Targeted learning ensembles for optimal individualized treatment rules with time-to-event outcomes

BY I. DÍAZ, O. SAVENKOV AND K. BALLMAN

*Division of Biostatistics, Weill Cornell Medicine, 402 East 67th Street, New York,
New York 10065, U.S.A.*

ild2005@med.cornell.edu ols2010@med.cornell.edu kab2053@med.cornell.edu

SUMMARY

We consider estimation of an optimal individualized treatment rule when a high-dimensional vector of baseline variables is available. Our optimality criterion is with respect to delaying the expected time to occurrence of an event of interest. We use semiparametric efficiency theory to construct estimators with properties such as double robustness. We propose two estimators of the optimal rule, which arise from considering two loss functions aimed at directly estimating the conditional treatment effect and recasting the problem in terms of weighted classification using the 0-1 loss function. Our estimated rules are ensembles that minimize the crossvalidated risk of a linear combination in a user-supplied library of candidate estimators. We prove oracle inequalities bounding the finite-sample excess risk of the estimator. The bounds depend on the excess risk of the oracle selector and a doubly robust term related to estimation of the nuisance parameters. We discuss the convergence rates of our estimator to the oracle selector, and illustrate our methods by analysis of a phase III randomized study testing the efficacy of a new therapy for the treatment of breast cancer.

Some key words: Double robustness; Individualized treatment rule; Oracle inequality; Super-learning.

1. INTRODUCTION

Individualized treatment rules play a fundamental role in precision medicine, whereby decisions are targeted to the individual based on their expected clinical response. A treatment rule is a function that maps an individual's pre-treatment covariates to an optimal treatment choice. In this paper, we discuss learning the optimal rules from data collected as part of an observational or randomized study, where optimality is defined as the maximum delay in the expected time of occurrence of an undesirable event.

Recent advances in biomedical imaging and gene expression technology have led to the generation of large amounts of data that can be used to tailor treatment to very specific patient characteristics. Methods for estimating the optimal rule when it is defined with respect to a single time-point outcome include those of [Qian & Murphy \(2011\)](#), [Rubin & van der Laan \(2012\)](#), [Zhao et al. \(2012\)](#), [McKeague & Qian \(2014\)](#) and [Song et al. \(2015\)](#). Several authors have developed techniques using survival outcomes subject to informative censoring. Early approaches were proposed by [Zhao et al. \(2011\)](#) and [Goldberg & Kosorok \(2012\)](#). The latter methods use Q-learning, relying on sequential support vector regressions, to estimate the optimal sequential treatment rule that optimizes a survival outcome under right-censoring. [Geng et al. \(2015\)](#) also dealt with estimation of the optimal rule in a survival setting using ℓ_1 regularization for the

outcome regression under the assumption that censoring is independent of covariates and the outcome, but their decision functions are linear. [Zhao et al. \(2015\)](#) generalized the weighted classification approach of [Zhao et al. \(2012\)](#) to allow for informative censoring and doubly robust loss functions, but their decision functions are restricted to support vector machines. [Bai et al. \(2016\)](#) presented methods for estimating optimal rules with a survival outcome subject to informative censoring. They considered two strategies, one based on estimating the conditional treatment effect and the other based on a classification perspective. Their methods are restricted to the use of parametric nuisance estimators, as well as decision functions that can be indexed by a Euclidean vector. These assumptions result in limited applicability of the methods to situations with complex data-generating mechanisms that preclude the correct specification of parametric models. All the above methods are potential candidates in the library of estimators that constitute our ensembles.

We propose two methods for constructing an ensemble of decision functions for the optimal rule. Our ensembles are linear combinations of estimators in a user-supplied library, where the coefficients in the linear combination are chosen to minimize the crossvalidated risk. We propose to use a doubly robust loss function inspired by efficient estimation theory for marginal causal effects ([Moore & van der Laan, 2009](#); [Díaz et al., 2018](#)). In our context, double robustness means that the estimated rules will have certain optimality properties under consistent estimation of at least one of two nuisance parameters: the hazard of the outcome at each time-point conditional on covariates and treatment, and the hazard of censoring and the treatment mechanism.

The library of candidate estimators may contain any of the algorithms discussed above. In light of the no-free-lunch theorems of [Wolpert \(2002\)](#) for supervised learning, our ensembles are expected to have better or equal generalization error compared to any of the individual candidates in the library for any given dataset. We prove this claim in the form of an oracle inequality, which bounds the excess risk of our estimator in terms of the excess risk of the oracle estimator, defined as the combination of estimators that would be chosen in a hypothetical world in which an infinite validation sample is available and at least one of the nuisance parameters is known. Our methods are developed under the assumption that censoring is random within strata of treatment and baseline variables. We also assume that treatment is randomized within strata of the covariates, either by nature or by experimentation.

The finite-sample bounds we present are inspired by developments in the targeted learning literature, which establish the optimality of crossvalidation in estimator selection for high-dimensional parameters ([van der Laan et al., 2006](#)). [Luedtke & van der Laan \(2016b\)](#) considered super-learning ensembles for estimating optimal dynamic treatment regimes at two time-points. They derived oracle inequalities for super-learning of the optimal rule using a loss function indexed by the treatment mechanism, which is assumed to be known. We generalize their results by providing oracle inequalities under a doubly robust loss function indexed by two nuisance parameters when neither of the nuisance parameters is known, by showing that the oracle inequalities inherit the double robustness property of the loss function, and by presenting comparable oracle inequalities for the 0-1 loss function. In addition, we discuss how these oracle inequalities are related to the convergence of the value of the rule under an assumption describing the behaviour of the conditional treatment effect on the boundary of the decision threshold.

2. DATA AND NOTATION

Assume that individuals are monitored at time-points $t = \{1, \dots, K\}$. Let T denote a time-to-event outcome taking values in $\{1, \dots, K\} \cup \{\infty\}$, where $T = \infty$ represents no event

occurring in the follow-up period. Let $C \in \{0, \dots, K\}$ denote the censoring time, defined as the time at which the individual is last observed in the study, and let $C = K$ represent administrative censoring. Let $A \in \{0, 1\}$ denote study arm assignment, and let W denote a vector of baseline variables, which may include gene expression as well as demographic, comorbidity, and other clinical data. Let $\mathbb{1}(\cdot)$ denote the indicator variable taking value 1 if the argument is true and 0 otherwise. The observed-data vector for each participant is $O = (W, A, \Delta, \tilde{T})$, where $\tilde{T} = \min(C, T)$ and $\Delta = \mathbb{1}(T \leq C)$ is the indicator that the participant's event time is observed and uncensored. Throughout the paper we use c_1, c_2 and c_3 to denote different universal constants; no two occurrences of c_1, c_2 and c_3 should be taken to mean the same constant.

We assume that the observed-data vector for participant i , $O_i = (W_i, A_i, \Delta_i, \tilde{T}_i)$, is an independent and identically distributed draw from an unknown distribution P_0 on $(W, A, \Delta, \tilde{T})$. The empirical distribution of $O_1, \dots, O_n$ is denoted by P_n . We assume that $P_0 \in \mathcal{M}$, where $\mathcal{M}$ consists of all continuous densities on O with respect to a dominating measure ν . We use P to denote a generic distribution $P \in \mathcal{M}$, and use $E_0(\cdot)$ to denote expectation with respect to P_0 . For a function $f(o)$, we write $Pf = \int f(o) dP(o)$ and $\|f\|^2 = P_0 f^2$. We write $a \lesssim b$ to mean that a is smaller than or equal to b up to a universal constant.

We can equivalently encode a single participant's data vector O using the longitudinal data structure

$$O = (W, A, R_0, L_1, R_1, L_2, \dots, R_{K-1}, L_K),$$

where $R_t = \mathbb{1}(\tilde{T} = t, \Delta = 0)$ and $L_t = \mathbb{1}(\tilde{T} = t, \Delta = 1)$ for $t \in \{0, \dots, K\}$. We denote the history through time of a random variable X by $\bar{X}_t = (X_0, \dots, X_t)$. For a given scalar x , the expression $\bar{X}_t = x$ denotes elementwise equality.

Define the following indicator variables for each $t \geq 1$: $I_t = \mathbb{1}(\bar{R}_{t-1} = 0, \bar{L}_{t-1} = 0)$ and $J_t = \mathbb{1}(\bar{R}_{t-1} = 0, \bar{L}_t = 0)$. The variable I_t is the indicator of a participant being at risk of the event observed at time t . Analogously, J_t is the indicator, based on the outcome data through time t and censoring data before time t , of a participant being at risk of censoring at time t . We define $J_0 = 1$.

Define the discrete hazard function for survival at time $m \in \{1, \dots, K\}$ by

$$h(m, a, w) = \text{pr}(L_m = 1 \mid I_m = 1, A = a, W = w),$$

among the population at risk at time m within strata of study arm and baseline variables. Similarly, for the censoring variable C , define the censoring hazard at time $m \in \{0, \dots, K\}$ by

$$g_R(m, a, w) = \text{pr}(R_m = 1 \mid J_m = 1, A = a, W = w).$$

We use the notation $g_A(a, w) = \text{pr}(A = a \mid W = w)$, $g = (g_A, g_R)$ and $\eta = (h, g_A, g_R)$, where the probabilities are computed under $P \in \mathcal{M}$. We add a subscript 0 to g and h to denote the corresponding quantities under P_0 .

3. OPTIMAL INDIVIDUALIZED TREATMENT RULES

3.1. Potential outcomes and causal parameters

Define the potential outcomes $T_a, a \in \{0, 1\}$, as the event times that would have been observed had study arm assignment $A = a$ and censoring time $C = K$ been externally set with probability 1. For a restriction time $\tau \in \{1, \dots, K\}$ of interest, we define the restricted survival time under

treatment arm $A = a$ as $\min(T_a, \tau)$. For a transformation Z of W , the treatment effect within strata of the covariates Z may be defined in terms of the so-called full-data blip function (see, e.g., [Robins, 1997](#)) of the restricted mean survival time:

$$\theta_c(z) = E\{\min(T_1, \tau) - \min(T_0, \tau) \mid Z = z\}.$$

The subscript c denotes a causal parameter, that is, a parameter of the distribution of the potential outcomes T_1 and T_0 . The transformation Z may represent a subset of covariates, such as gene expression, or the whole vector W . The marginal treatment effect is defined as $E\{\min(T_1, \tau) - \min(T_0, \tau)\}$. It can be shown (see [Díaz et al., 2018](#)) that $E\{\min(T_a, \tau) \mid Z = z\} = \sum_{t=0}^{\tau-1} S_c(t, a, z)$, where $S_c(t, a, z) = \text{pr}(T_a > t \mid Z = z)$ is the survival probability corresponding to the potential outcome under assignment to arm $A = a$ within strata $Z = z$. As a result,

$$\theta_c(z) = \sum_{t=0}^{\tau-1} \{S_c(t, 1, z) - S_c(t, 0, z)\}, \quad (1)$$

since $S_c(0, a, z) = 1$ for $a \in \{0, 1\}$ and for all z .

An individualized treatment rule d maps the covariate values z of a participant to a personalized treatment decision in $\{0, 1\}$. The potential time to event under a rule d is defined as $T_d = d(Z)T_1 + \{1 - d(Z)\}T_0$. Accordingly, the restricted mean survival time under a treatment rule that assigns treatment according to $d(Z)$ is

$$E\{\min(T_d, \tau)\} = E[d(Z)\{\min(T_1, \tau) - \min(T_0, \tau)\}] + E\{\min(T_0, \tau)\}.$$

Because the last term does not depend on $d(z)$, we define the value of the rule d as

$$V_c(d) = E[d(Z)\{\min(T_1, \tau) - \min(T_0, \tau)\}] = E\{d(Z)\theta_c(Z)\}. \quad (2)$$

Expression (2) provides the basis for the definition of an optimal rule as

$$d_c = \arg \max_{d \in \mathcal{D}} V_c(d),$$

where $\mathcal{D}$ is the space of functions that map the range of Z to a treatment decision in $\{0, 1\}$. The optimal rule $d_c(z)$ is given pointwise by $\mathbb{1}\{\theta_c(z) > 0\}$. We define optimality of a rule with respect to the restricted mean survival time, though other effect measures could also be used.

3.2. Identification of parameters in terms of P_0

In this section we show how the blip function $\theta_c(z)$, the value function $V_c(d)$, and the optimal rule $d_c(z)$, which are defined above in terms of the distribution of potential outcomes, can be equivalently expressed as functions $\theta_0(z)$, $V_0(d)$, and $d_0(z)$ of the observed-data distribution $P_0(W, A, \Delta, \tilde{T})$ under Assumptions 1–4 below. This is useful since the potential outcomes are not always observed, in contrast to the observed-data vector $(W, A, \Delta, \tilde{T})$ for each participant, whose distribution we can make direct statistical inferences about.

We make the following assumptions:

Assumption 1 (Consistency). $T = \mathbb{1}(A = 0)T_0 + \mathbb{1}(A = 1)T_1$;

Assumption 2 (Randomization). A is independent of T_a conditional on W , for each $a \in \{0, 1\}$;

Assumption 3 (Random censoring). C is independent of T_a conditional on (A, W) , for each $a \in \{0, 1\}$;

Assumption 4 (Strong positivity). for each $a \in \{0, 1\}$ and $t \in \{0, \dots, \tau - 1\}$, $\text{pr}\{g_{A,0}(a, W) > \epsilon\} = 1$ and $\text{pr}\{g_{R,0}(t, a, W) < 1 - \epsilon\} = 1$, for some $\epsilon > 0$.

Assumptions 1–4 will be assumed to hold throughout. Denote the survival and censoring functions for T at time $t \in \{1, \dots, \tau - 1\}$ conditioned on study arm a and baseline variables w by

$$S(t, a, w) = \text{pr}(T > t \mid A = a, W = w), \quad G(t, a, w) = \text{pr}(C \geq t \mid A = a, W = w).$$

Under Assumptions 1–3, we have $T \perp\!\!\!\perp C \mid A, W$, so

$$S(t, a, w) = \prod_{m=1}^t \{1 - h(m, a, w)\}, \quad G(t, a, w) = \prod_{m=0}^{t-1} \{1 - g_R(m, a, w)\}.$$

Under Assumptions 1–4, the potential outcome survival function $S_c(t, a, z)$ can equivalently be represented in terms of the observed-data distribution as $S(t, a, z) = E\{S(t, a, W) \mid Z = z\}$. It follows from (1) that the causal parameter $\theta_c(z)$ is equal to the observed-data blip function

$$\theta(z) = \sum_{t=1}^{\tau-1} E \left[\prod_{m=1}^t \{1 - h(m, 1, W)\} - \prod_{m=1}^t \{1 - h(m, 0, W)\} \mid Z = z \right]. \quad (3)$$

Thus, the value $V_c(d)$ of a rule d is equal to $V(d) = E\{d(Z)\theta(Z)\}$, and a corresponding optimal treatment rule is equal to $d_0(z) = \mathbb{1}\{\theta_0(z) > 0\}$.

In addition to Assumptions 1–4, we sometimes make the following assumption, which is common in the classification literature for plug-in estimators.

Assumption 5 (Margin). There exists a constant $\lambda \geq 0$ such that $\text{pr}\{0 < |\theta_0(Z)| \leq t\} \lesssim t^\lambda$ for all $t > 0$.

The case $\lambda = 0$ is trivial. As $\lambda \rightarrow \infty$, the probability mass around $t = 0$ decreases, yielding as an extreme case a function $\theta_0(Z)$ bounded away from zero. Assumption 5 characterizes the behaviour of the decision function on the boundary, and has been shown to be crucial in establishing the convergence of certain classifiers (Audibert et al., 2007; Luedtke & Chambaz, 2017).

4. PLUG-IN ESTIMATION OF THE BLIP FUNCTION AND THE OPTIMAL RULE

In this section we discuss various estimators for $\theta_0(z)$, which can be mapped to a plug-in estimator of $d_0(z)$ through $d_0(z) = \mathbb{1}\{\theta_0(z) > 0\}$. Our general strategy relies on the censoring unbiased transformation given in Definition 1 below. This was introduced by Fan & Gijbels (1994) and is discussed in Rubin & van der Laan (2007), among other works.

DEFINITION 1 (Unbiased transformation). The transformation $D : \mathcal{O} \rightarrow \mathbb{R}$ is said to be unbiased for $\theta_0(z)$ if $E\{D(O) \mid Z = z\} = \theta_0(z)$.

Definition 1 motivates the construction of estimators of $\theta_0(z)$ by regressing the transformation $D(O)$ on the covariates Z . A common complication in this step is that most unbiased transformations typically depend on unknown nuisance parameters which must be estimated prior to the analysis. We focus on the doubly robust censoring unbiased transformation D_η defined in Lemma 1. In addition to being a doubly robust unbiased transformation for $\theta_0(z)$, D_η is an efficient estimating function in the nonparametric model in the sense that it may be used to construct efficient estimators of the marginal treatment effect (Díaz et al., 2018).

LEMMA 1 (Doubly robust censoring unbiased transformation). *Define*

$$D_\eta(O) = \sum_{m=1}^{\tau-1} [I_m Z(m, A, W) \{L_m - h(m, A, W)\} + S(m, 1, W) - S(m, 0, W)],$$

where $Z(m, A, W) = Z_1(m, A, W) - Z_0(m, A, W)$ with

$$Z_a(m, A, W) = - \sum_{t=m}^{\tau-1} \frac{\mathbb{1}\{A = a\}}{g_A(a, W) G(m, a, W)} \frac{S(t, a, W)}{S(m, a, W)} \quad (a = 0, 1).$$

Let $\eta = (h, g_A, g_R)$ be such that $h = h_0$ or $(g_A, g_R) = (g_{A,0}, g_{R,0})$. Then D_η is an unbiased transformation for $\theta_0(z)$, that is, $E_0\{D_\eta(O) \mid Z = z\} = \theta_0(z)$.

Lemma 1 implies that the expected value of the quadratic loss function $L_\eta(O; \theta) = \{D_\eta(O) - \theta(Z)\}^2$ is minimized at θ_0 if $\eta = (h, g)$ is such that either $h = h_0$ or $g = g_0$.

We denote the expected value of a loss function L_η by $R_{0,\eta}(\theta) = E_0\{L_\eta(O; \theta)\}$ and refer to it as the risk. We now discuss the construction of super-learning ensembles of candidate estimators for θ that target minimization of the quadratic risk. Consider a library of estimation algorithms for estimating θ_0 , $\mathcal{L} = \{\hat{\theta}_j : j = 1, \dots, J\}$. For an estimator $\hat{\eta}$ of η_0 , in light of § 3.2, $\mathcal{L}$ may be constructed by considering any predictive algorithm that minimizes the quadratic risk for prediction of the doubly robust unbiased transformation $D_{\hat{\eta}}(O)$. The literature provides us with a wealth of algorithms that may be used, including random forests, Bayesian regression trees, generalized additive models, local polynomial regression, multivariate adaptive regression splines, support vector machines, and neural networks.

Consider the following crossvalidation set-up. Let $\mathcal{V}_1, \dots, \mathcal{V}_K$ denote a random partition of the index set $\{1, \dots, n\}$ into K folds of approximately the same size: $\mathcal{V}_k \subset \{1, \dots, n\}$, $\bigcup_{k=1}^K \mathcal{V}_k = \{1, \dots, n\}$, and $\mathcal{V}_k \cap \mathcal{V}_{k'} = \emptyset$. In addition, for each k , the associated training sample is $\mathcal{T}_k = \{1, \dots, n\} \setminus \mathcal{V}_k$. Let $\hat{\eta}_k$ denote the estimator of η_0 trained using only data from $\mathcal{T}_k$. Likewise, let $\hat{\theta}_{j,k}$ denote the estimator of θ_0 obtained by training the j th predictive algorithm in $\mathcal{L}$ using only data from the sample $\mathcal{T}_k$. We let $k(i)$ denote the index of the validation set that contains observation i . The crossvalidated prediction risk of $\hat{\theta}_j$ is defined as

$$\hat{R}_{\hat{\eta}}(\hat{\theta}_j) = \frac{1}{K} \sum_{i=1}^n \frac{1}{|\mathcal{V}_{k(i)}|} L_{\hat{\eta}_{k(i)}}(O_i, \hat{\theta}_{j,k(i)}).$$

In this paper we consider an ensemble learner given by a convex combination

$$\hat{\theta}_\alpha(z) = \sum_{j=1}^J \alpha_j \hat{\theta}_j(z), \quad \alpha_j \geq 0, \quad \sum_{j=1}^J \alpha_j = 1, \quad (4)$$

with weights α_j chosen to minimize the crossvalidated risk of (4):

$$\hat{\alpha} = \arg \min_{\alpha} \sum_{i=1}^n \frac{1}{|\mathcal{V}_{k(i)}|} \left\{ D_{\hat{\eta}_{k(i)}}(O_i) - \sum_{j=1}^J \alpha_j \hat{\theta}_{j,k(i)}(V_i) \right\}^2 \quad \text{subject to} \quad \alpha_j \geq 0, \quad \sum_{j=1}^J \alpha_j = 1. \quad (5)$$

Expression (5) is a weighted ordinary least squares problem with constraints on the coefficients, and may be solved using standard software. We denote this super-learner by $\hat{\theta}_{\text{sl}} = \hat{\theta}_{\hat{\alpha}}$.

The optimality of general crossvalidation selection procedures is discussed in [van der Laan et al. \(2006\)](#) and [van der Vaart et al. \(2006\)](#). Here optimality is defined in terms of asymptotic equivalence to the oracle risk, which we define as the risk computed when one of the components of the nuisance parameter η_0 is known and a validation sample of infinite size is available to assess the performance of the estimator. We describe this specifically in the following definition.

DEFINITION 2 (Oracle risk and oracle selector). *Let $\eta_1 = (g_1, h_1)$, where either $g_1 = g_0$ or $h_1 = h_0$. The oracle risk of a candidate $\hat{\theta}_{\alpha}$ is defined as*

$$\tilde{R}_{\eta_1}(\hat{\theta}_{\alpha}) = \frac{1}{K} \sum_{k=1}^K \int \{D_{\eta_1}(o) - \hat{\theta}_{\alpha,k}(z)\}^2 dP_0(o),$$

where $\hat{\theta}_{\alpha,k}$ denotes the estimator $\hat{\theta}_{\alpha}$ applied to sample $\mathcal{T}_k$. This is a random quantity. The oracle selector is

$$\tilde{\alpha} = \arg \min_{\alpha} \tilde{R}_{\eta_1}(\hat{\theta}_{\alpha}) \quad \text{subject to} \quad \alpha_j \geq 0, \quad \sum_{j=1}^J \alpha_j = 1,$$

and the corresponding oracle blip function is denoted by $\hat{\theta}_{\text{or}} = \hat{\theta}_{\tilde{\alpha}}$.

The risk $\tilde{R}_{\eta_1}(\theta_0) = \int L_{\eta_1}(o; \theta_0) dP_0(o)$ is the optimal risk with respect to the loss function L_{η_1} , achieved by the true θ_0 . Theorem 1 provides a bound on the excess risk of the estimator $\hat{\theta}_{\text{sl}}$ and the excess risk of $\hat{\theta}_{\text{or}}$. The excess risk for a selector $\hat{\alpha}$ is defined as the difference between the oracle risk of the selector $\hat{\alpha}$ and the optimal risk, i.e.,

$$\mathcal{E}^2(\hat{\theta}_{\hat{\alpha}}) = E\{\tilde{R}_{\eta_1}(\hat{\theta}_{\hat{\alpha}}) - \tilde{R}_{\eta_1}(\theta_0)\} = E\{P_0(\hat{\theta}_{\hat{\alpha}} - \theta_0)^2\},$$

where the expectation is taken over draws of $O_1, \dots, O_n$. Theorem 1 shows that the square root of $\mathcal{E}^2(\hat{\theta}_{\text{sl}})$ is bounded by two terms: one depending on the excess risk of the oracle selector $\mathcal{E}^2(\hat{\theta}_{\text{or}})$ and the other depending on doubly robust terms associated with estimation of η_0 .

THEOREM 1 (Oracle inequality for the super-learner of the blip function). *Let $\eta_1 = (g_1, h_1)$ denote the elementwise $L_2(P_0)$ limit of $\hat{\eta}$ as $n \rightarrow \infty$, and assume that either $g_1 = g_0$ or $h_1 = h_0$. Define*

$$B_1(\hat{\eta}, \eta_0) = E\|(\hat{g} - g_0)(\hat{h} - h_0)\|,$$

$$B_2(\hat{\eta}, \eta_0) = E\{\mathbb{1}(g_1 = g_0)\|\hat{g} - g_0\| + \mathbb{1}(h_1 = h_0)\|\hat{h} - h_0\|\}^2.$$

Then, for $\delta > 0$,

$$\begin{aligned} \mathcal{E}(\hat{\theta}_{\text{sl}}) &\leq (1 + 2\delta)^{1/2} \mathcal{E}(\hat{\theta}_{\text{or}}) + c_1 \{(1 + \log n)/n\}^{1/2} \\ &\quad + c_2 B_1(\hat{\eta}, \eta_0) + c_3 (\log n/n)^{1/4} \{B_2(\hat{\eta}, \eta_0)\}^{1/2} \end{aligned}$$

for universal constants c_1 , c_2 and c_3 that depend on δ , defined in the [Supplementary Material](#).

The terms $B_1(\hat{\eta}, \eta_0)$ and $B_2(\hat{\eta}, \eta_0)$ converge to zero if either $\hat{g}$ or $\hat{h}$ converge to g_0 or h_0 , respectively, in the $L_2(P_0)$ -norm, so the doubly robust property of $D_\eta(o)$ is transferred to the oracle inequality. To the best of our knowledge, this result has not previously been shown.

The super-learner $\hat{\theta}_{\text{sl}}$ may be used to construct a plug-in estimator $\hat{d}(z) = \mathbb{1}\{\hat{\theta}_{\text{sl}}(z) > 0\}$ of the optimal rule. The following remark discusses the convergence rates of the value of the selected rule $\hat{d}(z)$ to the value of the oracle rule $\tilde{d}(z) = \mathbb{1}\{\hat{\theta}_{\text{or}}(z) > 0\}$.

Remark 1 (Convergence rates to the oracle value). Assume that

$$B_1(\hat{\eta}, \eta_0) = O\{(\log n/n)^{1/2}\}, \quad B_2(\hat{\eta}, \eta_0) = O\{(\log n/n)^{1/2}\}.$$

Lemma 5.3 of [Audibert et al. \(2007\)](#), along with Jensen's inequality, shows that under Assumption 5,

$$E\{V_0(\tilde{d}) - V_0(\hat{d})\} \lesssim \{\mathcal{E}^2(\hat{\theta}_{\text{sl}}) - \mathcal{E}^2(\hat{\theta}_{\text{or}})\}^{(1+\lambda)/(2+\lambda)}.$$

This yields the convergence rate

$$E\{V_0(\tilde{d}) - V_0(\hat{d})\} = O\{(\log n/n)^{(1+\lambda)/(2+\lambda)}\}.$$

An example giving the above rate would be a randomized study with no censoring. In this case, substituting the true values $g_{A,0}(w) = q \in (0, 1)$ and $g_{R,0}(t, a, w) = 0$ for $\hat{g}_A(a)$ and $\hat{g}_R(t, a, w)$, and assuming that $\hat{h}$ is inconsistently estimated, yields $B_1(\hat{\eta}, \eta_0) = B_2(\hat{\eta}, \eta_0) = 0$. Under $\lambda = 0$, implying no assumption, we get a convergence rate of $(\log n/n)^{1/2}$. Under $\lambda = \infty$, implying that $\theta_0(z)$ is bounded away from zero, we get a rate of $\log n/n$. Alternatively, a logistic regression fit of A on $\mathcal{W}$ containing at least an intercept would yield an estimator satisfying $\|\hat{g}_A - g_{A,0}\|^2 = O_p(n^{-1})$ and bias terms $B_1(\hat{\eta}, \eta_0) = O(n^{-1/2})$ and $B_2(\hat{\eta}, \eta_0) = O(n^{-1/2})$. It has been established that fitting an estimator to the treatment mechanism when it is known improves the efficiency of marginal effect estimates ([Robins et al., 1994](#); [van der Laan & Robins, 2003](#)).

Remark 1 establishes the convergence of $V(\hat{d})$ to the oracle $V(\tilde{d})$. This is different from the typical result in the classification literature, which concerns convergence to the optimal value $V_0(d_0)$. The latter kind of result often involves fast learning rates, sometimes faster than n^{-1} . Such fast rates often require the class of allowable blip functions to be Hölder ([Audibert et al., 2007](#)) or the class of allowable rules to be Donsker ([Luedtke & Chambaz, 2017](#)). Our results impose neither of these restrictions, and thus entail slower convergence rates.

5. SUPER-LEARNER ENSEMBLES USING CLASSIFICATION

5.1. Estimators using the 0-1 loss function

In this section we discuss a classification approach that aims at directly estimating the optimal rule $d_0(z)$. Our approach here differs from § 4 in that we do not attempt to estimate the blip function. Instead, we introduce the concept of a decision function $f : Z \rightarrow \mathbb{R}$, which yields a

treatment rule $d_f(z) = \mathbb{1}\{f(z) > 0\}$. With a slight abuse of notation we write $V(f)$ to refer to the value of the rule d_f . Any function f_0 such that $\text{sign}\{f_0(z)\theta_0(z)\} = 1$ for all z has optimal value $V_0(d_0)$. This provides intuition about the benefits of directly optimizing the value of the loss function rather than the risk of the blip function: an inconsistent estimator of the blip function may provide an optimal rule, as long as its sign is correct. For a given rule d_f , in light of Lemma 1, we have that $V_0(f) = E_0\{d_f(Z)D_\eta(O)\}$ if η is such that either $h = h_0$ or $g = g_0$. Thus, a decision function that optimizes the value of the rule d_f is

$$f_0 \in \arg \max_f \int d_f(z) D_\eta(o) dP_0(o).$$

For a binary value $b \in \{0, 1\}$ and any X , we have $bX = \mathbb{1}(X > 0)|X| - |X|\mathbb{1}\{\mathbb{1}(X > 0) \neq b\}$. Thus, the optimization problem may be recast as $f_0 \in \mathcal{F}_0$, where $\mathcal{F}_0 = \arg \min_f \int L_\eta(o; f) dP_0$ and

$$L_\eta(o; f) = |D_\eta(o)| \mathbb{1}[\mathbb{1}\{D_\eta(o) > 0\} \neq d_f(z)]. \quad (6)$$

Expression (6) is a weighted classification loss function in which we aim to classify the binary outcome $\mathbb{1}\{D_\eta(O) > 0\}$ based on data Z , using the 0-1 loss function with weights given by $|D_\eta(O)|$. The objective is to classify an individual who benefits from treatment arm $A = 1$, i.e., an individual with $D_\eta(O) > 0$, as requiring treatment, i.e., $d_f(Z) = 1$, while penalizing for the loss $|D_\eta(O)|$ incurred if the individual were misclassified.

In what follows we consider a library of algorithms for estimating the decision function $\mathcal{L} = \{\hat{f}_j(z) : j = 1, \dots, J\}$. The most natural choice for a decision function is the blip function $\theta_0(z)$, but the library may contain other decision functions such as support vector machines (Zhao et al., 2015) and parametric decision functions (Bai et al., 2016).

We construct an ensemble of the decision functions as

$$\hat{f}_\alpha(z) = \sum_{j=1}^J \alpha_j \hat{f}_j(z), \quad \alpha_j \geq 0. \quad (7)$$

In this way, we generate an ensemble optimal rule as $\hat{d}_\alpha(z) = \mathbb{1}\{\hat{f}_\alpha(z) > 0\}$. As in the previous section, we define the super-learner selector as

$$\hat{\alpha} \in \arg \min_{\alpha} \sum_{i=1}^n \frac{1}{|\mathcal{V}_{k(i)}|} L_{\hat{\eta}_k(i)}(O_i, \hat{f}_{\alpha, k(i)}) \quad \text{subject to} \quad \alpha_j \geq 0,$$

where $\hat{f}_{\alpha, k(i)}$ represents (7) with $\hat{f}_j(z)$ replaced by $\hat{f}_{j, k(i)}(z)$, the j th decision function estimated using the training sample $\mathcal{T}_{k(i)}$. The super-learner of the decision function is defined as $\hat{f}_{\text{sl}}(z) = \hat{f}_{\hat{\alpha}}(z)$, and the corresponding optimal rule is defined as $\hat{d}_{\text{sl}} = \mathbb{1}\{\hat{f}_{\text{sl}}(z) > 0\}$.

For $\eta_1 = (g_1, h_1)$ such that either $g_1 = g_0$ or $h_1 = h_0$, the oracle risk of the decision function is defined as

$$\tilde{R}_{\eta_1}(\hat{f}) = \frac{1}{K} \sum_{k=1}^K \int L_{\eta_1}(o, \hat{f}_k) dP_0(o).$$

The oracle selector of α is thus defined as $\tilde{\alpha} \in \arg \min_{\alpha} \tilde{R}_{\eta_1}(\hat{f}_{\alpha})$, and we write $\hat{f}_{\text{or}} = \hat{f}_{\tilde{\alpha}}$. The excess risk of an estimator $\hat{f}$ is equal to

$$\mathcal{E}(\hat{f}) = E\{\tilde{R}_{\eta_1}(\hat{f}) - \tilde{R}_{\eta_1}(f_0)\} = V_0(f_0) - E\{V_0(\hat{f})\}.$$

Theorem 2 provides bounds on $\mathcal{E}(\hat{f}_{\text{sl}})$ in terms of the excess risk of the oracle selector $\mathcal{E}(\hat{f}_{\text{or}})$ and the bias terms $B_1(\hat{\eta}, \eta_0)$ and $B_2(\hat{\eta}, \eta_0)$ defined in Theorem 1.

THEOREM 2 (Oracle inequality for the super-learner of the optimal rule). *Assume the conditions of Theorem 1. In addition, assume that $\hat{\alpha}$ is computed on a grid of size Mn^q for some $M > 0$ and $q > 0$. Then*

$$0 \leq \mathcal{E}(\hat{f}_{\text{sl}}) \leq \mathcal{E}(\hat{f}_{\text{or}}) + c_1(\log n/n)^{1/2} + c_2 B_1(\hat{\eta}, \eta_0).$$

If Assumption 5 holds with $\lambda = \infty$, then for $\delta > 0$,

$$0 \leq \mathcal{E}(\hat{f}_{\text{sl}}) \leq (1 + 2\delta) \mathcal{E}(\hat{f}_{\text{or}}) + c_1(1 + \log n)/n + c_2 B_1(\hat{\eta}, \eta_0) + c_3(\log n/n)^{1/2} B_2(\hat{\eta}, \eta_0),$$

where c_1 , c_2 and c_3 are universal constants that depend on δ , defined in the [Supplementary Material](#). Here B_1 and B_2 are defined as in Theorem 1.

Remark 2. Let $B_1(\hat{\eta}, \eta_0)$ converge as in Remark 1. An immediate consequence of the above theorem is that under no margin assumption, we have $E\{V_0(\hat{d}_{\text{or}}) - V_0(\hat{d}_{\text{sl}})\} = O\{(\log n/n)^{1/2}\}$. Under the strong margin assumption, Assumption 5, with $\lambda = \infty$, we have $E\{V_0(\hat{d}_{\text{or}}) - V_0(\hat{d}_{\text{sl}})\} = O(\log n/n)$. These rates are identical to those obtained in Remark 1 for the plug-in estimator. The question of whether analogous convergence rates can be obtained for other values of λ under Assumption 5 remains an open problem.

Remark 3. [Athey & Wager \(2017\)](#) proposed estimating the optimal rule as

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \sum_{i=1}^n \frac{1}{|\mathcal{V}_{k(i)}|} L_{\hat{\eta}_k(i)}(O_i, f).$$

They showed that the value of this estimator may achieve convergence rates as fast as $O_p(n^{-1/2})$, under assumptions on the entropy bound of the class $\mathcal{F}$, such as those satisfied by Vapnik–Chervonenkis classes.

In comparison to Theorem 1, Theorem 2 also requires that the optimization of the loss function be carried out on a finite grid. However, the size of the grid is allowed to grow polynomially in n , in order to minimize any possible asymptotic implications of this additional condition. Inspection of the proofs of the theorems reveals why the additional assumption is needed: the 0-1 loss function is nonsmooth and the Lipschitz condition used in the proof of Theorem 1 does not apply. This assumption is likely to have few practical consequences, but it is unclear to us whether it can be removed.

5.2. Using a surrogate for the 0-1 loss

Minimizing (6) is generally difficult due to the discontinuity and nonconvexity of the 0-1 loss. A common approach to mitigating the issues that arise from the discontinuity and nonconvexity

of the 0-1 loss function is to use a surrogate, such as the logistic loss $\phi(x) = \log\{1 + \exp(-x)\}$ or the hinge loss $\phi(x) = \max(1 - x, 0)$. The following result tells us that any decision function $d_{f_0}(z)$ based on a decision function $f_0 \in \mathcal{F}_{\phi,0}$ has the same performance as the optimal rule $d_0(z)$.

LEMMA 2. Assume η is such that either $h = h_0$ or $g = g_0$. Define

$$\mathcal{F}_{\phi,0} = \arg \min_f \int L_{\phi,\eta}(o;f) dP_0(o),$$

where the surrogate loss is

$$L_{\phi,\eta}(o;f) = |D_\eta(o)| \phi(f(z)[2\mathbb{1}\{D_\eta(o) > 0\} - 1]).$$

Then $\mathcal{F}_{\phi,0} \subseteq \mathcal{F}_0$, where $\mathcal{F}_0 = \arg \min_f \int L_\eta(o;f) dP_0$.

6. OPTIMAL TREATMENT FOR BREAST CANCER PATIENTS

6.1. Description of the problem

Different types of human breast cancer tumours respond differently to treatments (Perou et al., 2000; Sotiriou & Pusztai, 2009). Amplification of the ERBB2 gene and associated overexpression of the human epidermal growth factor receptor HER2 encoded by this gene occur in 25–30% of breast cancers (Slamon et al., 2001). HER2-positive breast cancer is an aggressive form of the disease and the prognosis for patients with this type of cancer is generally poor (Slamon et al., 1987; Seshadri et al., 1993). The clinical efficacy of adjuvant trastuzumab, a recombinant monoclonal antibody, in early stage HER2-positive patients has been demonstrated by several large clinical trials (Romond et al., 2005; Perez et al., 2011). Despite significant improvement in disease-free and overall survival of patients treated with trastuzumab, 20–25% of patients relapse within three to five years (Perez et al., 2011). In this paper we use data from the North Central Cancer Treatment Group, NCCTG, N9831 study, a phase III randomized clinical trial testing the addition of trastuzumab to chemotherapy in stage I–III HER2-positive breast cancer, to estimate the optimal treatment rule for HER2-positive cancer patients as a function of mRNA, demographic, and clinical data.

The total number of patients enrolled in the NCCTG N9831 trial was 3505. Samples from 1390 patients for whom there was available tissue were used to quantify mRNA from a custom codeset of 730 genes created by experts. The available baseline variables may thus be categorized into three classes: demographic, such as race, age and ethnicity; clinical, such as tumour grade, tumour size, nodal status, and hormone receptor status; and gene expression. Among the 1390 patients, 483 received chemotherapy alone and 907 received chemotherapy plus trastuzumab.

The clinical challenge is to identify genetic and demographic profiles of patients with HER2-positive breast cancer who are unlikely to benefit from adjuvant trastuzumab.

In order to estimate and assess the performance of the estimated rule using different datasets, we split our data into training and validation datasets, of sizes 1000 and 390, respectively.

6.2. Estimators of h and g_R

Our theoretical results show that the optimality of the estimated treatment rules hinges upon consistent estimation of at least one of the nuisance parameters h and g_R , so it is crucial to employ flexible methods capable of unveiling complex patterns which are not visible to the human eye. As demonstrated in § 6.4, simple models fail to detect such complex relations.

Table 1. *Coefficients of the super-learner ensemble for estimation of g_R and h*

		RF	XGB	MLP	GLM	MARS	LASSO
$\hat{g}_R$	$A = 1$	0.100	0.056	0.000	0.000	0.312	0.532
	$A = 0$	0.000	0.301	0.000	0.000	0.294	0.405
$\hat{h}$	$A = 1$	0.023	0.050	0.000	0.000	0.237	0.691
	$A = 0$	0.154	0.086	0.098	0.060	0.123	0.480

RF, random forests; XGB, extreme gradient boosting; MLP, multilayer perceptron; GLM, logistic regression; MARS, multivariate adaptive splines; LASSO, L_1 -regularized logistic regression.

In order to accurately estimate the nuisance parameters, we use an ensemble learner known as the super-learner for prediction ([van der Laan et al., 2007](#)). We train the ensemble separately using data from each treatment arm to account fully for treatment-covariate interactions. Like our rule ensembles, super-learning predictors build a combination of candidate predictors that minimize a crossvalidated user-supplied risk function. Since g_R and h are conditional probabilities, we focus on logistic regression ensembles and the negative loglikelihood loss function, using the R ([R Development Core Team, 2018](#)) implementation in the SuperLearner package ([Polley et al., 2016](#)). The candidate estimators included in the ensembles are listed in Table 1, along with the coefficients of each predictor in the ensemble, when trained on the complete dataset using five-fold crossvalidation.

For random forests, extreme gradient boosting, and multilayer perceptron, the tuning parameters are tuned using data splitting with the aid of the R package caret ([Kuhn et al., 2016](#)). To avoid $p > n$, logistic regression and multivariate adaptive splines are estimated with a variable screening algorithm which computes univariate t -statistics and keeps only the 50 variables with larger values.

6.3. Candidate estimators for the optimal treatment rule

According to our discussion in § 4 and § 5, there are at least three types of estimators for the optimal rule d_0 . The first type is a simple substitution estimator, obtained by inspection of equation (3), which consists in regressing the blip function $\hat{B}(W) = \sum_{t=1}^{T-1} \{\hat{S}(t, 1, W) - \hat{S}(t, 0, W)\}$ on Z , where $\hat{S}$ is the estimator of the survival function corresponding to the estimator $\hat{h}$ described in § 6.2. The second type is obtained through regression of the unbiased transformation $D_{\hat{\eta}}(O)$ on Z . The third type of estimation method is based on (6), and is obtained by classifying the binary outcome $\mathbb{1}\{D_{\hat{\eta}}(O) > 0\}$ as a function of Z , with weights given by $|D_{\hat{\eta}}(O)|$. Here $\hat{\eta} = (\hat{g}_R, \hat{h})$, where the components of $\hat{\eta}$ are as described in § 6.2. Any regression or supervised classification technique available in the statistical learning literature may be used as a candidate for solving these problems.

The coefficients of each candidate estimator in each ensemble are presented in Table 2. The rows represent the candidates for estimating d_0 . The columns represent the loss functions used to train the super-learning ensembles. These coefficients were computed using the Subplex routine, developed in a 1990 University of Texas at Austin PhD thesis by T. Rowan and implemented in the nonlinear optimization package NLOpt in R. For improved robustness, the 0-1 loss was optimized using 1000 different random starting values.

6.4. Assessing the performance of the estimated treatment rule

Once each rule has been estimated using only data from the training dataset, its value $V(\hat{d})$ is estimated on the validation dataset. We use a targeted minimum loss-based estimator of the restricted mean survival time ([Moore & van der Laan, 2011](#); [Díaz et al., 2018](#)).

Table 2. Coefficients of each candidate in each ensemble standardized to sum to 1; the rows represent the candidates, and the columns represent the ensembles

	SL-Reg	SL-Class-0-1	SL-Class-Hinge	SL-Class-Log
D-Class-RF	0.000	0.005	0.000	0.000
D-Class-XGB	0.792	0.001	0.945	0.869
D-Class-GLM	0.000	0.031	0.000	0.000
D-Reg	0.007	0.017	0.001	0.006
B-Reg	0.201	0.947	0.054	0.125

B-Reg, regression of the blip function $\hat{B}(W)$ using super-learning with candidate learners as described in Table 1; D-Reg, regression of the doubly robust transformation $D_{\hat{\eta}}(O)$ using super-learning with candidate learners as described in Table 1; D-Class-RF, weighted classification of $\mathbb{1}\{D_{\hat{\eta}}(O) > 0\}$ using random forests; D-Class-XGB, weighted classification of $\mathbb{1}\{D_{\hat{\eta}}(O) > 0\}$ using extreme gradient boosting; D-Class-GLM, weighted classification of $\mathbb{1}\{D_{\hat{\eta}}(O) > 0\}$ using logistic regression; SL-Reg, regression ensemble minimizing the expected quadratic loss function; SL-Class-0-1, classification ensemble minimizing the expected 0-1 loss function; SL-Class-Hinge, classification ensemble with surrogate hinge loss function; SL-Class-Log, classification ensemble with surrogate log loss function.

Fig. 1. Restricted mean survival time estimated in the validation set, for different estimated rules. The bars represent 95% confidence intervals. Letters indicate the point estimate and represent different estimator types: R for regression, C for classification, E for ensemble, and S for static.

Figure 1 presents the estimated restricted mean survival time obtained using each estimated rule, along with 95% confidence intervals. We also report the value of two static rules of interest: never treat and always treat. The best algorithm in our application is regression of the blip function. All super-learning ensembles yield a similar value, demonstrating the oracle property of the super-learner. Treating patients according to the optimal rule yields a restricted mean survival of 157.1 months, with a standard error of 3.1. In comparison with the always treat rule, which gives 151.2 months with a standard error of 3.3, the optimal rule improves mean patient survival by six months.

According to Table 2, only the super-learning ensemble based on the 0-1 loss assigns a large weight to the best algorithm. In fact, its restricted mean survival time is identical to that of the optimal rule. The other ensembles assign more weight to the second-best algorithm, weighted classification using extreme gradient boosting, and have slightly smaller restricted mean survival time.

The estimated rules based on weighted classification using random forests, logistic regression, and regression of the function $D_{\hat{\eta}}$ yield restricted mean survival times smaller than or equal to the restricted mean survival time of the static rule of always treat.

Table S1 in the [Supplementary Material](#) shows the p -values for pairwise comparisons of the value of each estimated rule. The ensemble using the 0-1 loss function outperforms the other ensembles; the difference in values is small but significant at the 5% level. The value of the optimal rule, which is obtained through regression of the blip function, is significantly different from the values given by all the other rules, except for the ensemble using the 0-1 loss function. This illustrates the theoretical property of super-learning, i.e., that the risk of the ensemble converges to the risk of the best candidate in the library. Weighted classification using logistic regression, which is often advocated because it yields parsimonious rules ([Zhang et al., 2015](#)), has a value significantly lower than the static rule of always treat.

In our application, we used data splitting to train and assess the performance of the estimated rules. This approach may be unnecessary, since the value of the rule may be assessed using the training dataset, under certain conditions derived by [Luedtke & van der Laan \(2016a\)](#).

7. DISCUSSION

Although we found no asymptotic differences between the plug-in estimator optimizing the prediction error of the blip function and a weighted classification approach, the classification ensemble using the 0-1 loss function yielded better treatment rules than the other approaches in our illustrative application. The superiority of the classification approach has been recognized before (e.g., [Zhao et al., 2012](#)), and is a consequence of the fact that it emphasizes optimizing the decision rule rather than prediction accuracy as in the blip approach.

We consider a survival time measured on a discrete time scale. Most clinical research studies measure time to event on a discrete scale. In our motivating application, time to relapse of cancer or death was measured in days. We foresee no technical difficulties in extending our approach to deal with a continuous time to event. This can be achieved by replacing discrete-time hazards with their continuous counterparts, as well as replacing certain sums over time by the appropriate martingale integrals ([Bai et al., 2016](#)) in the definition of the censoring unbiased transformation $D_{\eta}(O)$. A potential practical limitation is that the software and literature for data-adaptive machine learning estimation of continuous-time hazards is scarce in comparison to that of binary classification, which may be used for estimation of discrete-time hazards. Among the few methods that can be used for this problem are semiparametric models such as Cox regression and accelerated failure time models. Available data-adaptive approaches include survival random forests and regularized Cox regression. If time is measured on a continuous scale, implementation of our methods requires discretization. The specific choice of the discretization intervals may be guided by what is clinically relevant. For example, in cancer research, the clinically relevant scale would typically be a week or a month. In the absence of clinical criteria to guide the choice of discretization level, a concern is that too coarse of a discretization may lead to loss of relevant information. A question for future research is how to optimally set the level of discretization in order to trade off information loss against estimator precision. Another area for future research is to consider discretization levels that get finer with sample size.

We have derived doubly robust oracle inequalities and convergence rates assuming an exposure of interest that occurs at baseline and censoring which is confounded with the time to event only by baseline variables. We conjecture that our general results apply to the more general case of a dynamic treatment regime with a time-varying treatment and time-varying confounders. Such settings will be the subject of future research.

In our definition of oracle risk and oracle selector we have used a nuisance parameter $\eta_1 = (g_1, h_1)$ satisfying either $h_1 = h_0$ or $g_1 = g_0$. Therefore these oracle quantities change depending on which of the two nuisance parameters is correctly specified. According to efficient estimation theory in semiparametric models, we expect the case $\eta_1 = \eta_0$ to yield oracle quantities with minimal variability. In the single-misspecification case in which $h_1 = h_0$ or $g_1 = g_0$ but not both, it is unclear to us whether misspecification of one of the models would yield better results than the other. Lastly, our set-up includes as a particular case the inverse probability weighted loss function, which may be obtained by using a constant estimator $\hat{h}(t, a, w) = 1$, as well as the g-computation loss function, which is obtained by using $\hat{g}_A(a, w) = 1$ and $\hat{g}_R(t, a, w) = 0$.

ACKNOWLEDGEMENT

We are grateful to the editor, associate editor and referees for their insightful comments. We would also like to thank Mark van der Laan for useful discussion. This research was supported in part by awards from the National Cancer Institute of the U.S. National Institutes of Health to the Alliance for Clinical Trials in Oncology.

SUPPLEMENTARY MATERIAL

[Supplementary material](#) available at *Biometrika* online includes proofs of the theoretical results presented here.

REFERENCES

- ATHEY, S. & WAGER, S. (2017). Efficient policy learning. *arXiv*: 1702.02896.
- AUDIBERT, J.-Y. & TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35**, 608–33.
- BAL, X., TSIATIS, A. A., LU, W. & SONG, R. (2016). Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime Data Anal.* **23**, 1–20.
- DÍAZ, I., COLANTUONI, E., HANLEY, D. F. & ROSENBLUM, M. (2018). Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. *Lifetime Data Anal.* doi: 10.1007/s10985-018-9428-5.
- FAN, J. & GIJBELS, I. (1994). Censored regression: Local linear approximations and their applications. *J. Am. Statist. Assoc.* **89**, 560–70.
- GENG, Y., ZHANG, H. H. & LU, W. (2015). On optimal treatment regimes selection for mean survival time. *Statist. Med.* **34**, 1169–84.
- GOLDBERG, Y. & KOSOROK, M. R. (2012). Q-learning with censored data. *Ann. Statist.* **40**, 529–60.
- KUHN, M., WING, J., WESTON, S., WILLIAMS, A., KEEFER, C., ENGELHARDT, A., COOPER, T., MAYER, Z., KENKEL, B., R DEVELOPMENT CORE TEAM et al. (2016). *caret: Classification and Regression Training*. R package version 6.0-73, <https://CRAN.R-project.org/package=caret>.
- LUEDTKE, A. & CHAMBAZ, A. (2017). Faster rates for policy learning. *arXiv*: 1704.06431.
- LUEDTKE, A. R. & VAN DER LAAN, M. J. (2016a). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44**, 713–42.
- LUEDTKE, A. R. & VAN DER LAAN, M. J. (2016b). Super-learning of an optimal dynamic treatment rule. *Int. J. Biostatist.* **12**, 305–32.
- MCKEAGUE, I. W. & QIAN, M. (2014). Estimation of treatment policies based on functional predictors. *Statist. Sinica* **24**, 1461–85.
- MOORE, K. L. & VAN DER LAAN, M. J. (2009). Increasing power in randomized trials with right censored outcomes through covariate adjustment. *J. Biopharm. Statist.* **19**, 1099–131.
- MOORE, K. L. & VAN DER LAAN, M. J. (2011). RCTs with time-to-event outcomes. In *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer, pp. 259–69.
- PEREZ, E. A., ROMOND, E. H., SUMAN, V. J., JEONG, J.-H., DAVIDSON, N. E., GEYER JR, C. E., MARTINO, S., MAMOUNAS, E. P., KAUFMAN, P. A. & WOLMARK, N. (2011). Four-year follow-up of trastuzumab plus adjuvant chemotherapy for operable human epidermal growth factor receptor 2-positive breast cancer: Joint analysis of data from NCCTG N9831 and NSABP B-31. *J. Clin. Oncol.* **29**, 3366–73.

- PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H., AKSLEN, L. A. et al. (2000). Molecular portraits of human breast tumours. *Nature* **406**, 747–52.
- POLLEY, E., LEDELL, E. & VAN DER LAAN, M. (2016). *SuperLearner: Super Learner Prediction*. R package version 2.0-19, <https://CRAN.R-project.org/package=SuperLearner>.
- QIAN, M. & MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39**, 1180–210.
- R DEVELOPMENT CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org>.
- ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*. New York: Springer, pp. 69–117.
- ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.
- ROMOND, E. H., PEREZ, E. A., BRYANT, J., SUMAN, V. J., GEYER JR, C. E., DAVIDSON, N. E., TAN-CHIU, E., MARTINO, S., PAIK, S., KAUFMAN, P. A. et al. (2005). Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *New Engl. J. Med.* **353**, 1673–84.
- RUBIN, D. & VAN DER LAAN, M. J. (2007). A doubly robust censoring unbiased transformation. *Int. J. Biostatist.* **3**, 1–19.
- RUBIN, D. B. & VAN DER LAAN, M. J. (2012). Statistical issues and limitations in personalized medicine research with clinical trials. *Int. J. Biostatist.* **8**, 1–20.
- SESHADRI, R., FIRGAIRA, F., HORSFALL, D., MCCAUL, K., SETLUR, V. & KITCHEN, P. (1993). Clinical significance of HER-2/neu oncogene amplification in primary breast cancer. The South Australian Breast Cancer Study Group. *J. Clin. Oncol.* **11**, 1936–42.
- SLAMON, D. J., CLARK, G. M., WONG, S. G., LEVIN, W. J., ULLRICH, A. & MCGUIRE, W. L. (1987). Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177–82.
- SLAMON, D. J., LEYLAND-JONES, B., SHAK, S., FUCHS, H., PATON, V., BAJAMONDE, A., FLEMING, T., EIERMANN, W., WOLTER, J., PEGRAM, M. et al. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New Engl. J. Med.* **344**, 783–92.
- SONG, R., KOSOROK, M., ZENG, D., ZHAO, Y., LABER, E. & YUAN, M. (2015). On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat* **4**, 59–68.
- SOTIRIOU, C. & PUSZTAI, L. (2009). Gene-expression signatures in breast cancer. *New Engl. J. Med.* **360**, 790–800.
- VAN DER LAAN, M., POLLEY, E. & HUBBARD, A. (2007). Super learner. *Statist. Appl. Genet. Molec. Biol.* **6**, article 25.
- VAN DER LAAN, M. J., DUDOIT, S. & VAN DER VAART, A. W. (2006). The cross-validated adaptive epsilon-net estimator. *Statist. Decis.* **24**, 373–95.
- VAN DER LAAN, M. J. & ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.
- VAN DER VAART, A. W., DUDOIT, S. & VAN DER LAAN, M. J. (2006). Oracle inequalities for multi-fold cross-validation. *Statist. Decis.* **24**, 351–71.
- WOLPERT, D. H. (2002). The supervised learning no-free-lunch theorems. In *Soft Computing and Industry: Recent Applications*, R. Roy, M. Köppen, S. Ovaska, T. Furuhashi & F. Hoffmann, eds. London: Springer, pp. 25–42.
- ZHANG, Y., LABER, E. B., TSIATIS, A. & DAVIDIAN, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* **71**, 895–904.
- ZHAO, Y., ZENG, D., RUSH, A. J. & KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Am. Statist. Assoc.* **107**, 1106–18.
- ZHAO, Y., ZENG, D., SOCINSKI, M. A. & KOSOROK, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* **67**, 1422–33.
- ZHAO, Y.-Q., ZENG, D., LABER, E. B., SONG, R., YUAN, M. & KOSOROK, M. R. (2015). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika* **102**, 151–68.

[Received on 14 February 2017. Editorial decision on 7 February 2018]