $v(1)$. Applying the DR-IPCW mapping (vanderLaan Dudoit, 2003) gives:

$$L_{2,Q,g}(\bar{Q}_2)(O) = \frac{A_2(0)}{g_{A(0)}(O)}\left(L^F_{2,D_2(Q,g)}(\bar{Q}_2) - E_Q\left[L^F_{2,D_2(Q,g)}(\bar{Q}_2)\middle|A(0),L(0)\right]\right) \quad (16)$$
$$+ \sum_{a_1(0)=0}^{1} E_Q\left[L^F_{2,D_2(Q,g)}(\bar{Q}_2)\middle|A(0)=(a_1(0),1),L(0)\right],$$

We will use the sign of the $\bar{Q}_2$ which minimizes $L_{2,Q,g}$ to estimate $d_{0,A(1)}$. For a given $d_{A(1)}$, define

$$D_1(d_{A(1)},Q,g)(O) = A_2(0)\frac{2A_1(0)-1}{g_{A(0)}(O)}\left(Y - E_Q\left[Y_{d_{A(1)}}\middle|L(0),A(0)\right]\right)$$
$$+ E_Q\left[Y_{d_{A(1)}}\middle|L(0),A(0)=(1,1)\right] - E_Q\left[Y_{d_{A(1)}}\middle|L(0),A(0)=(0,1)\right].$$

Let $L^F_{1,D_1(d_{A(1)},Q,g)}$, be some loss that satisfies:

$$E_{P_0,d_{A(1)}}\left[D_1(d_{A(1)},Q,g)\middle|V(0)=\cdot\right] = \arg\min_{Q_1} P_{0,d_{A(1)}}L^F_{1,D_1(d_{A(1)},Q,g)}(\bar{Q}_1),$$

Our proposed loss function is obtained by applying the DR-IPCW mapping to the above loss function:

$$L_{1,d_{A(1)},Q,g}(\bar{Q}_1)(O) = \frac{I(A(1)=d_{A(1)}(A(0),V(1)))}{g_{A(1)}(O)}L^F_{1,D_1(d_{A(1)},Q,g)}(\bar{Q}_1) \quad (17)$$
$$- \frac{I(A(1)=d_{A(1)}(A(0),V(1)))}{g_{A(1)}(O)}E_Q\left(L^F_{1,D_1(d_{A(1)},Q,g)}(\bar{Q}_1)\middle|\bar{A}(1),\bar{L}(1)\right)$$
$$+ E_Q\left(L^F_{1,D_1(d_{A(1)},Q,g)}(\bar{Q}_1)\middle|A(0),A(1)=d_{A(1)}(A(0),V(1),\bar{L}(1)\right),$$

We now state a theorem that gives conditions under which the above loss functions allow us to learn the optimal rule d_0.

Theorem 3

(DR Version). Suppose the positivity assumption holds at g and g_0 and either $Q = Q_0$ *or* $g = g_0$. Then:

$$P_0\left\{L_{2,Q,g}(\bar{Q}_2) - L_{2,Q,g}(\bar{Q}_{20})\right\} = \sum_{a(0)} P_{0,a(0)}\left(L^F_{2,D_2(Q,g)}(\bar{Q}_2) - L^F_{2,D_2(Q,g)}(\bar{Q}_{20})\right),$$
$$P_0\left\{L_{1,d_{0,A(1)},Q,g}(\bar{Q}_1) - L_{1,d_{0,A(1)},Q,g}(\bar{Q}_{10})\right\}$$
$$= P_{0,d_{0,A(0)}}\left(L^F_{1,D_1(d_{0,A(1)},Q,g)}(\bar{Q}_2) - L^F_{1,D_1(d_{0,A(1)},Q,g)}(\bar{Q}_{10})\right),$$

where $a(0) \in \{0, 1\} \times \{1\}$. As a consequence:

$$\bar{Q}_{20} = \arg\min_{\bar{Q}_2} P_0 L_{2,Q,g}(\bar{Q}_2),$$
$$\bar{Q}_{10} = \arg\min_{\bar{Q}_1} P_0 L_{1,d_{0,A(1)},Q,g}(\bar{Q}_1).$$

The condition that $Q = Q_0$ can be weakened so that only the needed conditional expectations Q are equal to the analogous expectations under Q_0. We state a slightly stronger form of double robustness than stated in the above theorem in Section 9.1 of the accompanying technical report. The stronger form shows that we have double robustness separately at each time point, so we need only have the portion of g_0 or that of Q_0 corresponding to each time point correctly specified. For example, we may have the treatment/censoring mechanism correctly specified at the first but not the second time point, but $L_{2,Q,g}$ is still a valid loss as long as the portion of Q corresponding to the second time point is correctly specified (even if Q is misspecified at the first time point!).

Proof of Theorem *(DR Version)*

Suppose $Q = Q_0$ or $g = g_0$. By the double robustness of DR-IPCW mapping:

$$E_{P_0} L_{2,Q,g}(\bar{Q}_2)(O) = \sum_{a(0)} E_{P_{0,a(0)}} L^F_{2,D_2(Q,g)}(\bar{Q}_2),$$
$$E_{P_0} L_{1,d_{0,A(1)},Q,g}(\bar{Q}_1) = E_{P_{0,d_{0,A(1)}}}\left[L^F_{1,D_1(d_{0,A(1)},Q,g)}(\bar{Q}_1)\right].$$

All claims again follow immediately by the choice of $L^F_{2,D_2(Q,g)}$, and $L^F_{1,D_1(d_{0,A(1)},Q,g)}$. □

Optimizing the double robust blip loss functions is not straightforward because of the final two terms in expressions in (16) and (17). Taking these terms to be 0, which is equivalent to misspecifying these needed conditional expectations under Q_0, allows for the use of weighted regression methods. We show in Section A.3 that optimizing the weighted classification losses does not encounter this difficulty.

A.2 Performance of rule

Define:

$$-\widetilde{L}^F_{2,Q,g}\left(d_{A(1)}\right)(O) = \frac{I(A(1) = d_{A(1)}(a(0), V(1)))}{g_{A(1)}(O)}(Y - E_Q[Y \mid \bar{L}(1), \bar{A}(1)]) + E_Q[Y \mid \bar{L}(1), A(0), A(1) = d_{A(1)}(a(0), V(1))].$$

Applying the DR-IPCW mapping [8] gives:

$$\widetilde{L}_{2,Q,g}\left(d_{A(1)}\right)(O) = \frac{A_2(0)}{g_{A(0)}(O)}\left(\widetilde{L}^F_{2,Q,g}\left(d_{A(1)}\right)(O) - E_Q\left[\widetilde{L}^F_{2,Q,g}\left(d_{A(1)}\right) \mid A(0), L(0)\right]\right) + \sum_{a_1(0)=0}^{1} E_Q\left[\widetilde{L}^F_{2,Q,g}\left(d_{A(1)}\right) \mid A(0) = (a_1(0), 1), L(0)\right].$$

Let $d_{A(1)}$ be a treatment rule for the second time point. Define:

$$-\widetilde{L}^F_{1,d_{A(1)},Q,g}\left(d_{A(0)}\right)(O) = \frac{I(A(0) = d_{A(0)}(V(0)))}{g_{A(0)}(O)}(Y - E_Q[Y_{d_{A(1)}} \mid L(0), A(0)]) + E_Q[Y_{d_{A(1)}} \mid L(0), A(0) = d_{A(0)}(V(0))].$$

Applying the DR-IPCW mapping gives:

$$\widetilde{L}_{1,d_{A(1)},Q,g}\left(d_{A(0)}\right)(O) = \frac{I(A(1) = d_{A(1)}(A(0), V(1)))}{g_{A(1)}(O)}\widetilde{L}^F_{1,d_{A(1)},Q,g}\left(d_{A(0)}\right) - \frac{I(A(1) = d_{A(1)}(A(0), V(1))}{g_{A(1)}(O)}E_Q\left(\widetilde{L}^F_{1,d_{A(1)},Q,g}\left(d_{A(0)}\right) \mid \bar{A}(1), \bar{L}(1)\right) + E_Q\left(\widetilde{L}^F_{1,d_{A(1)},Q,g}\left(d_{A(0)}\right) \mid A(0), A(1) = d_{A(1)}(A(0), V(1)), \bar{L}(1)\right)$$

Theorem 4 (DR Version)

Suppose the positivity assumption holds at g and g_0 and either $Q = Q_0$ or $g = g_0$. Then:

$$P_0\left\{\widetilde{L}_{2,Q,g}(d_{A(1)}) - \widetilde{L}_{2,Q,g}(d_{0,A(1)})\right\} = \sum_{a(0)} P_0 I(d_{A(1)} \neq d_{0,A(1)}) \left|\bar{Q}_{20}\right|(a(0), V_{a(0)}(1)),$$

$$P_0\left\{\widetilde{L}_{1,d_{0,A(1)},Q,g}(d_{A(0)}) - \widetilde{L}_{1,d_{0,A(1)},Q,g}(d_{0,A(0)})\right\} = P_0 I(d_{A(0)} \neq d_{0,A(0)}) \left|\bar{Q}_{20}\right|(V(0)),$$

where the sum is over $a(0) \in \{0, 1\} \times \{1\}$. It follows that the minimizer of $P_0\widetilde{L}_{2,Q,g}(d_{A(1)})$ over rules $d_{A(1)}$ is an optimal second time point rule, and the minimizer of $P_0\widetilde{L}_{1,d_{0,A(1)},Q,g}(d_{A(0)})$ over rules $d_{A(0)}$ is an optimal first time point rule.

Proof of Theorem 4 *(DR Version)*

For all $d_{A(1)}$:

$$\begin{aligned}
&P_0\left(\widetilde{L}_{2,Q,g^{(d_{A(1)})}} - \widetilde{L}_{2,Q,g^{(d_{0,A(1)})}}\right)\\
&= \sum_{a(0)} P_{0,a(0)}\left(\widetilde{L}^F_{2,Q,g^{(d_{A(1)})}} - \widetilde{L}^F_{2,Q,g^{(d_{0,A(1)})}}\right)\\
&= \sum_{a(0)} P_{0,a(0)}\left(E_{P_0,a(0)}\left[\widetilde{L}^F_{2,Q,g^{(d_{A(1)})}} - \widetilde{L}^F_{2,Q,g^{(d_{0,A(1)})}} \middle| V_{a(0)}\right]\right)\\
&= \sum_{a(0)} P_{0,a(0)} I(d_{A(1)} \neq d_{0,A(1)})(a(0), V_{a(0)}) \left|\bar{Q}_{20}(a(0), V_{a(0)})\right|.
\end{aligned}$$

where the sums are over $a(0) \in \{0, 1\} \times \{1\}$. Because $\left|\bar{Q}_{20}\right| \geq 0$, the above is minimized at $d_{A(1)} = d_{0,A(1)}$. For any first time point treatment rule $d_{A(0)}$:

$$\begin{aligned}
&P_0 r\left\{\widetilde{L}_{1,d_{0,A(1)},Q,g^{(d_{A(0)})}} - \widetilde{L}_{1,d_{0,A(1)},Q,g^{(d_{0,A(0)})}}\right\}\\
&= P_{0,d_{0,A(1)}}\left\{\widetilde{L}^F_{1,d_{0,A(1)},Q,g^{(d_{A(0)})}} - \widetilde{L}^F_{1,d_{0,A(1)},Q,g^{(d_{0,A(0)})}}\right\}\\
&= P_{0,d_{0,A(1)}}\left\{E_{P_0,d_{0,A(1)}}\left[\widetilde{L}^F_{1,d_{0,A(1)},Q,g^{(d_{A(0)})}} - \widetilde{L}^F_{1,d_{0,A(1)},Q,g^{(d_{0,A(0)})}} \middle| V(0)\right]\right\}\\
&= P_0 I\left(d_{A(0)} \neq d_{0,A(0)}\right)(V(0)) \left|\bar{Q}_{10}(V(0))\right|.
\end{aligned}$$

The above expression is minimized at $d_{A(0)} = d_{0,A(0)}$. □

A.3 Weighted classification

We will use the definitions of Q, D_1, and D_2 from the Section A.1.

Define:

$$\begin{aligned}
K_{2,Q,g}(O) = &\frac{A_2(0)}{g_{A(0)}(O)}(D_2(Q,g) - E_Q[D_2(Q,g) \mid A(0), L(0)])\\
&+ \sum_{a_1(0)=0}^{1} E_Q[D_2(Q,g) \mid A(0) = (a_1(0), 1), L(0)].
\end{aligned}$$

Also define:

$$\hat{L}_{2,Q,g^{(d_{A(1)})}}(O) = \left|K_{2,Q,g}(O)\right| I(d_{A(1)}(A(0), V(0)) \neq (Z \circ K_{2,Q,g}(O), 1)).$$

Similarly, let:

$$K_{1,d_{A(1)},Q,g}(O) = \frac{I(A(1) = d_{A(1)}(A(0), V(1)))}{g_{A(1)}(O)} D_1(d_{A(1)}, Q, g))$$
$$- \frac{I(A(1) = d_{A(1)}(A(0), V(1))}{g_{A(1)}(O)} E_Q\left(D_1(d_{A(1)}, Q, g) \middle| \bar{A}(1), \bar{L}(1)\right)$$
$$+ E_Q\left(D_1(d_{A(1)}, Q, g) \middle| A(0), A(1) = d_{A(1)}(A(0), V(1)), \bar{L}(1)\right),$$

and:

$$\hat{L}_{1,d_{A(1)},Q,g}(d_{A(0)})(O) = \left|K_{1,d_{A(1)},Q,g}(O)\right| I(d_{A(0)}(A(0), V(0)) \neq (Z \circ K_{1,d_{A(1)},Q,g}(O), 1)).$$

We have the following theorem:

Theorem 5 (DR Version)

Suppose the positivity assumption holds at g and g_0. Then for any $(d_{A(0)}, d_{A(1)}) \in \mathcal{D}$:

$$\hat{L}_{2,Q,g}(d_{A(1)}) = \tilde{L}_{2,Q,g}(d_{A(1)}) + C_{2,Q,g},$$
$$\hat{L}_{1,d_{A(1)},Q,g}(d_{A(0)}) = \tilde{L}_{1,d_{A(1)},Q,g}(d_{A(0)}) + C_{1,d_{A(1)},Q,g},$$

where $C_{2,Q,g}(O)$ and $C_{1,d_{A(1)},Q,g}(O)$do not rely on $d_{A(1)}$ *or* $d_{A(0)}$, respectively. It follows that $\hat{L}_{2,Q,g}$ and $\hat{L}_{1,d_{A(1)},Q,g}$, are valid loss functions for sequentially estimating $d_{0,A(1)}$ and $d_{0,A(0)}$ if either either $Q = Q_0$ or $g = g_0$.

Proof of Theorem 5 *(DR Version)*

For all realizations $o \in \mathcal{O}$, define:

$$C_{2,Q,g}(o) = -\tilde{L}_{2,Q,g}((Z \circ K_{2,Q,g}(o), 1))(o),$$

where $\tilde{L}_{2,Q,g}((Z \circ K_{2,Q,g}(o), 1))$ represents $\tilde{L}_{2,Q,g}$ evaluated at the static decision rule where everyone is given the treatment $Z \circ K_{2,Q,g}(o) \in \{0, 1\}$ without censoring.

Checking all values of $d_{A(1)} \in \{0, 1\} \times \{1\}$, $Z \circ K_{2,Q,g} \in \{0, 1\}$, $a(0), a(1) \in \{0, 1\}^2$ shows that:

$$\left|K_{2,Q,g}\right| I(d_{A(1)} \neq (Z \circ K_{2,Q,g}, 1)) - \tilde{L}_{2,Q,g}(d_{A(1)}) = C_{2,Q,g}.$$

For the first time point, we define:

$$C_{1,d_{A(1)},Q,g}(o) = -\tilde{L}^{F}_{1,d_{A(1)},Q,g}\left((Z \circ K_{1,d_{A(1)},Q,g}(o),1)\right)(o).$$

Checking all values of $d_{A(0)} \in \{0,1\} \times \{1\}$, $Z \circ K_{1,d_{A(1)},Q,g} \in \{0,1\}$, and $a(0), a(1) \in \{0,1\}^2$ shows that:

$$\left|K_{1,d_{A(1)},Q,g}\right| I\left(d_{A(0)} \neq (Z \circ K_{1,d_{A(1)},Q,g},1)\right) - \tilde{L}_{1,d_{A(1)},Q,g}(d_{A(0)}) = C_{1,d_{A(1)},Q,g}.$$

The claim that $\hat{L}_{2,Q,g}$ and $\hat{L}_{1,d_{A(1)},Q,g}$ are valid loss functions for the sequential estimation of d_0 follows by the double robust version of Theorem 4. □

B Convex surrogate for the weighted 0–1 loss

We now present a simple result which motivates future work to apply general results on surrogate loss functions like those in Bartlett, Jordan, and McAuliffe [35] to the above weighted classification problem. Zhao et al. [29] present a specific result with a weighted hinge loss function in the single time point case. The result below can be extended naturally using the methods in Bartlett et al. 06 but already covers many interesting cases. The theorem uses the double robust versions of $K_{2,g}$ and $\hat{L}_{2,g}$ as presented in Appendix A.3.

Theorem 6

Suppose the positivity assumption holds at g and $0 < E|K_{2,Q,g(O)}| < \infty$. Let $\phi: \mathbb{R} \to [0,\infty)$ be some convex function that is differentiable at 0 with $\phi'(0) < 0$. Define

$$L_{2,\phi,Q,g}(f_2)(O) = \left|K_{2,Q,g}(O)\right| \phi\left(f_2(A(0),V(1))(2Z \circ K_{2,d_{A(1)},Q,g}(O) - 1)\right)$$

for some latent function f with range $\mathbb{R}$. Let $f_{2,i}$ be some sequence of functions and $d_{A(1),i}$ be a sequence of functions such that $d^{(i)}_{A(1)}(A(0),V(1))$ gives treatment $I(f_{2,i}(A(0),V(1)) \geq 0)$ without censoring. Then:

$$P_0 L_{2,\phi,Q,g}(f_{2,i}) \xrightarrow{i \to \infty} \inf_{\tilde{f}_2} P_0 L_{2,\phi,Q,g}(\tilde{f}_2) \Rightarrow P_0 \hat{L}_{2,Q,g}(d^{(i)}_{A(i)}) \xrightarrow{i \to \infty} P_0 \hat{L}_{2,Q,g}(d_{0,A(1)}),$$

where the infimum is over all measurable functions $\tilde{f}_2$ that take $A(0), V(1)$ as input.

Proof of Theorem 6

By the law of total expectation, for all $d_{A(1)}$ that set observations to uncensored:

$$E_{P_0}\hat{L}_{2,Q,g}\left(d_{A(1)}\right) = E_{P_0}\left[E_{P_0}\left[\left|K_{2,Q,g}\right| A(0), V(1), Z \circ K_{2,Q,g}(O)\right] I\left(d_{A(1),1}(A(0), V(1)) \neq Z \circ K_{2,Q,g}\right)\right],$$

where $d_{A(1),1}$ is the treatment index of the optimal rule. Let $\widetilde{P}_0$ be the probability measure with:

$$Pr_{\widetilde{P}_0}\left((A(0), V(1), Z \circ K_{2,Q,g}) \in B\right)$$
$$= \frac{1}{E_{P_0}\left|K_{2,Q,g}\right|}\int_B E_{P_0}\left[\left|K_{2,Q,g}\right| A(0), V(1), Z \circ K_{2,Q,g}(O)\right] dP_0$$

for all measurable sets B. Note that $\widetilde{P}_0$ is a probability distribution over values of $A(0)$, $V(1)$, $Z \circ K_{2,Q,g}$ and that $\widetilde{P}_0$ is absolutely continuous with respect to P_0. Also note that

$$E_{P_0}\hat{L}_{2,Q,g}\left(d_{A(1)}\right) = E_{\widetilde{P}_0} I\left(d_{A(1),1}(A(0), V(1)) \neq Z \circ K_{2,Q,g}\right),$$

so we can now consider a simple 0–1 loss under the distribution $\widetilde{P}_0$.

By Theorem 4 in Bartlett et al., ϕ is classification-calibrated according to the definition in the paper. By part (c) of Theorem 3 in the same paper, it follows that:

$$\lim_{i \to \infty} \widetilde{P}_0 \phi\left(f_{2,i}(A(0), V(1))(2Z \circ K_{1,d_{A(1)},Q,g} - 1)\right) = \inf_{\widetilde{f}_2} \widetilde{P}_0 \phi\left(\widetilde{f}_2(A(0), V(1))(2Z \circ K_{1,d_{A(1)},Q,g} - 1)\right)$$
$$\lim_{i \to \infty} \widetilde{P}_0 I\left(I(f_{2,i}(A(0), V(1)) > 0) \neq Z \circ K_{2,Q,g}\right) = \inf_{\widetilde{f}_2} \widetilde{P}_0 I\left(I(\widetilde{f}_2(A(0), V(1)) > 0) \neq Z \circ K_{2,Q,g}\right),$$

Writing the above expectations under $\widetilde{P}_0$ as expectations under P_0 weighted by $d\widetilde{P}_0/dP_0$ and multiplying by the constant $E_{P_0}\left|K_{2,Q,g}\right|$ gives the desired result. □

Examining the above proof shows that the conditions on ϕ can be weakened to the condition that ϕ is classification-calibrated according to the definition in Bartlett, Jordan, and McAuliffe [35].

If Q or g is correctly specified and the infimum of $P_0 L_{2,\phi,Q,g}(\cdot)$ is achievable at some f_2^* then it follows immediately that $d_{0,A(1)}$ has the same performance as the rule $(A(0), V(1)) \to I(f_2^*(A(0), V(1)) > 0)$ under $P_{0,d(0)}$. This shows that weighted surrogate loss functions are valid for $d_{0,A(1)}$.

An analogous result holds for the first time point using the loss

$$L_{1,\phi,d_{A(1)},Q,g}(f_1) = \left| K_{1,\phi,d_{A(1)},Q,g}(O) \right| \phi\left(f_1(V(0))(2Z \circ K_{1,d_{A(1)},Q,g}(O) - 1) \right).$$

C Example 4 proof

Proof that (5) holds in Example 4. Note that

$$\begin{aligned} &Var_{P_0}\left(L_{2,g_0}(f_2) - L_{2,g_0}(f_{20}) \right) \leq E_{P_0}\left(L_{2,g_0}(f_2) - L_{2,g_0}(f_{20}) \right)^2 \\ &= E_{P_0}\left[I(I(f_2 \geq 0) \neq I(f_{20} \geq 0))(A(0), V(1)) \frac{A_2(0)}{g_{A(0)}(O)^2} \frac{A_2(1)}{g_{A(1)}(O)^2} Y^2 \right] \\ &\leq \delta^{-2} \sum_{a(0)} \sum_{a(1)} E_{P_0}\left[I(I(f_2 \geq 0) \neq I(f_{20} \geq 0))(a(0), V_{a(0)}(1)) Y^2_{a(0),a(1)} \right], \end{aligned}$$

where the sums are over $\{0, 1\} \times \{1\}$. For all $k_\delta > 0$, Theorem 4 shows that:

$$\begin{aligned} &k_\delta Var_{P_0}\left(L_{2,g_0}(f_2) - L_{2,g_0}(f_{20}) \right) - E_{P_0}\left[L_{2,g_0}(f_2) - L_{2,g_0}(f_{20}) \right] \\ &\leq \max_{a(1)} \sum_{a(0)} E_{P_0}\left[I(I(f_2 \geq 0) \neq I(f_{20} \geq 0)) \left(2k_\delta Y^2_{a(0),a(1)} - \left| \bar{Q}_{20} \right| \right)(a(0), V_{a(0)}) \right], \end{aligned}$$

where the maximum is over $a(1) \in \{0, 1\} \times \{1\}$. By 4, we can choose $k_\delta > 0$ small enough so that $2K_\delta E_{P_0}\left[Y^2_{a(0),a(1)} \middle| V_{a(0)} \right] - \left| \bar{Q}_{20}(a(0), V_{a(0)}(1)) \right| \leq 0$ almost surely for all $a(0)$, $a(1) \in \{0, 1\} \times \{1\}$. The law of total expectation applied to the above then shows that, for $k_\delta > 0$ sufficiently small:

$$k_\delta Var_{P_0}\left(L_{2,g_0}(f_2) - L_{2,g_0}(f_{20}) \right) - E_{P_0}\left[L_{2,g_0}(f_2) - L_{2,g_0}(f_{20}) \right] \leq 0.$$

Condition (5) follows immediately, thus completing the proof. □

D Non-sequential super-learner targeted directly at mean outcome

We now sketch a non-sequential super-learner which seeks to maximize the mean outcome under the entire estimated rule $d_n = (d_{n,A(0)}, d_{n,A(1)})$. This estimator is a direct application of the CV-TMLE presented in Section 7.1 and Appendix B.2 of van der Laan and Luedtke [33]. Suppose we have libraries of sequential candidate latent function estimators $(P \mapsto \hat{f}_{1,j}(P) : j = 1, \ldots, J_1)$ and $(P \mapsto \hat{f}_{2,j}(P) : j = 1, \ldots, J_2)$ for the first and second time points. The latent function estimators for the first time point rely on nuisance function fits for the second time point rule, but there is no requirement that this nuisance function be the same as the final output rule $d_{n,A(1)}$ at the second time point. For each fold u we can compute a

sequential super-learner for the second time point on training set u, which yields an estimate $d^{nuis}_{n,u,A(1)}$ of $d_{0,A(1)}$. In learning each $d^{nuis}_{n,u,A(1)}$ we have estimated latent functions resulting from estimators $\hat{f}_{2,j_2}, j_2 = 1,\ldots,J_2$, applied to all of the training $(\cdot)$ samples. We can get estimates resulting from applying the sequential estimators $\hat{f}_{1,j_1}, j_1 = 1,\ldots,J_1$, to each training set u, where the treatment at the second time point is set to $d^{nuis}_{n,u,A(1)}$.

We now have estimates resulting from estimators and $\hat{f}_{1,j_1}$ and $\hat{f}_{2,j_2}$ applied to each training sample for all j_1, j_2. We can simultaneously optimize over $a_{A(0)}$ and $a_{A(1)}$ to maximize the CV-TMLE of the mean outcome under the fitted rule. The final estimated latent functions at the first and second time points are given by $\sum_j \alpha_{n,A(0),j}\hat{f}_{1,j}(P_n)$ and $\sum_j \alpha_{n,A(1),j}\hat{f}_{2,j}(P_n)$, respectively. This method seems to be most targeted towards our goal, namely maximizing the mean outcome under the estimated rule. We note that $a_{A(1)}$ need not equal any of the convex combinations $\alpha^{nuis}_{u,A(1)}$ used to obtain each $d^{nuis}_{n,u,A(1)}$, but we can establish oracle inequalities will ensure that that $a_{P,A(1)}$ performs at least as well as each $\alpha^{nuis}_{u,A(1)}$ in terms of mean outcome for the final output optimal rule. We leave deeper consideration of this cross-validation scheme to future work.

References

1Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. Comput Math Appl. 1987; 14:139s–161s.

2Robins JM. Proc Biopharm Sect American Statistical Association; 1993 Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers; 26

3Robins JM. Causal inference from complex longitudinal data. In: Berkane EM, editorLatent variable modeling and applications to causality New York: Springer Verlag; 1997 69117

4Holland PW. Statistics and causal inference. J Am Stat Assoc. 1986; 81:945–60.

5Neyman J. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (In Polish). In: Dabrowska DM, , Speed TP, translatorsStat Sci Vol. 5. 1923 46580(1990)

6Pearl J. Causality: models, reasoning and inference 2nd. New York: Cambridge University Press; 2009

7Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974; 66:688–701.

8van der Laan MJ, , Dudoit S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples Division of Biostatistics, University of California; Berkeley: 2003Technical Report 130

9van der Laan MJ, , Dudoit S, , van der Vaart AW. The cross-validated adaptive epsilon-net estimator, Technical Report 142 3, Division of Biostatistics University of California; Berkeley: 2004

10van der Laan MJ, Polley E, Hubbard A. Super learner. Stat Appl Genet Mol. 2007; 6 Article 25.

11Murphy S. An experimental design for the development of adaptive treatment strategies. Stat Med. 2005; 10:24.

12Cotton C, Heagerty P. A data augmentation method for estimating the causal effect of adherence to treatment regimens targeting control of an intermediate measure. Stat Biosc. 2011; 3:28–44.

13Orellana L, Rotnitzky A, Robins JM. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part I: proofs and additional results. Int J Biostat. 2010; 6 Article 8.

14Robins JM, Orallana L, Rotnitzky A, Orellana L. Estimation and extrapolation of optimal treatment and testing strategies. Stat Med. 2008; 27:4678–721. [PubMed: 18646286]

15Petersen ML, van der Laan MJ, Napravnik S, Eron JJ, Moore RD, Deeks SG. Long-term consequences of the delay between virologic failure of highly active antiretroviral therapy and regimen modification. AIDS. 2008; 22:2097–106. [PubMed: 18832873]

16Petersen ML, Deeks SG, Martin JN, van der Laan MJ. History-adjusted marginal structural models for estimating time-varying effect modification. Am J Epidemiol. 2007; 166:985–93. [PubMed: 17875580]

17Moodie E, Platt R, Kramer M. Estimating response-maximized decision rules with applications to breastfeeding. J Am Stat Assoc. 2009; 104:155–65.

18Murphy SA. Optimal dynamic treatment regimes. J R Stat Soc Ser B. 2003; 65:331–6.

19Robins JM. Discussion of “Optimal dynamic treatment regimes” by Susan A. Murphy. J R Stat Soc Ser B. 2003; 65:355–66.

20Sutton R, , Sung H. Reinforcement learning: an introduction Cambridge, MA: MIT Press; 1998

21Robins JM. Optimal structural nested models for optimal sequential decisions. In: Lin DY, , P Heagerty, editorsProc Second Seattle Symp Biostat Vol. 179. 2004 189326

22Dudoit S, van der Laan MJ. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. Stat Methodol. 2005; 2:131–54.

23Ernst D, Geurts P, Wehenkel L. Tree-based batch mode reinforcement learning. J Mach Learn Res. 2005; 6:503–56.

24Ormoneit D, Sen S. Kernel-based reinforcement learning. Mach Learn. 2002; 49:161–78.

25van der Laan MJ, , Robins JM. Unified methods for censored longitudinal data and causality New York Berlin Heidelberg: Springer; 2003

26Qian M, Murphy S. Performance guarantees for individualized treatment rules. Ann Statist. 2011; 39:1180–210.

27Zhao Y, Zeng D, Rush A, Kosorok M. Estimating individual treatment rules using outcome weighted learning. J Am Stat Assoc. 2012; 107:1106–18. [PubMed: 23630406]

28Rubin DB, van der Laan MJ. Statistical issues and limitations in personalized medicine research with clinical trials. Int J Biostat. 2012; 8 Article 18.

29Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. Statistics. 2012; 68:103–14.

30Zhao Y-Q, Zeng D, Laber EB, Kosorok MR. New statistical learning methods for estimating optimal dynamic treatment regimes. J Am Stat Assoc. 2014; 110:583–98.

31Polley EC, , Rose S, , van der Laan MJ. Super learning. In: van der Laan MJ, , Rose S, editorsTargeted Learning: Causal Inference for Observational and Experimental Data New York, Dordrecht, Heidelberg, London: Springer, New York; 2011 chapter 3

32van der Laan MJ, , Rose S. Targeted learning: causal inference for observational and experimental data New York: Springer, New York; 2011

33van der Laan MJ, , Luedtke AR. Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome Division of Biostatistics, University of California; Berkeley: 2014aTechnical Report 329available at http://www.bepress.com/ucbbiostat/

34van der Laan MJ, Luedtke AR. Targeted learning of the mean outcome under an optimal dynamic treatment rule. J Causal Inference. 2014b; 3:61–95.

35Gill RD, Robins JM. Causal inference in complex longitudinal studies: continuous case. Ann Stat. 2001; 29:1785–811.

36Chakraborty B, , Moodie EE. Statistical Methods for Dynamic Treatment Regimes Berlin Heidelberg New York: Springer; 2013

37Bartlett PL, Jordan MI, McAuliffe JD. Convexity, classification, and risk bounds. J Am Stat Assoc. 2006; 101:138–56.

38Zheng W, , van der Laan MJ. Asymptotic theory for cross-validated targeted maximum likelihood estimation Division of Biostatistics, University of California; Berkeley: 2010Technical Report 273

39Diaz I, , van der Laan MJ, , Daz I. Targeted data adaptive estimation of the causal dose response curve Division of Biostatistics, University of California; Berkeley: 2013Technical Report 306submitted to JCI

40Polley E, van der Laan MJ. Super Learner Prediction. 2012

41Ypma J. The NLopt nonlinear-optimization package. 2014

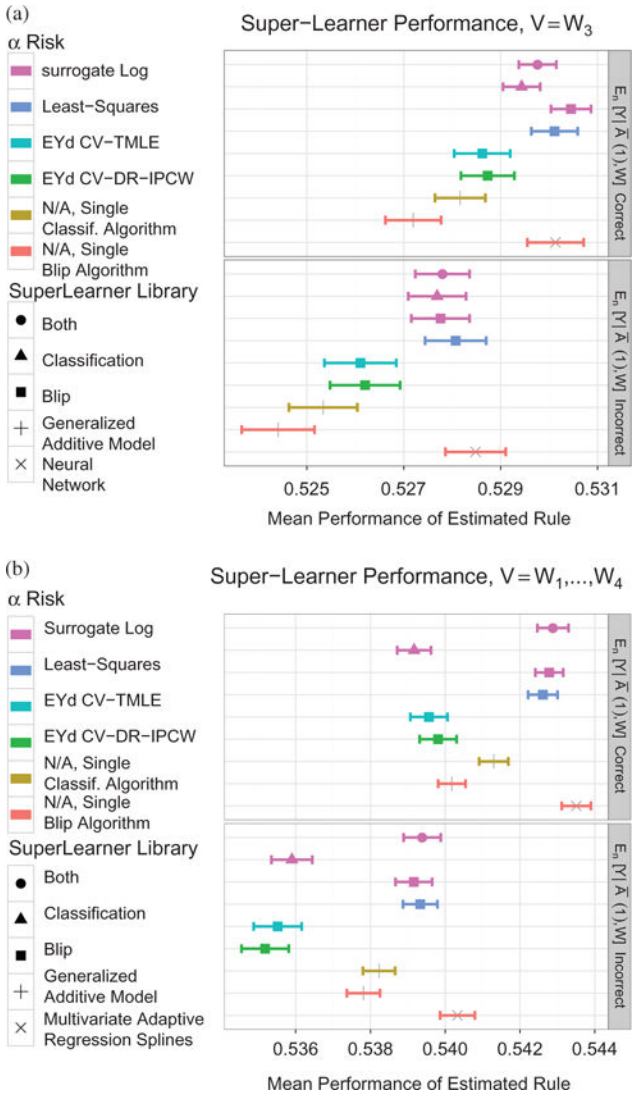

Figure 1.

Mean performance of the estimated rule when the estimate $E_n[Y|\bar{A}(1), W]$ of $E_{P_0}[Y|\bar{A}(1), W]$, is correctly and incorrectly specified. Error bars indicate 95% confidence intervals to account for uncertainty from the finite number of Monte Carlo draws in our simulation. (a) V(1) = W_3, (b) V(1) = W_1, …, W_4.

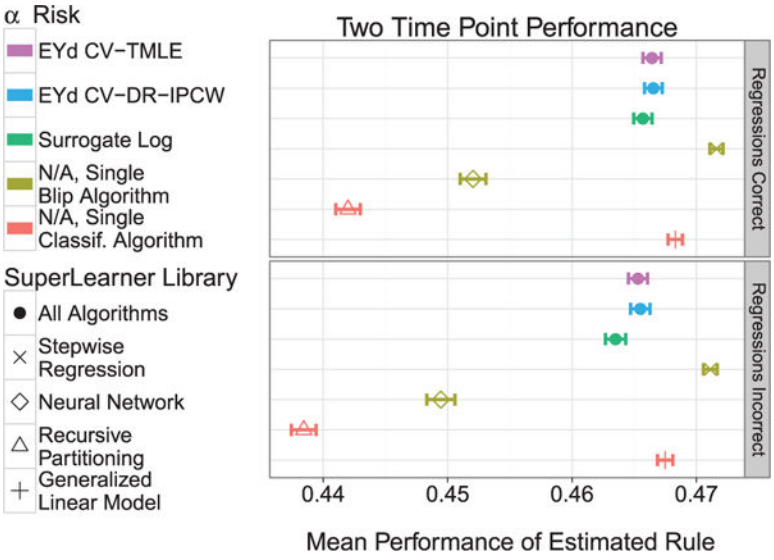

Figure 2.

Mean performance of the estimated rule when $E_{P_0}[Y \mid \bar{A}(1), \bar{L}(1)]$ and $E_{P_0}[Y_d \mid A(0), L(0)]$ are specified correctly and incorrectly. Error bars indicate 95% confidence intervals to account for uncertainty from the finite number of Monte Carlo draws in our simulation.

Table 1

Candidate estimators used to estimate $\bar{Q}_{2,0}$. See the SuperLearner package documentation for more details (Polley vanderLaan12). SL.earth only used for $V = (W_1, \ldots, W_4)$.

Loss function	Method	R function
Squared error	Bayesian GLM	`SL.bayesglm`
	Generalized additive model	`SL.gam`
	Generalized linear model	`SL.glm`
	Generalized linear model, interactions	`SL.glm.interaction`
	Multivariate adaptive regression splines	`SL.earth`
	Sample mean	`SL.mean`
	Neural network	`SL.nnet`
	Stepwise regression	`SL.step`
	Forward stepwise regression	`SL.step.forward`
	Stepwise regression, interactions	`SL.step.interaction`
Weighted log	Generalized additive model	`SL.gam`
	Generalized linear model	`SL.glm`
	Generalized linear model, interactions	`SL.glm.interaction`
	Neural network	`SL.nnet`
	Recursive partitioning	`SL.rpart`