The Annals of Applied Statistics
2019, Vol. 13, No. 3, 1847–1883
https://doi.org/10.1214/19-AOAS1261

OBLIQUE RANDOM SURVIVAL FORESTS[1]

BY BYRON C. JAEGER*, D. LEANN LONG*, DUSTIN M. LONG*, MARIO SIMS†, JEFF M. SZYCHOWSKI*, YUAN-I MIN†, LESLIE A. MCCLURE‡, GEORGE HOWARD* AND NOAH SIMON§

University of Alabama at Birmingham, University of Mississippi Medical Center†, Dornsife School of Public Health Drexel University ‡ and University of Washington§*

We introduce and evaluate the oblique random survival forest (ORSF). The ORSF is an ensemble method for right-censored survival data that uses linear combinations of input variables to recursively partition a set of training data. Regularized Cox proportional hazard models are used to identify linear combinations of input variables in each recursive partitioning step. Benchmark results using simulated and real data indicate that the ORSF's predicted risk function has high prognostic value in comparison to random survival forests, conditional inference forests, regression and boosting. In an application to data from the Jackson Heart Study, we demonstrate variable and partial dependence using the ORSF and highlight characteristics of its ten-year predicted risk function for atherosclerotic cardiovascular disease events (ASCVD; stroke, coronary heart disease). We present visualizations comparing variable and partial effect estimation according to the ORSF, the conditional inference forest, and the Pooled Cohort Risk equations. The `obliqueRSF` R package, which provides functions to fit the ORSF and create variable and partial dependence plots, is available on the comprehensive R archive network (CRAN).

1. Introduction. Since Breiman (2001) introduced the random forest (RF), it has become recognized as a flexible and accurate tool for classification, regression, and right-censored time-to-event (i.e., survival) analysis (Strobl, Malley and Tutz (2009)). RFs are characterized by an ensemble of decision trees, each of which

Received November 2018; revised April 2019.

[1]The REasons for Geographic and Racial Differences in Stroke (REGARDS) study is supported by a cooperative agreement (U01 NS041588) from the National Institute of Neurological Disorders and Stroke, National Institutes of Health, Department of Health and Human Service. Additional support was provided by grant R01 HL080477 from the National Heart, Lung and Blood Institute. The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I/HHSN26800001) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD).

Key words and phrases. Random forest, survival, machine learning, penalized regression, cardiovascular disease.

are grown through recursive partioning (Breiman (1984)). Recursive partitioning is performed in steps, where each step involves splitting a set of data into two descendant subsets, beginning with the full set. The data are split based on an input variable X and a splitting value c such that one descendant set contains observations in the current data with $X < c$ and the other set contains observations in the current data with $X \geq c$. The splitting variable and value are chosen to maximize the difference between the descendant sets in an outcome variable, and the splitting variable can be either a single input variable or a linear combination of input variables (LCIVs). Descendant data sets that are split into smaller sets are referred to as "nonterminal nodes" in the tree. Descendant data sets that are too small to split further (i.e., data sets with less than a prespecified minimum number of unique observations or events) are not split any further and are referred to as "terminal nodes" (i.e., leaves) of the tree.[2] Predictions for a single test observation are computed by determining which terminal node the observation falls to and then aggregating the outcomes of participants in the training data who were mapped to the same node.

Breiman (2001) showed that RFs often achieve lower generalization error when LCIVs are used instead of a single variable (i.e., oblique splitting versus axis-based splitting) to split the training data in each recursive partitioning step. Hothorn and Lausen (2003) incorporated LCIVs by applying linear discriminant analysis to out-of-bag data and using predicted values from the discriminant model for in-bag data as a candidate splitting variable. Menze et al. (2011) introduced oblique RFs, and showed that Breiman's method of constructing LCIVs could be improved by fitting regularized regression models (Hastie, Tibshirani and Friedman (2001), Section 3.4) to the data in nonterminal nodes and synthesizing a splitting variable using the model's predictions. Zhu, Zeng and Kosorok (2015) proposed to construct LCIVs using the most important K variables, which they identified by fitting ensembles of extremely randomized trees (Geurts, Ernst and Wehenkel (2006)) to the data in each nonterminal node. For classification and regression problems, RFs with LCIVs have achieved excellent generalization error (Breiman (2001), Section 5) for several data sets in public repositories (Dheeru and Karra Taniskidou (2017)).

Random survival forests (RSFs) (Ishwaran et al. (2008)) and conditional inference forests (CIFs) (Hothorn, Hornik and Zeileis (2006)) extend Breiman's RF to right-censored survival data, but neither of these recursive partitioning algorithms incorporate LCIVs. Additionally, to our knowledge, there are no studies examining the use of LCIVs for recursive partitioning in the context of right-censored survival outcomes (Bou-Hamad, Larocque and Ben-Ameur (2011)). Therefore, we developed and evaluated a method to incorporate LCIVs in binary decision trees

[2]The term "node" refers to a descendant data set in the decision tree. The term "splitting" a node refers to partitioning the descendant data into two nonoverlapping subsets.

for right-censored survival data. Zhu (2013) found that extremely randomized survival trees did not produce optimal multiplicative coefficients for LCIVs. Thus, we extended the approach of Menze et al. (2011) by using regularized (Zou and Hastie (2005)) Cox proportional hazards (PH) (Cox (1992), Simon et al. (2011)) models to identify multiplicative coefficients for LCIVs in each recursive partitioning step. Following the notation of Menze et al. (2011), we refer to our method as the oblique RSF (ORSF).

The purpose of this article is to describe the ORSF and assess the prognostic value of its predicted risk fucntion. After a brief summary of RSFs and CIFs in Section 2, ORSFs are described in Section 3. Simulated and real data are used in Sections 4 and 5, respectively, to assess performance of the ORSF in comparison to the RSF, CIF, gradient boosted decision trees and Cox PH models. Performance and mean time required for computation over the entire collection of data from Sections 4 and 5 is assessed in Section 6. In Section 7, we apply the `obliqueRSF` R package to conduct exploratory analyses using data from participants in the Jackson Heart Study (JHS). Our analysis of the JHS data focuses on atherosclerotic cardiovascular disease (ASCVD) and its dependence on four key variables: age, systolic blood pressure, estimated glomerular filtration rate and left ventricular mass. We conclude with a summary and discussion of our results in Section 8.

2. Ensemble tree methods for right-censored survival data. Here we summarize two separate implementations of ensemble tree methods for right-censored survival data:

1. Random survival forest (RSF) (Ishwaran et al. (2008)).
2. Conditional inference forest (CIF) (Hothorn, Hornik and Zeileis (2006)).

We present an ad-hoc summary of the steps taken to grow ensembles of binary decision trees for right-censored survival data (i.e., survival trees) in Section 2.1. Section 2.2 contains notation to describe the framework of survival ensembles. Last, we present summaries of the RSF and CIF in Sections 2.3 and 2.4, respectively.

2.1. *How to grow ensembles of survival trees.*

Step 1 Draw B subsamples[3] from the data: $\{D_1, \ldots, D_B\}$
Step 2 For $b = 1, \ldots, B$, grow a survival tree using D_b:

(a) Initiate $\mathcal{N}_G = \{D_b\}$, the set of "nodes to grow."
(b) If $\mathcal{N}_G = \varnothing$, go to (c). Otherwise, for each node in $\mathcal{N}_G$, do the following:

[3]The RSF applies bootstrap sampling with replacement, whereas the CIF uses sampling without replacement to achieve unbiasedness.

(i) Define the set of candidate splitting variables for the current node by selecting a random subset of the available predictor variables.[4]

(ii) [5]If there is no evidence of statistical association between the survival outcome and any of the candidate splitting variables, then (1) remove the current node from $\mathcal{N}_G$, (2) label the current node as "terminal," and (3) skip steps (iii) and (iv) below.

(iii) Split the current node into descendant nodes,[6] A_1 and A_2, using the candidate splitting variable that maximizes a log-rank statistic comparing survival outcomes between A_1 and A_2.

(iv) Remove the current node from $\mathcal{N}_G$. If A_1 has at least a minimum number of unique observations (i.e., `minsplit` or `nodesize`), add A_1 to $\mathcal{N}_G$. Otherwise, label A_1 as "terminal." Do the same for A_2.

(c) Define a predicted survival or cumulative hazard function for each terminal node based on the observed survival times in the node.

Step 3 Aggregate predicted survival[7] or cumulative hazard functions from the B survival trees to compute ensemble predictions.

2.2. *Notation*. Consider right-censored survival data from N participants: $(y_1, \boldsymbol{x}_1, \delta_1), \ldots, (y_N, \boldsymbol{x}_N, \delta_N)$. For participant i, $1 \le i \le N$, y_i represents survival time if $\delta_i = 1$ and time to censoring if $\delta_i = 0$, $\boldsymbol{x}_i$ is a length p vector of predictors: $(x_{i,1}, \ldots, x_{i,p})$. Denote the jth column of the matrix $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]'$ as $\boldsymbol{x}_{(j)}$. Let $t_1 < \cdots < t_m$ denote the m unique event times (assume there are no ties). Denote the bth survival tree as $\mathcal{T}_b$, let $\mathcal{T}_b(\boldsymbol{x}_i)$ identify the terminal node in $\mathcal{T}_b$ participant i is mapped to, and let c_{ib} denote the number of times the data from participant i occurs in the bth bootstrap sample. Following the counting process notation of Andersen et al. (2012), define

$$N_i(s) = I(y_i \le s, \delta_i = 1); \qquad Y_i(s) = I(y_i > s), \tag{2.1}$$

where $0 \le s \le t_m$ and $I(\cdot)$ is the indicator function. Define

$$N_b^*(s, \boldsymbol{x}) = \sum_{i=1}^{N} c_{ib} \cdot I\big(\boldsymbol{x}_i \in \mathcal{T}_b(\boldsymbol{x})\big) \cdot N_i(s) \tag{2.2}$$

[4]Software packages generally use the term `mtry` to denote the size of the random subset of predictors.

[5]This step only occurs in the CIF.

[6]For survival trees, descendant nodes must have at least `nmin` unique observations that are not censored.

[7]The CIF applies a nearest neighbor aggregation scheme to aggregate predictions from each tree, whereas the RSF weights all trees equally.

and

$$Y_b^*(s, \boldsymbol{x}) = \sum_{i=1}^{N} c_{ib} \cdot I\big(\boldsymbol{x}_i \in \mathcal{T}_b(\boldsymbol{x})\big) \cdot Y_i(s) \tag{2.3}$$

as the number of uncensored events prior to and number of participants at risk at time s, respectively, in $\mathcal{T}_b(\boldsymbol{x})$.

2.3. *The random survival forest* (*RSF*). The RSF extends the prescription described by Breiman and Cutler (2003) to the context of survival analyses. The log-rank statistic (Segal (1988), Section 3.2) is applied to determine which candidate variable and cut-point should be used. A Nelson–Aalen estimator of the cumulative hazard function is formed in each terminal node based on the survival times of observations in the node:

$$\widehat{H}_b(t \mid \boldsymbol{x}) = \int_0^t \frac{N_b^*(ds, \boldsymbol{x})}{Y_b^*(s, \boldsymbol{x})}. \tag{2.4}$$

In turn, the ensemble survival function is

$$\widehat{S}_{\mathrm{RSF}}(t \mid \boldsymbol{x}) = \exp\left[-\frac{1}{B}\sum_{i=1}^{B} \widehat{H}_b(t \mid \boldsymbol{x})\right]. \tag{2.5}$$

2.4. *The conditional inference forest* (*CIF*). The CIF avoids variable selection bias (Breiman (1984), pg. 42) by applying permutation tests (Strasser and Weber (1999)) to compare candidate splitting variables. During each recursive partitioning step, if there is no evidence of association between any of the candidate variables and the response, the current node is not split and is labeled terminal. For ensemble prediction, the CIF applies a weighted Kaplan–Meier estimate (Hothorn et al. (2004)) based on all training observations in the B leaves containing the new observation:

$$\widehat{S}_{\mathrm{CIF}}(t \mid \boldsymbol{x}) = \prod_{s \le t}\left[1 - \frac{\sum_{b=1}^{B} N_b^*(ds, \boldsymbol{x})}{\sum_{b=1}^{B} Y_b^*(s, \boldsymbol{x})}\right]. \tag{2.6}$$

3. The oblique random survival forest (ORSF). Here we describe the ORSF, beginning with the use of regularized Cox PH models (Section 3.1) and synthesis of candidate splitting variables using regularized Cox PH models (Section 3.2). Additional details related to sampling and prediction are provided in Section 3.3.

3.1. *The regularized Cox PH model*. The ORSF embeds regularized Cox PH models into the nonterminal nodes of its survival trees. Suppose there are K observations in the current nonterminal node. The embedded Cox PH model assumes a semi-parametric form for the hazard:

$$h_k(t) = h_0(t)e^{\boldsymbol{x}_k^T \boldsymbol{\beta}}, \tag{3.1}$$

where $k = 1, \ldots, K$, $h_k(t)$ is the hazard for observation k at time t, $h_0(t)$ is the baseline hazard function, and $\boldsymbol{\beta}$ is a vector of $\texttt{mtry} \leq p$ coefficients, where `mtry` is the number of randomly selected predictor variables for the current node and p is the total number of predictor variable. Estimation of $\boldsymbol{\beta}$ is carried out using the partial likelihood,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{e^{\boldsymbol{x}_{j(i)}^T \boldsymbol{\beta}}}{\sum_{j \in R_i} e^{\boldsymbol{x}_j^T \boldsymbol{\beta}}}, \tag{3.2}$$

where R_i is the set of indices, j, with $y_j \geq t_i$ (i.e., those still at risk at time t_i), and $j(i)$ is the index of the observation for which an event occurred at time t_i. Simon et al. (2011) proposed to maximize (3.2) subject to the *elastic net* penalty (Zou and Hastie (2005)):

$$\alpha \sum_{i=1}^{p} |\beta_i| + (1 - \alpha) \sum_{i=1}^{p} \beta_i^2 \leq s, \tag{3.3}$$

where $s \geq 0$ is a constant that controls shrinkage and has one-to-one correspondence with the complexity parameter, $\lambda \geq 0$. Setting $\alpha = 1$ and $\alpha = 0$ in (3.3) results in the classic "Ridge" and "Lasso" penalties, respectively (Hastie, Tibshirani and Friedman (2001), Section 3.4). Each pair of values for α and λ corresponds to a different solution for $\boldsymbol{\beta}$. Notably, with a ridge penalty, every node will develop a solution for $\boldsymbol{\beta}$ with nonzero regression coefficients for all `mtry` randomly selected predictor variables. On the other hand, a lasso penalty will usually identify sparse solutions that set most of the `mtry` coefficients in $\boldsymbol{\beta}$ to zero.

3.2. *Linear combinations of input variables* (*LCIVs*). LCIVs are synthesized using embedded Cox PH models and used as splitting variables for internal nodes in the ORSF. The LCIVs are synthesized using $\boldsymbol{x}^T \widehat{\boldsymbol{\beta}}$, where $\boldsymbol{x}^T = \{\boldsymbol{x}_1^T, \ldots \boldsymbol{x}_{\boldsymbol{K}}{}^T\}$ is the stacked matrix of input vectors and $\widehat{\boldsymbol{\beta}}$ is selected from a number of candidate solutions given by the embedded Cox PH model. The maximum number of predictor variables with nonzero regression coefficients in these candidate solutions is `mtry`. The process that identifies candidate solutions for $\widehat{\boldsymbol{\beta}}$ is governed by (1) whether cross-validation is applied, and (2) the value of α.

If cross-validation is applied (`use.cv = TRUE` in the `ORSF` function), then 5-fold cross-validation will be used in each nonterminal node to identify the following λ values:

λ_{CV}: The value of λ maximizing cross-validated partial likelihood (van Houwelingen et al. (2006)).

λ_{SE}: The highest value of λ within one standard error of λ_{CV}.

The choice of α affects the candidate solutions identified by λ_{CV} and λ_{SE}. For example, when $\alpha = 1$, both λ_{CV} and λ_{SE} identify solutions that use *all* of the

randomly selected candidate splitting variables; however, the solution identified by λ_{SE} will impose a stronger penalty on regression coefficients. If α is set at a value close but not equal to 1, for example, 0.90, then λ_{CV} and λ_{SE} will both identify "ridge-like" solutions that estimate nonzero regression coefficients for *nearly all* of the candidate splitting variables.

If the analyst chooses not to use cross-validation, a regularization path is fitted to the current node's data to identify $\lambda_1, \ldots, \lambda_{\texttt{mtry}}$, where λ_i is the maximum value of lambda such that the model has i effective degrees of freedom (Efron et al. (2004)). To ensure these values of λ can be found, we require $\alpha < 1$ when cross-validation is not used. In this context, the choice of α affects how quickly the regularization path transitions from a minimally to maximally complex solution. For example, when α is close to 0, the regularization path will be "lasso-like" and identify more solutions with a small number of nonzero regression coefficients compared to a "ridge-like" regularization path, where all or nearly all solutions have nonzero regression coefficients for all candidate variables.

Regardless of whether or not cross-validation is used, each candidate solution for $\widehat{\boldsymbol{\beta}}$ is evaluated as follows:

1. $\widehat{\boldsymbol{\eta}} = \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}$ is computed for each observation in the current node.
2. `nsplit` candidate cut-points, $c_1 \ldots, c_{\texttt{nsplit}}$ are selected at random from the unique values of $\widehat{\boldsymbol{\eta}}$.
3. For $i = 1, \ldots, \texttt{nsplit}$, a log-rank statistic comparing survival curves between observations with $\boldsymbol{x}_k^T \widehat{\boldsymbol{\beta}} \leq c_i$ and observations with $\boldsymbol{x}_k^T \widehat{\boldsymbol{\beta}} > c_i$ is computed.

When cross-validation is not applied, log-rank statistics are penalized by a scaling factor of $(1 + \gamma \cdot \text{df})$, where $\gamma > 0$ is a tuning parameter and df is the effective degrees of freedom of the model that generated the candidate LCIV. Larger values of γ will result in selection of less complex LCIVs with fewer variables. Setting $\gamma = 0$ imposes no penalty on the complexity of LCIVs and will result in selection of LCIVs with a larger number of variables. If the maximal log-rank statistic does not exceed a prespecified threshold, early stopping is applied (i.e., the node is not split and is labeled terminal). Otherwise, the cut-point and candidate solution maximizing the log-rank statistic are used to split the node.

3.3. *Additional details*. The ORSF applies subsampling rather than bootstrap sampling with replacement. Ensemble predictions from an ORSF are formed using the same weighted aggregation scheme as in the CIF.

4. Simulation study. Here we describe and summarize results from a simulation study following the protocol described by Morris, White and Crowther (2019). The primary aim of the simulation is to compare the prognostic value of the ORSF's predicted risk function with a variety of competing learning algorithms in three general settings. Data generating mechanisms are summarized in

Section 4.1. The primary estimands of the simulation study are described in Section 4.3. Competing learning algorithms and their corresponding hyper-parameter settings are introduced in Section 4.2. Tabular summaries of results are presented in Section 4.5. Computational strategies are reported in Section 4.4. Source codes for the current simulation study are available from the first author's GitHub site and Supplementary Material Jaeger et al. (2019).

4.1. *Data generating mechanisms*. We generated right-censored survival data from three scenarios using a training sample of $N = 500$ and a testing sample of $M = 1000$. The number of predictor variables used to generate survival outcomes, p, was set at 25 and 50 for each scenario. For $i = 1, \ldots, N + M$, $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$ was generated from a zero mean, unit variance multivariate normal distribution with $\text{cov}(x_{i,u}, x_{i,v}) = (1/3)^{|u-v|}$. Multiplicative coefficient vectors (i.e., $\boldsymbol{\beta}$) for variables 1 through p were generated as a uniform sequence of length p, beginning from -1 and ending at $+1$. The three scenarios were as follows:

A: The length p vector $\boldsymbol{\beta}$ contained one effect for each predictor. Simulated outcomes followed a Weibull distribution.

B: The length $2p$ vector $\boldsymbol{\beta}$ contained one main effect for each predictor and, additionally, each predictor was used in two bi-variate interaction terms. The multiplicative coefficients for interaction variables were generated as a uniform sequence of length p, beginning from -1 and ending at $+1$.[8] Simulated outcomes followed the same distribution as in scenario A.

C: The same $\boldsymbol{\beta}$ vector was used as described scenario B, and each (y_i, δ_i) followed one of three distinct Weibull distributions, depending on the value of $\boldsymbol{x}_i\boldsymbol{\beta}$.

Scenario A is designed to favor the PH model, whereas scenario B will favor survival tree ensembles. Scenario C is of particular interest as the proportional hazards assumption does not hold.

4.2. *Competing methods and tuning parameters*.

4.2.1. *Decision tree ensembles*. ORSF, RSF and CIF ensembles were composed of 1000 survival trees. The minimum number of unique observations needed to split a node (i.e., `minsplit` or `nodesize`) was 10. For the ORSF and RSF, the minimum number of events per terminal node was 1, splitting values were selected by comparing 25 randomly selected cut-points, and the number of variables tried at each node (i.e., `mtry`) was set to the smallest integer $\geq \sqrt{p}$. The CIF

[8]For example, consider $p = 4$. Variable 1 interacts with variable 2 (effect size: -1), variable 2 interacts with variable 3 (effect size: $-1/3$), variable 3 interacts with variable 4 (effect size: $1/3$), and variable 4 interacts with variable 1 (effect size: 1) giving four main effects and four interaction effects, that is, $2p$ terms in $\boldsymbol{\beta}$.

used tuning parameters consistent with an unbiased recursive partitioning framework (Hothorn et al. (2019), Strobl et al. (2007)). Two ORSF models were fitted as follows:

ORSF: The ORSF algorithm is applied with $\alpha = 0.50$ and without using cross-validation to synthesize candidate splitting variables (i.e., setting `use.cv = FALSE` as described in Section 3.2).

ORSF_{CV}: The ORSF algorithm is applied with $\alpha = 0.05$ (a ridge-like penalty) and using cross-validation to synthesize candidate splitting variables (i.e., setting `use.cv = TRUE` as described in Section 3.2).

Gradient boosted decision trees were fitted using cross-validation to determine (1) the optimal number of boosting steps and (2) an optimal set of hyper-parameter values (Chen and Guestrin (2016), Friedman (2001)). For (2), we generated 10 sets of hyper-parameter values randomly (i.e., values for tree depth, column subsampling, row subsampling, minimum child weight and minimum loss reduction required to make a further partition on a leaf node). The hyper-parameter set that maximized cross-validated partial log-likelihood was used to fit an ensemble of gradient boosted decision trees to the training data.

4.2.2. *Cox proportional hazards (PH) models.* Regularized Cox PH models were fitted using Lasso-like ($\alpha = 0.90$) and Ridge-like ($\alpha = 0.10$) penalties, and using the value of λ that maximized cross-validated partial likelihood (Friedman, Hastie and Tibshirani (2010)). Unregularized Cox PH models were fitted using stepwise variable selection based on Akaike's information criteria (Burnham and Anderson (2004)). Boosted Cox PH models were fitted using cross-validation to determine the optimal number of boosting steps and penalized score statistics to determine parameter updates in each boosting step (Tutz and Binder (2007)).

4.3. *Prediction error.* Predicted risk for right-censored survival outcomes can be assessed using the time dependent concordance index and Brier score (Gerds et al. (2013), Graf et al. (1999)). For $t > 0$, the inverse probability of censoring weighted concordance index and Brier score are estimated using

$$\widehat{\text{C}}(t) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} I(t_i < t_j) \cdot I(\widehat{S}(t \mid \boldsymbol{x}_i) > \widehat{S}(t \mid \boldsymbol{x}_j)) \cdot I(t_i < t) \cdot \delta_i \cdot \widehat{W}_{ij}^{-1}}{\sum_{i=1}^{M}\sum_{j=1}^{M} I(t_i < t_j) \cdot I(t_i < t) \cdot \delta_i \cdot \widehat{W}_{ij}^{-1}}$$

and

$$\widehat{\text{BS}}(t) = \frac{1}{M}\sum_{i=1}^{M}\{\widehat{S}(t \mid \boldsymbol{x}_i)^2 \cdot I(y_i < t, \delta_i = 1) \cdot \widehat{G}(y_i \mid \boldsymbol{x}_i)^{-1} + [1 - \widehat{S}(t \mid \boldsymbol{x}_i)]^2 \cdot I(y_i > t) \cdot \widehat{G}(y_i \mid \boldsymbol{x}_i)^{-1}\}$$

respectively, where for participant i in the testing data, $\widehat{S}(t \mid \boldsymbol{x}_i)$ is the estimated probability of survival at time t, $\boldsymbol{x}_i$ is the input information, and $\widehat{G}(y_i \mid \boldsymbol{x}_i)$ is the estimated probability of censoring. Additionally, $\widehat{W}_{ij} = \widehat{G}(y_i \mid \boldsymbol{x}_j) \cdot \widehat{G}(y_i - \mid \boldsymbol{x}_i)$ defines the probability of censoring weights for concordance.

The concordance index measures the probability at time t that a randomly selected participant who has experienced an event has a higher model-based prediction than a randomly selected subject who has not experienced an event. Concordance indices of 1.00 and 0.50 correspond to perfect and worthless discrimination, respectively. Gerds et al. (2013) demonstrated asymptotic bias of the concordance index introduced by Harrell et al. (1982) and proposed a method to estimate a time-dependent concordance index for models with covariate dependent censoring (i.e., $\widehat{\mathrm{C}}(t)$). Additionally, Blanche, Kattan and Gerds (2019) showed that the time dependent area under the receiver operating characteristic curve rather than the concordance index introduced by Harrell et al. (1982) should be used to assess t-year predicted risk.

For a single observation at time t, the Brier score is the squared difference between observed survival status (e.g., 1 = alive at time t and 0 = dead at time t) and a model-based prediction of survival at time t. The expected Brier score of a prediction model which ignores all predictor variables (i.e., the Kaplan–Meier estimate of survival calculated with all training samples) is a reference value that can be used as a benchmark Brier score for prediction models. For clarity and ease of interpretation, we tabulate both the unscaled and scaled Brier score estimates, where the scaled Brier score is computed as 1 minus the ratio of the unscaled Brier score to the reference Brier score. Similar to the R^2 statistic, a scaled Brier score of 1 and 0 indicate a perfect and worthless model, respectively.

4.3.1. *Summary measures of prediction error.* The concordance index proposed by Gerds et al. (2013) uses survival time as response and differs conceptually from the the time-dependent receiver-operator characteristic curve proposed by Heagerty, Lumley and Pepe (2000). The latter incident statistic measures a prediction model's ability to classify survival status at a given time point, while the former cumulative statistic measures a prediction model's ability to order the survival times. Heagerty and Zheng (2005) have established a formal relationship between these two measures. Following the precedent of Ishwaran et al. (2008), we define concordance error as $\widehat{\mathrm{C}}_e(t) = 1 - \widehat{\mathrm{C}}(t)$.

As $\widehat{\mathrm{BS}}(t)$ is time dependent, integration from baseline to a specified follow-up time provides a summary measure of performance that can be tabulated for direct comparisons between competing learning methods. The integrated BS is defined as

$$\widehat{\mathcal{BS}}(T) = \frac{1}{T} \int_0^T \widehat{\mathrm{BS}}(t)\, dt, \tag{4.1}$$

where T is the specified follow-up time. In our results, we set T equal to the median event time in the testing data. We also present values of $\widehat{\mathcal{BS}}(T)$ and $\widehat{\mathrm{C}}_e(t)$

that are scaled by a factor of 100 to avoid an unnecessary amount of leading zeros (i.e., we present 0.5 instead of 0.005).

4.4. *Computational notes and software.* All analyses were performed in R version 3.5.0. Right-censored survival outcomes were generated using the `simsurv` R package (Brilleman (2018)). We used the `RandomForestSRC` (Ishwaran and Kogalur (2019)), `party` (Hothorn et al. (2019)), and `obliqueRSF` (Jaeger (2018)) R packages to fit RSFs, CIFs, and ORSFs, respectively, the `xgboost` (Chen et al. (2019)) package to fit gradient boosted decision trees, the `glmnet` (Friedman, Hastie and Tibshirani (2010)) package to fit regularized Cox PH models, the `survival` (Therneau (2015)) package to fit classical Cox PH models, the `MASS` (Venables and Ripley (2002)) package to perform forward stepwise selection using Akaike's information criteria, and the `CoxBoost` package to fit boosted Cox PH models (Binder (2013)). To compute $\widehat{\mathcal{BS}}(T)$ and $\widehat{\mathrm{C}}(t)$, we used the `pec` package (Mogensen, Ishwaran and Gerds (2012)). Unadjusted Kaplan–Meier estimates were used to estimate inverse probability of censoring weights throughout (Mogensen, Ishwaran and Gerds (2012), Section 6.2).

4.5. *Results.* The mean rankings according to $\widehat{\mathrm{C}}_e(t)$ for $\mathrm{ORSF}_{\mathrm{CV}}$ and ORSF were 2.17 and 2.67, respectively, out of the nine learning algorithms we applied (Table 1). As expected, the ORSF and $\mathrm{ORSF}_{\mathrm{CV}}$ provided the lowest values of $\widehat{\mathrm{C}}_e(t)$ in scenario B. Notably, the ORSF and $\mathrm{ORSF}_{\mathrm{CV}}$ also provided the lowest values of $\widehat{\mathrm{C}}_e(t)$ in scenario C, despite the invalidity of the PH assumption. In comparison to the RSF and CIF, both the ORSF and $\mathrm{ORSF}_{\mathrm{CV}}$ provided lower values of $\widehat{\mathrm{C}}_e(t)$ in each of the six simulated analyses. The absolute (percent) reduction in the mean value of $\widehat{\mathrm{C}}_e(t)$ from using the $\mathrm{ORSF}_{\mathrm{CV}}$ instead of the RSF and using the $\mathrm{ORSF}_{\mathrm{CV}}$ instead of the CIF was 3.13 (9.62%) and 0.52 (1.58%), respectively.

The mean rankings according to $\widehat{\mathcal{BS}}(T)$ for $\mathrm{ORSF}_{\mathrm{CV}}$ and ORSF were 4.67 and 4.17, respectively, out of the nine learning algorithms we applied (Table 2). In comparison to the RSF and CIF, both the ORSF and $\mathrm{ORSF}_{\mathrm{CV}}$ recorded lower mean values of $\widehat{\mathcal{BS}}(T)$ in each of the six simulated analyses. The absolute (percent) increase in the mean *scaled value* (see Section 4.3) of $\widehat{\mathcal{BS}}(T)$ using the $\mathrm{ORSF}_{\mathrm{CV}}$ instead of the RSF and using the $\mathrm{ORSF}_{\mathrm{CV}}$ instead of the CIF were 1.57 (42.39%) and 0.45 (9.37%), respectively.

5. Application to real data. Here we describe and summarize results from a resampling experiment using data from six independent studies. We summarize each study, separately, in Section 5.1. We describe tuning parameters and the resampling procedure we applied in Sections 5.2 and 5.3, respectively. We present and summarize results in Section 5.5.

TABLE 1

Mean concordance errors for competing learning methods, aggregated over 100 *simulations in three scenarios. The minimum concordance error for each simulated analysis is written in bold text*

		Ensemble Survival Trees[§]					Proportional Hazards			
Scenario[†]	p[‡]	ORSF	ORSF_{CV}	CIF	RSF	Xgboost	CoxBoost	Lasso	Ridge	Step
					Concordance error: $100 \cdot \widehat{\text{C}}_e(t)$					
A	25	30.79	30.69	31.45	33.90	31.32	30.35	31.31	**30.13**	30.68
B	25	31.32	31.14	31.92	35.69	32.41	31.65	31.32	**30.90**	32.10
C	25	29.88	**29.86**	30.49	32.52	30.64	31.87	31.87	31.63	32.35
A	50	31.00	**30.97**	31.33	34.06	31.60	33.06	32.90	32.43	33.64
B	50	**35.13**	35.20	35.71	37.75	36.02	38.35	38.39	38.01	38.98
C	50	**37.36**	37.39	37.44	40.11	38.15	39.71	39.80	39.14	40.49
					Monte Carlo Standard Error of $100 \cdot \widehat{\text{C}}_e(t)$					
A	25	0.103	0.106	0.107	0.103	0.106	0.116	0.118	0.119	0.111
B	25	0.098	0.102	0.104	0.093	0.104	0.115	0.118	0.114	0.108
C	25	0.057	0.063	0.068	0.072	0.097	0.081	0.077	0.081	0.080
A	50	0.099	0.087	0.082	0.090	0.100	0.091	0.092	0.096	0.088
B	50	0.085	0.088	0.081	0.061	0.098	0.093	0.093	0.096	0.088
C	50	0.069	0.070	0.063	0.095	0.076	0.101	0.096	0.086	0.104
					Percent increase in $\widehat{\text{C}}_e(t)$*, relative to minimum (0.00)*					
A	25	2.2	1.9	4.4	12.5	4.0	0.8	0.6	**0.0**	1.8
B	25	1.4	0.8	3.3	15.5	4.9	2.4	1.4	**0.0**	3.9
C	25	0.0	**0.0**	2.1	8.9	2.6	6.7	6.7	5.9	8.3
A	50	0.1	**0.0**	1.2	10.0	2.0	6.7	6.2	4.7	8.6
B	50	**0.0**	0.2	1.7	7.5	2.5	9.2	9.3	8.2	11.0
C	50	**0.0**	0.1	0.2	7.4	2.1	6.3	6.5	4.8	8.4
					Rankings based on $\widehat{\text{C}}_e(t)$					
A	25	6	5	8	9	7	3	2	**1**	4
B	25	4	2	6	9	8	5	3	**1**	7
C	25	2	**1**	3	9	4	6	7	5	8
A	50	2	**1**	3	9	4	7	6	5	8
B	50	**1**	2	3	5	4	7	8	6	9
C	50	**1**	2	3	8	4	6	7	5	9
		2.67	**2.17**	4.33	8.17	5.17	5.67	5.50	3.83	7.50

[†] Descriptions of Scenarios A, B and C are provided in Section 4.1.

[‡] p represents the number of predictor variables in each simulated data set.

[§] RSF = random survival forest; CIF = conditional inference forest; ORSF = oblique random survival forest; ORSF_{CV} = oblique random survival forest with internal cross-validation (see Section 3.2).

[‡] Apparent ties in concordance errors are a result of rounding errors.

TABLE 2

Mean integrated Brier scores for competing learning methods, aggregated over 100 *simulations in three scenarios. The minimum Brier score for each simulated analysis is written in bold text*

		Ensemble Survival Trees[§]					Proportional Hazards				
Scenario[†]	p[‡]	ORSF	ORSF_{CV}	CIF	RSF	Xgboost	CoxBoost	Lasso	Ridge	Step	Reference*
					Integrated Brier score: $100 \cdot \widehat{\mathcal{BS}}(T)$						
A	25	8.89	8.89	8.93	9.07	8.82	8.73	8.73	**8.71**	8.78	9.48
B	25	8.93	8.93	9.02	9.17	8.87	8.81	8.78	**8.74**	8.93	9.43
C	25	8.91	8.92	8.96	9.00	**8.78**	8.93	8.93	8.94	8.95	9.67
A	50	9.15	9.16	9.24	9.27	**8.99**	9.13	9.11	9.09	9.22	9.73
B	50	11.15	11.16	11.16	11.31	**11.14**	11.32	11.33	11.31	11.37	11.61
C	50	11.38	11.39	11.40	11.58	**11.38**	11.47	11.48	11.45	11.61	11.68
					Monte Carlo Standard Error of $100 \cdot \widehat{\mathcal{BS}}(T)$						
A	25	0.014	0.014	0.015	0.012	0.012	0.011	0.010	0.010	0.009	0.016
B	25	0.008	0.008	0.008	0.007	0.008	0.008	0.008	0.008	0.009	0.010
C	25	0.012	0.012	0.012	0.011	0.013	0.013	0.012	0.012	0.013	0.014
A	50	0.019	0.019	0.020	0.018	0.018	0.019	0.019	0.019	0.018	0.021
B	50	0.017	0.017	0.017	0.016	0.016	0.017	0.018	0.018	0.018	0.018
C	50	0.016	0.016	0.016	0.014	0.014	0.017	0.017	0.016	0.016	0.017
					Scaled $\widehat{\mathcal{BS}}(T)$ *values*: $100 \cdot [1 - \widehat{\mathcal{BS}}(T)/\text{Reference}]$						
A	25	6.26	6.28	5.80	4.37	6.93	7.88	7.91	**8.07**	7.35	0.00
B	25	5.31	5.35	4.42	2.84	5.97	6.61	6.94	**7.33**	5.38	0.00
C	25	7.91	7.81	7.35	6.92	**9.24**	7.64	7.65	7.61	7.45	0.00
A	50	5.96	5.85	5.05	4.71	**7.62**	6.17	6.40	6.62	5.30	0.00
B	50	3.95	3.85	3.85	2.54	**4.03**	2.45	2.42	2.53	2.04	0.00
C	50	2.60	2.50	2.46	0.84	**2.61**	1.78	1.74	1.96	0.65	0.00
					Rankings based on $\widehat{\mathcal{BS}}(T)$						
A	25	7	6	8	9	5	3	2	**1**	4	10
B	25	7	6	8	9	4	3	2	**1**	5	10
C	25	2	3	8	9	**1**	5	4	6	7	10
A	50	5	6	8	9	**1**	4	3	2	7	10
B	50	2	4	3	5	**1**	7	8	6	9	10
C	50	2	3	4	8	**1**	6	7	5	9	10
		4.17	4.67	6.50	8.17	**2.17**	4.67	4.33	3.50	6.83	10.00

[†] Descriptions of Scenarios A, B and C are provided in Section 4.1.

[‡] p represents the number of predictor variables in each simulated data set.

[§] RSF = random survival forest; CIF = conditional inference forest; ORSF = oblique random survival forest; ORSF_{CV} = oblique random survival forest with internal cross-validation (see Section 3.2).

* The reference values are expected Brier scores of a prediction model that ignores all predictor variables, that is, the Kaplan–Meier estimate calculated using the training data.

[‡] Apparent ties in Brier scores are a result of rounding errors.

TABLE 3
Data set summary

Data	Event Type	Follow-up times[§] 25%	50%	75%	% Censored	N	M	p
A25	Simulated	3.9	10.0	10.0	55.4	500	1000	25
B25	Simulated	3.9	10.0	10.0	57.5	500	1000	25
C25	Simulated	2.8	10.0	10.0	50.6	500	1000	25
A50	Simulated	3.8	10.0	10.0	55.2	500	1000	50
B50	Simulated	3.5	10.0	10.0	56.2	500	1000	50
C50	Simulated	2.6	10.0	10.0	49.5	500	1000	50
PBC	Death	3.0	4.7	7.2	61.5	209	209	17
GBSG2	Death or relapse	1.6	3.0	4.6	56.4	342	342	8
GEBC	Death or relapse	1.8	2.8	4.1	78.2	307	307	1690
MBC	Death or relapse	2.3	5.3	8.8	56.4	39	39	4707
JHS1	CHD or stroke	10.8	11.7	12.6	90.8	915	914	58
JHS2	Heart Failure	9.2	10.0	10.0	93.8	904	904	58
JHS3	Death	13.0	13.9	14.7	81.3	944	943	58
REGARDS1	CHD or stroke	5.2	8.1	9.9	92.2	4485	4485	67
REGARDS2	Heart Failure	5.0	8.1	9.9	97.3	4485	4484	67
REGARDS3	Death	6.0	8.4	10.0	87.2	4485	4485	67

[§] All follow-up times are in years.

5.1. *Data sets*. For each of the data sets described in the following sections, study participants who were lost to follow-up or died from causes unrelated to the primary event(s) were censored at time of last contact or time of death, respectively. Characteristics (e.g., number of participants, number of predictors, percent censored) of each data set are tabulated and presented alongside the characteristics of simulated datasets in Table 3.

5.1.1. *Primary biliary cirrhosis* (*PBC*) *data*. The PBC data and their description are taken from Appendix D of Fleming and Harrington (2011). PBC of the liver is a rare and fatal disease of unknown cause. These data were collected for the Mayo Clinic trial in PBC of the liver conducted between January 1974 and May 1984 comparing the drug D-penicillamine with a placebo.

5.1.2. *German breast cancer study group* (*GBSG*2) *data*. The GBSG2 data and their description are taken from Schumacher et al. (1994). In 1984, the GBSG2 started a multicenter randomized clinical trial to compare the effectiveness of three versus six cycles of 500 mg/m^2 cyclophosphamide, 40 mg/m^2 methotrexate, and 600 mg/m^2 fluorouracil on day 1 and 8 starting perioperatively with or without tamoxifen (3×10 mg/d for two years).

5.1.3. *Gene expression breast cancer* (*GEBC*) *data*. The GEBC data and their description are taken from Gene Expression Omnibus database and the analysis by Ternès et al. (2017), respectively, to identify treatment-effect modifiers in 614 breast cancer patients (Desmedt et al. (2011), Hatzis et al. (2011)) receiving anthracycline-based adjuvant chemotherapy with ($n = 507$) or without ($n = 107$) taxane. For the current analysis, we use preprocessed data from Ternès et al. (2017), who applied frozen robust multiarray (McCall, Bolstad and Irizarry (2010)) and cross-platform normalization (Shabalin et al. (2008)), and standardized the remaining 1689 genes.

5.1.4. *Microarray breast cancer* (*MBC*) *data*. The MBC data and their description are taken from (Van't Veer et al. (2002)). Gene expression profiling was conducted on 78 sporadic lymph-node-negative patients to search for a prognostic signature in their gene expression profiles. 44 patients remained free of breast cancer after their initial diagnosis for an interval of at least five years (good prognosis group, mean follow-up of 8.7 years), and 34 patients developed distant metastases within five years (poor prognosis group, mean time to metastases 2.5 years).

5.1.5. *Jackson heart study* (*JHS*) *data*. The JHS is a population-based prospective cohort study designed to examine the etiology of cardiovascular disease (CVD) (i.e., stroke, coronary heart disease, heart failure) and related risk factors among African Americans (Taylor Jr et al. (2005)). In brief, 5,306 noninstitutionalized African-American participants aged ≥ 20 years were recruited from the Jackson, Mississippi, metropolitan area between 2000 and 2004. For the current analyses, we incorporated data from JHS participants who provided complete records of age, sex, anthropometric measures, alcohol/smoking/dietary/exercise habits, medication use, zip code, blood pressure, diabetes, cholesterol, high and low density lipo-proteins, triglycerides, electrocardiograms, estimated glomerular filtration rate (eGFR) (Levey et al. (2009)), insurance, and history of CVD at baseline. We conducted three separate analyses using these data; JHS1 considered composite events of stroke or coronary heart disease, JHS2 considered heart failure events and JHS3 considered all-cause mortality events.

5.1.6. *The REasons for geographic and racial differences in stroke* (*REGARDS*) *study.* The REGARDS study was designed to investigate reasons underlying the higher rate of stroke mortality among blacks compared with whites, and among residents of the Southeastern US compared with other US regions (Howard et al. (2005)). A total of 30,239 adults from the 48 contiguous US states and the District of Columbia were enrolled between January 2003 and October 2007. For the current analysis, we incorporated data from 8,970 participants who provided complete records of age, sex, race, anthropomorphic measures, alcohol/smoking/dietary/exercise habits, quality of life, self-reported history of CVD, echocardiogram, medication use, blood measures, blood pressure, urine

albumin and creatinine, and eGFR. We conducted three separate analyses using these data; REGARDS1 considered composite events of stroke or coronary heart disease (Safford et al. (2012)), REGARDS2 considered heart failure events and REGARDS3 considered all-cause mortality events.

5.2. *Tuning parameters*. We applied the same learning algorithms and corresponding tuning parameter specifications as described in Section 4.2.

5.3. *Resampling*. Results from the current resampling experiment are based on 250 replicates of bootstrap cross-validation (Mogensen, Ishwaran and Gerds (2012), Section 4.2). In each replicate, for each analysis (e.g., PBC, GBSG2, JHS1, etc.), we:

1. Randomly allocated roughly one half of the data to a training set, and used the rest of the data for testing.
2. Trained each competing method using the training data set.
3. Computed $\widehat{\mathcal{BS}}(T)$ and $\widehat{\mathrm{C}}_e(t)$ using each method's predicted survival curves for observations in the testing data.

5.4. *Monte Carlo error relative to forest size*. Using the PBC data, we assessed the Monte Carlo error (seed effect) of the predicted survival curves relative to the number of trees in the ORSF. We first identified three classes of patients using the `protoclust` package Bien and Tibshirani (2019). We identified one prototype patient in each class, and used these three patients as a hold-out set. We developed an ORSF with 10, 100 and 1000 trees, and then predicted survival curves for patients in the hold-out set. We replicated this step 250 times and plotted each survival curve to visualize Monte Carlo error for the three separate forest sizes.

5.5. *Results*. The three methods that achieved the lowest values of $\widehat{\mathrm{C}}_e(t)$ were the $\mathrm{ORSF}_{\mathrm{CV}}$, the ORSF and the CIF (mean ranks of 1.70, 1.80 and 3.20, respectively) (Table 4). Distributions of $\widehat{\mathrm{C}}_e(t)$ were consistent among competing learning methods with the exception of the penalized Cox PH models, which had higher variability of $\widehat{\mathrm{C}}_e(t)$ in analyses of JHS and REGARDS data (Figure 1). The difference in $\widehat{\mathrm{C}}_e(t)$ between the $\mathrm{ORSF}_{\mathrm{CV}}$ and the RSF was minimal in the GBSG2 analysis (27.86 and 28.71, respectively) and maximal in the GEBC analysis (25.08 and 31.19, respectively). The absolute (percent) increase in the mean value of $\widehat{\mathrm{C}}_e(t)$ from using the RSF instead of the $\mathrm{ORSF}_{\mathrm{CV}}$ or the CIF instead of the $\mathrm{ORSF}_{\mathrm{CV}}$ was 1.79 (9.08%) and 0.16 (0.82%), respectively. The variability and overall shape of time-dependent trajectories of $\widehat{\mathrm{C}}_e(t)$ were similar among the RSF, CIF and ORSF (Figure 2).

The three methods that achieved the lowest values of $\widehat{\mathcal{BS}}(T)$ were the ORSF, gradient boosted decision trees and the $\mathrm{ORSF}_{\mathrm{CV}}$ (mean ranks of 2.30, 2.50 and 3.90, respectively) (Table 5). Distributions of $\widehat{\mathcal{BS}}(T)$ were consistent among the

TABLE 4

Mean concordance index error for competing learning methods, aggregated over 250 *replicates of bootstrap cross-validation in five independent data sets. The minimum concordance error value for each analysis is written in bold text*

	Ensemble Survival Trees[§]					Proportional Hazards			
Scenario[†]	ORSF	ORSF_{CV}	CIF	RSF	Xgboost	CoxBoost	Lasso	Ridge	Step
				Concordance error: $100 \cdot \widehat{\text{C}}_e(t)$					
GEBC	**24.98**	25.08	25.79	31.19	27.08	28.62	28.59	26.47	–
GBSG2	28.10	**27.86**	28.40	28.71	28.09	30.24	30.30	29.99	29.88
JHS1	17.67	**17.58**	17.74	19.31	17.84	18.69	18.68	18.37	18.74
JHS3	17.61	**17.57**	17.91	18.40	17.63	18.54	18.90	19.66	18.07
JHS2	**13.13**	13.16	13.42	13.45	13.61	17.51	18.86	14.43	14.83
PBC	**12.25**	12.29	12.80	13.90	12.87	13.53	13.72	12.82	15.15
REGARDS1	**21.81**	21.83	21.84	22.69	22.15	22.68	22.57	22.35	24.58
REGARDS3	18.69	18.80	18.82	19.01	**18.55**	19.27	20.33	30.62	19.32
REGARDS2	13.91	**13.86**	14.21	16.08	14.01	15.64	17.87	29.38	16.93
MBC	29.41	29.27	**27.99**	32.47	31.28	40.62	39.79	35.14	–
				Percent increase in $\widehat{\text{C}}_e(t)$*, relative to minimum (0.00)*					
GEBC	**0.0**	0.4	3.3	24.9	8.4	14.6	14.5	6.0	–
GBSG2	0.9	**0.0**	1.9	3.0	0.8	8.5	8.7	7.6	7.2
JHS1	0.5	**0.0**	0.9	9.8	1.5	6.3	6.3	4.5	6.6
JHS3	0.2	**0.0**	1.9	4.7	0.3	5.5	7.5	11.9	2.8
JHS2	**0.0**	0.2	2.2	2.5	3.7	33.4	43.7	10.0	13.0
PBC	**0.0**	0.3	4.5	13.5	5.1	10.5	12.0	4.7	23.7
REGARDS1	**0.0**	0.1	0.1	4.1	1.6	4.0	3.5	2.5	12.7
REGARDS3	0.7	1.4	1.5	2.5	**0.0**	3.9	9.6	65.0	4.2
REGARDS2	0.4	**0.0**	2.6	16.1	1.1	12.8	29.0	112.0	22.2
MBC	5.1	4.6	**0.0**	16.0	11.7	45.1	42.2	25.5	–
				Rankings based on $\widehat{\text{C}}_e(t)$					
GEBC	**1**	2	3	8	5	7	6	4	–
GBSG2	3	**1**	4	5	2	8	9	7	6
JHS1	2	**1**	3	9	4	7	6	5	8
JHS3	2	**1**	4	6	3	7	8	9	5
JHS2	**1**	2	3	4	5	8	9	6	7
PBC	**1**	2	3	8	5	6	7	4	9
REGARDS1	**1**	2	3	8	4	7	6	5	9
REGARDS3	2	3	4	5	**1**	6	8	9	7
REGARDS2	2	**1**	4	6	3	5	8	9	7
MBC	3	2	**1**	5	4	8	7	6	–
Mean	1.80	**1.70**	3.20	6.40	3.60	6.90	7.40	6.40	7.25

[§] RSF = random survival forest; CIF = conditional inference forest; ORSF = oblique random survival forest; ORSF_{CV} = oblique random survival forest with internal cross-validation (see Section 3.2).

[†] GEBC = gene expression breast cancer; MBC = microarray breast cancer; GBSG2 = German breast cancer study group; PBC = primary biliary cirrhosis; JHS = Jackson heart study; REGARDS = REasons for Geographic And Racial Differences in Stroke.

[‡] Apparent ties in concordance errors are a result of rounding errors.

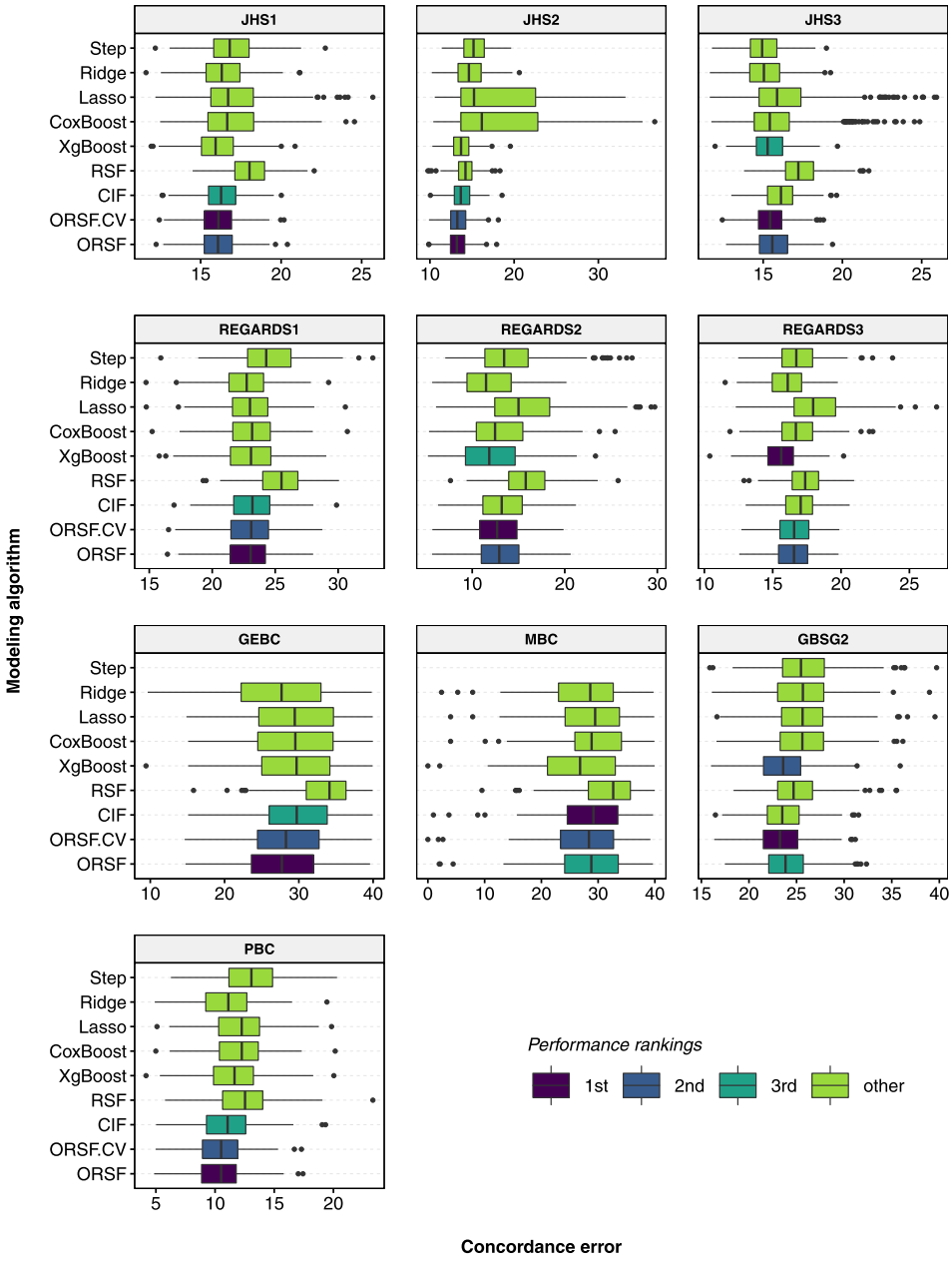

FIG. 1. *Concordance error values, aggregated over* 250 *replications of bootstrap cross-validation, for competing methods in ten analyses of data with right-censored time-to-event outcomes.*

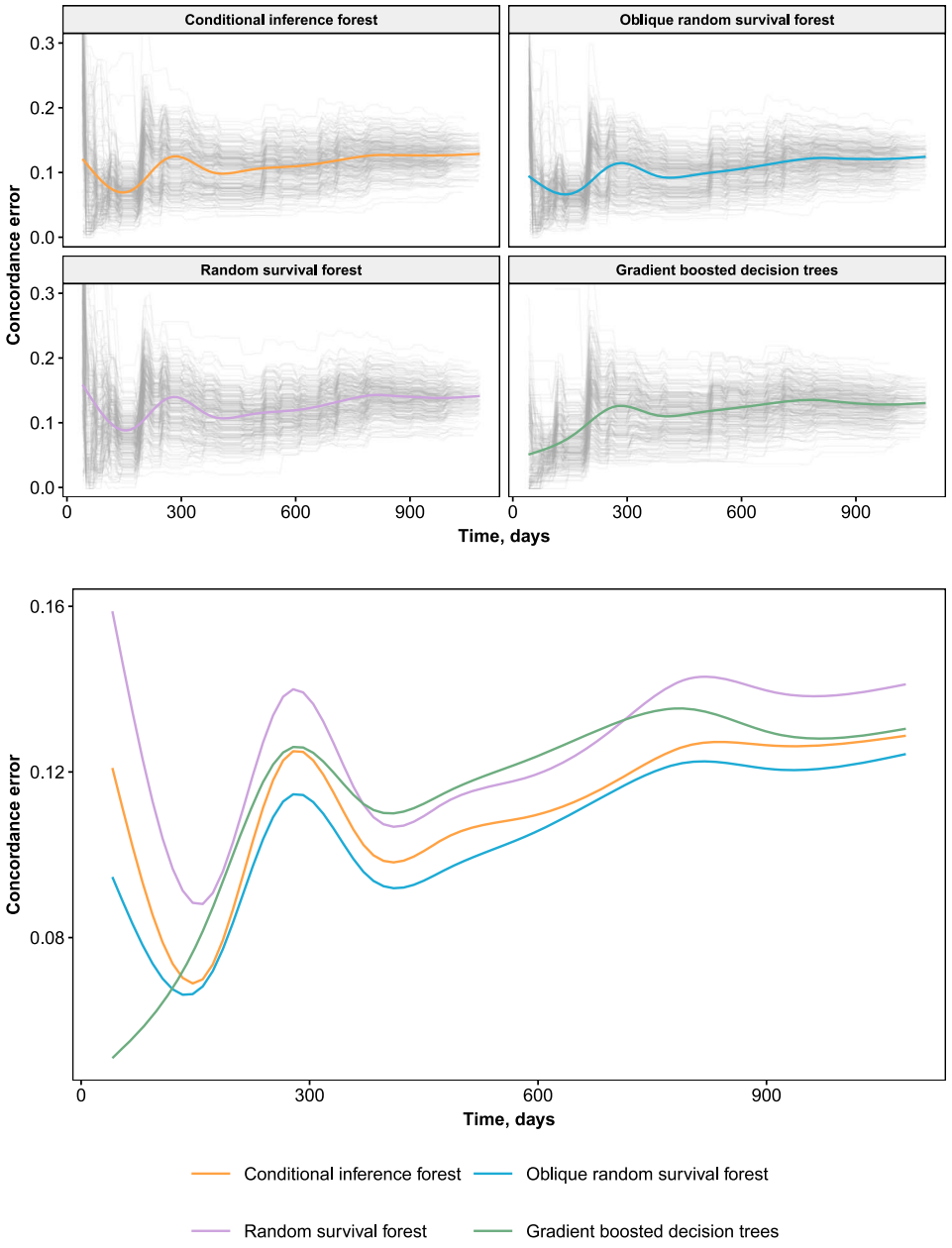

FIG. 2. *Concordance error values for the primary biliary cirrhosis analysis. Results from individual replications of bootstrap cross-validation are shown as grey trajectories. Smoothed estimates of average values are colored. Error values are plotted from baseline to median event time.*

TABLE 5

Mean integrated Brier scores for competing learning methods, aggregated over 250 replicates of bootstrap cross-validation in five independent data sets. The minimum Brier score for each analysis is written in bold text

	Ensemble Survival Trees[§]					Proportional Hazards				
Scenario[†]	ORSF	ORSF_{CV}	CIF	RSF	Xgboost	CoxBoost	Lasso	Ridge	Step	Reference
					Integrated Brier score: $100 \cdot \widehat{\mathcal{BS}}(T)$					
GEBC	**4.20**	4.20	4.23	4.44	4.25	4.27	4.27	4.22	–	4.36
GBSG2	7.10	7.12	**7.10**	7.21	7.20	7.51	7.51	7.49	7.47	7.77
JHS1	2.31	2.31	2.33	2.32	**2.31**	2.35	2.36	2.34	2.35	2.49
JHS3	3.23	3.23	3.28	3.23	**3.15**	3.20	3.23	3.20	3.17	3.67
JHS2	2.30	2.30	2.33	**2.29**	2.31	2.41	2.41	2.37	2.38	2.55
PBC	**5.71**	5.82	5.79	5.77	5.73	6.08	6.12	5.92	6.45	8.09
REGARDS1	1.69	1.69	1.69	1.70	**1.68**	1.70	1.70	1.70	1.73	1.76
REGARDS3	2.21	2.22	2.23	2.20	**2.16**	2.18	2.21	2.26	2.21	2.43
REGARDS2	0.63	0.63	0.64	0.64	**0.63**	0.64	0.64	0.64	0.66	0.66
MBC	9.07	9.08	9.11	9.62	9.16	9.65	9.67	9.33	–	**8.91**
					Scaled $\widehat{\mathcal{BS}}(T)$ *values*: $100 \cdot [1 - \widehat{\mathcal{BS}}(T)/\text{Reference}]$					
GEBC	**3.58**	3.53	3.03	−1.94	2.51	1.96	1.95	3.20	–	0.00
GBSG2	8.60	8.36	**8.68**	7.18	7.33	3.40	3.29	3.55	3.82	0.00
JHS1	7.46	7.37	6.75	6.84	**7.46**	5.67	5.53	6.21	5.59	0.00
JHS3	12.12	12.05	10.79	12.01	**14.27**	12.84	12.10	12.95	13.80	0.00
JHS2	9.95	9.74	8.54	**10.39**	9.40	5.45	5.33	7.01	6.63	0.00
PBC	**29.40**	28.03	28.42	28.59	29.10	24.81	24.29	26.83	20.28	0.00
REGARDS1	3.99	3.79	3.58	3.11	**4.27**	3.20	3.23	3.43	1.41	0.00
REGARDS3	8.93	8.50	7.85	9.16	**11.04**	10.24	8.86	6.63	8.88	0.00
REGARDS2	4.38	4.14	3.55	3.78	**5.31**	3.74	2.97	2.47	0.76	0.00
MBC	−1.76	−1.85	−2.27	−7.90	−2.73	−8.22	−8.49	−4.66	–	**0.00**
					Ranks in each data set of competing methods based on $\widehat{\mathcal{BS}}(T)$					
GEBC	**1**	2	4	9	5	6	7	3	–	8
GBSG2	2	3	**1**	5	4	8	9	7	6	10
JHS1	2	3	5	4	**1**	7	9	6	8	10
JHS3	5	7	9	8	**1**	4	6	3	2	10
JHS2	2	3	5	**1**	4	8	9	6	7	10
PBC	**1**	5	4	3	2	7	8	6	9	10
REGARDS1	2	3	4	8	**1**	7	6	5	9	10
REGARDS3	4	7	8	3	**1**	2	6	9	5	10
REGARDS2	2	3	6	4	**1**	5	7	8	9	10
MBC	2	3	4	7	5	8	9	6	–	**1**
Mean	**2.30**	3.90	5.00	5.20	2.50	6.20	7.60	5.90	6.88	8.90

[§] RSF = random survival forest; CIF = conditional inference forest; ORSF = oblique random survival forest; ORSF_{CV} = oblique random survival forest with internal cross-validation (see Section 3.2).

[†] GEBC = gene expression breast cancer; MBC = microarray breast cancer; GBSG2 = German breast cancer study group; PBC = primary biliary cirrhosis; JHS = Jackson heart study; REGARDS = REasons for Geographic And Racial Differences in Stroke.

[*] The reference values are expected Brier scores of a prediction model that ignores all predictor variables, that is, the Kaplan–Meier estimate calculated using the training data.

[‡] Apparent ties in Brier scores are a result of rounding errors.

competing learning methods with the exception of the penalized Cox PH models, which had higher variability of $\widehat{\mathcal{BS}}(T)$ in analyses of JHS and REGARDS data (Figure 3). The difference in $\widehat{\mathcal{BS}}(T)$ between the ORSF and the RSF was minimal in the JHS3 analysis (3.228 and 3.232, respectively) and maximal in the GEBC analysis (4.20 and 4.44, respectively). The absolute (percent) increase in the mean *scaled value* (see Section 4.3) of $\widehat{\mathcal{BS}}(T)$ using the ORSF instead of the RSF or instead of the CIF was 1.54 (21.67%) and 0.77 (9.79%), respectively.

Monte Carlo error was adequate when 1000 trees were used to fit the ORSF (Figure 4). The variability in predicted survival curves was noticeably higher when 10 or 100 trees were used to fit an ORSF.

6. Overall performance comparisons. Here we describe a formal comparison of the performance and computational requirements of each method that was applied in Section 5, incorporating results from each of the real data sets described. To assess the relative performance of each method, we ranked the methods in each data set, separately, giving a rank of 1 to the method with the lowest error, a rank of 2 to the method with the second lowest error, and so on. We recorded the rankings in this manner using both $\widehat{\mathrm{C}}_e(t)$ and $\widehat{\mathcal{BS}}(T)$ as the metric for error. In Section 6.1, we describe the procedure used to draw inferences from these rankings. We summarize performance of the learning algorithms in Section 6.2 and the computational resources required to run each algorithm in Section 6.3.

6.1. *Statistical comparisons of ranks.* We applied a modification of Friedman's nonparametric rank test (Friedman (1937)), which compares the average ranks among a set of classifiers over a collection of data sets. Let r_i^j be the rank of the jth of k classifiers on the ith of D datasets. Friedman's test compares the average ranks of the competing classifiers, $R_j = \frac{1}{D}\sum_{i=1}^{D} r_i^j$. Under the null hypothesis of equivalent performance, each classifier will have a mean rank of $(k+1)/2$, with variance $(k^2-1)/(12D)$ and

$$F_F = \frac{(D-1)\chi_F^2}{D(k-1)-\chi_F^2} \sim F_{k-1,(k-1)(D-1)}, \tag{6.1}$$

where $F_{p,q}$ denotes an F distribution with p and q numerator and denominator degrees of freedom, respectively, and

$$\chi_F^2 = \frac{12D}{k(k+1)}\left[\left(\sum_{j=1}^{k} R_j^2\right) - \frac{k(k+1)^2}{4}\right]. \tag{6.2}$$

Iman and Davenport (1980) derived F_F to correct the overly conservative χ_F^2 statistic (Demšar (2006)). If the null hypothesis of overall equivalent performance is rejected, post-hoc pairwise comparisons can be conducted using

$$z = \frac{R_i - R_j}{\sqrt{k(k+1)/(6D)}} \tag{6.3}$$

to compare the average ranking of the ith and jth classifiers.

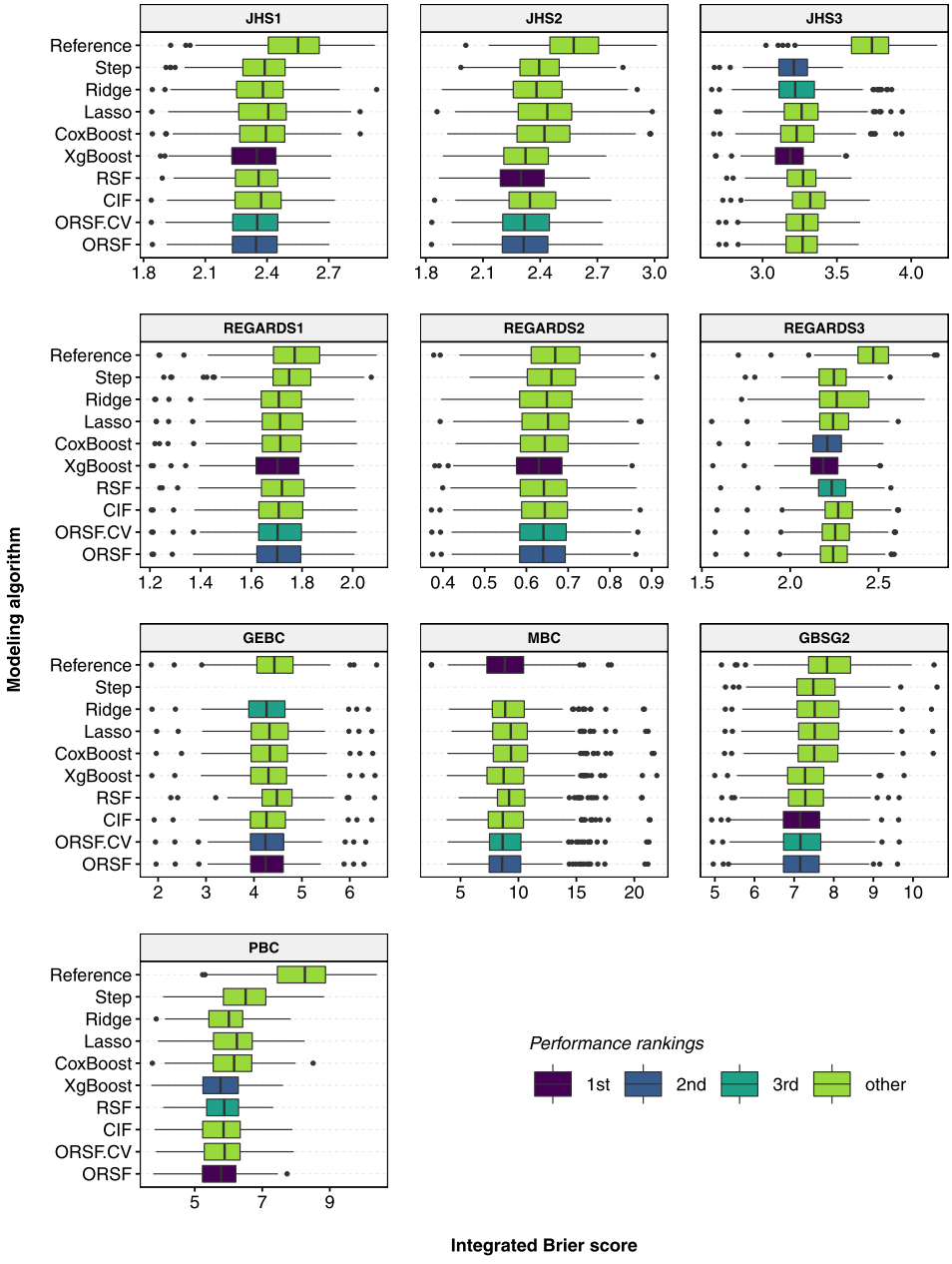

FIG. 3. *Integrated Brier score values, aggregated over* 250 *replications of bootstrap cross-validation, for competing methods in ten analyses of data with right-censored time-to-event outcomes.*

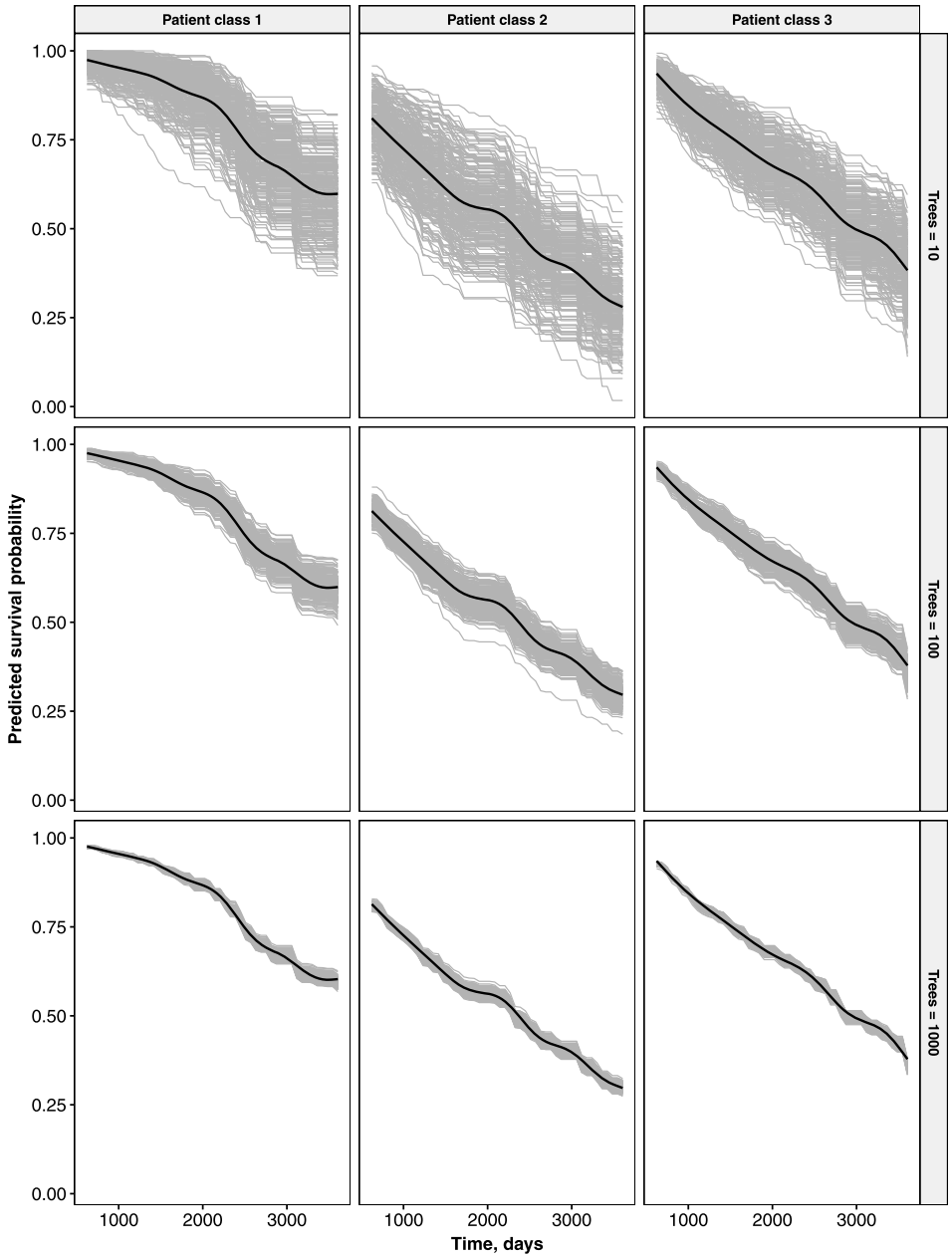

FIG. 4. *Monte Carlo error* (*seed effect*) *for the primary biliary cirrhosis analysis. Results are shown for three prototypical patients. Grey curves are individual predictions from an individual forest. Smoothed estimates of average values are drawn as black curves.*

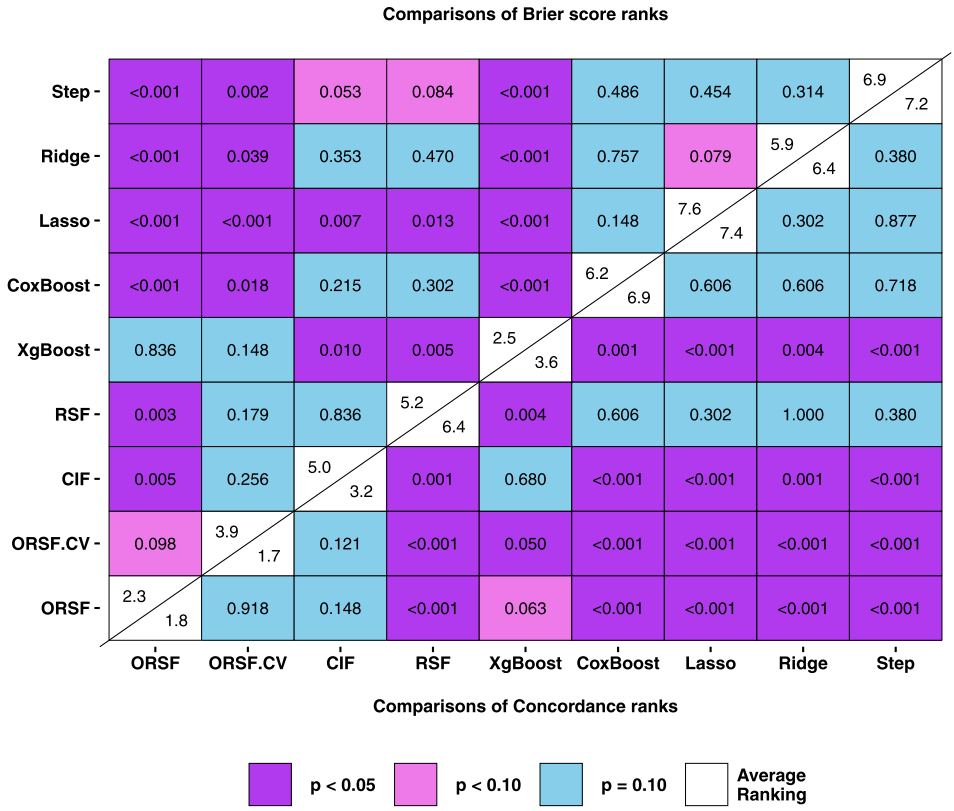

FIG. 5. *Overall performance comparison between oblique random survival forests and competing learning methods for prediction of the data described in Sections 4 and 5.*

6.2. *Results.* Among the nine learning algorithms we compared, the hypothesis of equivalence in performance was rejected ($F_F = 74.1$ [$p < 0.001$] and 17.2 [$p < 0.001$] for concordance error and integrated Brier scores, respectively) using the overall F_F test statistic proposed by Iman and Davenport (1980). Figure 5 shows a color-coded comparison matrix. Diagonal cells are split in two, with the mean rank according to $\widehat{\mathcal{BS}}(T)$ printed in the upper left and the mean rank according to $\widehat{\mathrm{C}}_e(t)$ printed in the lower right. The upper-left and lower-right sections of the matrix show two sided p-values corresponding to pairwise comparisons of the mean ranks between two methods using the rankings according to $\widehat{\mathcal{BS}}(T)$ and $\widehat{\mathrm{C}}_e(t)$, respectively. For example, the mean rankings according to $\widehat{\mathrm{C}}_e(t)$ of the $\mathrm{ORSF}_{\mathrm{CV}}$ and the CIF were 1.70 and 3.20, respectively, and the p-value corresponding to a test of rank equivalence between these two methods was 0.121.

After ranking each method in each analysis according to $\widehat{\mathrm{C}}_e(t)$, the $\mathrm{ORSF}_{\mathrm{CV}}$ had the lowest mean ranking: 1.70. The ORSF and CIF had the second and third lowest mean rankings, 1.80 and 3.20, respectively. For all pairwise comparisons involving

TABLE 6
Mean computation time, in seconds, required to fit each learning algorithm in each analysis

Analysis	ORSF	ORSF.CV	CIF	RSF	Xgboost	CoxBoost	Lasso	Ridge	Step
GEBC	1265.7	5848.1	9.8	6.5	275.1	126.3	20.0	14.4	0.0
GBSG2	401.6	1033.3	6.0	1.9	234.4	24.9	0.3	0.3	0.8
PBC	204.0	693.1	2.1	0.8	276.5	8.2	0.2	0.3	4.2
MBC	101.8	291.9	19.2	0.8	316.7	27.7	3.7	4.0	0.0
JHS1	1621.8	5005.0	56.3	6.4	293.6	203.0	3.3	3.5	41.5
JHS3	2556.5	7592.1	62.2	10.6	431.6	413.3	2.9	3.1	49.5
JHS2	1291.9	4385.3	50.5	5.4	265.3	159.3	3.3	3.3	36.4
REGARDS1	1852.3	5142.9	132.4	11.8	533.6	267.7	8.4	8.0	52.6
REGARDS3	2655.5	6873.5	140.6	19.1	525.7	476.2	4.7	5.1	67.5
REGARDS2	909.3	3127.5	120.1	6.3	342.7	109.5	16.2	12.8	48.9
SIM3.25	1448.9	4316.8	8.5	7.1	239.7	79.8	0.3	0.3	5.6
SIM3.50	1593.7	5123.4	7.9	7.7	191.1	92.9	0.4	0.4	15.0
SIM2.25	1084.2	3605.3	6.7	4.4	158.1	56.8	0.3	0.3	6.1
SIM2.50	1311.2	4360.8	7.8	6.6	199.7	82.5	0.4	0.5	28.3
SIM1.25	1176.6	3735.7	7.8	5.8	242.5	68.7	0.3	0.3	9.6
SIM1.50	1342.1	4583.2	8.3	7.5	313.8	87.0	0.5	0.5	30.7
Overall	1301.1	4107.4	40.4	6.8	302.5	142.7	4.1	3.6	24.8

$\widehat{\mathrm{C}}_e(t)$ for the $\mathrm{ORSF}_{\mathrm{CV}}$, excluding comparison with the ORSF ($p = 0.918$) and CIF, a statistically significant ($p < 0.05$) difference in mean rankings was observed.

After ranking each method in each analysis according to $\widehat{\mathcal{BS}}(T)$, the ORSF had the lowest mean ranking (2.30) followed by gradient boosted decision trees (mean rank of 2.50) and the $\mathrm{ORSF}_{\mathrm{CV}}$ (mean rank of 3.90). For all pairwise comparisons involving $\widehat{\mathcal{BS}}(T)$ for the ORSF, excluding comparisons with the $\mathrm{ORSF}_{\mathrm{CV}}$ ($p = 0.098$) and gradient boosted decision trees ($p = 0.836$), a statistically significant ($p < 0.05$) difference in mean rankings was observed.

6.3. *Computing time*. We recorded the mean amount of time required to fit each learning algorithm in each analysis using real and simulated data (Table 6). As fitting an ORSF requires fitting regularized Cox PH models in each nonterminal node, the mean time of computation for the ORSF was roughly 192 and 32 times that of the RSF and CIF, respectively. When nested cross-validation was used, the mean time of computation for the ORSF was roughly 605 and 102 times that of the RSF and CIF, respectively.

7. Application to Jackson heart study. Here we apply the ORSF and the CIF in a comparative example using data from the Jackson Heart Study (JHS). A description of the JHS is given in Section 5.1.5. In the current example, we included 5126 JHS participants who consented to provide follow-up information on ASCVD events. We randomly split the baseline data from these JHS participants

into a training set ($N = 3000$) and a testing set ($M = 2126$). We imputed missing values in both sets, separately, using a k-nearest neighbors algorithm (Kowarik and Templ (2016)). Using the training data, we developed prediction rules based on the ORSF and CIF, separately. These prediction rules were applied to compute ten-year predicted risk of ASCVD events for JHS participants in the testing data. Using these values of predicted risk, we created variable dependence and partial dependence plots (Section 7.1) that directly compare expected predictions from the ORSF and CIF without and with adjustment for confounding effects, respectively (Friedman (2001)). For variable dependence plots, we included ten-year predicted risk of ASCVD events according to the Pooled Cohort risk equations, a well-known risk prediction equation that is recommended by Whelton et al. (2018) for clinical assessment of ten-year predicted risk for ASCVD events. Last, we compared predicted risk curves according to the ORSF and CIF for four individual JHS participants from the testing data.

7.1. *Variable dependence and partial dependence.* Variable dependence plots render the expected value of an outcome (y-axis) relative to the observed values of an input variable (x-axis), using the unaltered training data. Although variable dependence plots illustrate observed relationships, they do not account for variation in covariates of interest apart from the input variable. *Partial* dependence plots show predicted risk as a function of the variables in a designated subset of input variables by averaging effects of the designated variables over the observed distribution of variables not in the designated set. Let

$$z_{\{r\}} = \{z_{(1)}, \ldots, z_{(r)}\} \subset \{\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(p)}\}$$

denote the designated subset of $r < p$ predictor variables. Define the complement set of predictor variables $z_{\{s\}}$ such that

$$z_{\{r\}} \cup z_{\{s\}} = \{\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(p)}\}.$$

The predicted risk function, $\widehat{F}(\boldsymbol{x})$, depends on variables in both subsets. Friedman (2001) shows that conditioning on the variables in $z_{\{s\}}$ allows $\widehat{F}(\boldsymbol{x})$ to be considered as a function of only the variables chosen in $z_{\{r\}}$. Moreover, the partial dependence of $\widehat{F}(\boldsymbol{x})$ on variables in $z_{\{r\}}$ can be computed by averaging over the observed values of $z_{\{s\}}$, that is,

$$\bar{F}(z_{\{r\}}) = E\big[\widehat{F}(\boldsymbol{x} \mid z_{\{s\}})\big] = \int \widehat{F}(\boldsymbol{x}) P(z_{\{s\}})\, dz_{\{s\}}. \tag{7.1}$$

7.2. *Results.* We identified four explanatory variables to analyze for the current application of the ORSF and CIF to data from the JHS: (1) left ventricular mass (LVM) in g/m^2, (2) age in years, (3) estimated glomerular filtration rate (eGFR) (Levey et al. (2009)) in ml/min/1.73m^2 and (4) systolic blood pressure

TABLE 7
Descriptive statistics for four variables in the Jackson Heart Study

Summary measures[§]	Min.	Percentile 25%	50%	75%	Max.	Mean	SD
Testing data							
Age, years	20.60	45.10	55.20	64.40	93.10	55.04	12.77
eGFR, ml/min/1.73 m^2	4.29	81.48	96.90	110.25	153.46	94.94	22.24
Left ventricular mass, g/m^2	20.84	59.55	69.17	81.33	215.50	72.88	20.27
Systolic blood pressure, mm Hg	77.98	115.58	125.66	136.67	228.36	127.35	16.91
Training data							
Age, years	21.00	45.70	56.10	65.20	95.50	55.58	12.93
eGFR, ml/min/1.73 m^2	4.14	80.47	95.24	109.09	151.88	93.59	21.81
Left ventricular mass, g/m^2	5.28	60.16	69.26	82.56	225.45	73.30	20.10
Systolic blood pressure, mm Hg	86.24	116.26	125.66	136.67	221.02	127.60	16.90

[§] Min = minimum; Max = maximum; SD = standard deviation.

(SBP) in mm Hg, measured in a clinical setting. We performed routine descriptive analyses of these variables in the training and testing datasets, separately (Table 7 and Figure 6).

We created variable and partial dependence plots for each explanatory variable and ASCVD risk, separately, using the ORSF and the CIF. Variable dependence plots indicated that predictions from the ORSF demonstrate stronger alignment than the CIF with the Pooled Cohort Risk equations (Figure 7). Partial dependence plots show that the ORSF's predicted risk function generally differs from the CIF in the upper or lower range of values for each explanatory variable (Figure 8). This pattern of difference in predicted risk may explain the observed differences (Figure 9, top row) and lack of differences (Figure 9, bottom row) in survival curves for four JHS participants in the testing data.

8. Discussion.

8.1. *Summary.* In this article, we have introduced the ORSF and assessed its predictive accuracy. The ORSF extends current implementations of ensemble methods for right-censored time-to-event analyses by applying a recursive partitioning algorithm that can incorporate LCIVs. Our results indicated that the ORSF may provide substantial improvement in discrimination (i.e., lower concordance error) and minor improvement in Brier scores compared to state-of-the-art learning algorithms. Using data from participants in the JHS, we compared dependence plots from the ORSF and CIF using four explanatory variables, separately, for ten-year predicted risk of ASCVD events. Our results demonstrated differences between predicted risk for ASCVD according to the ORSF and CIF. Results also

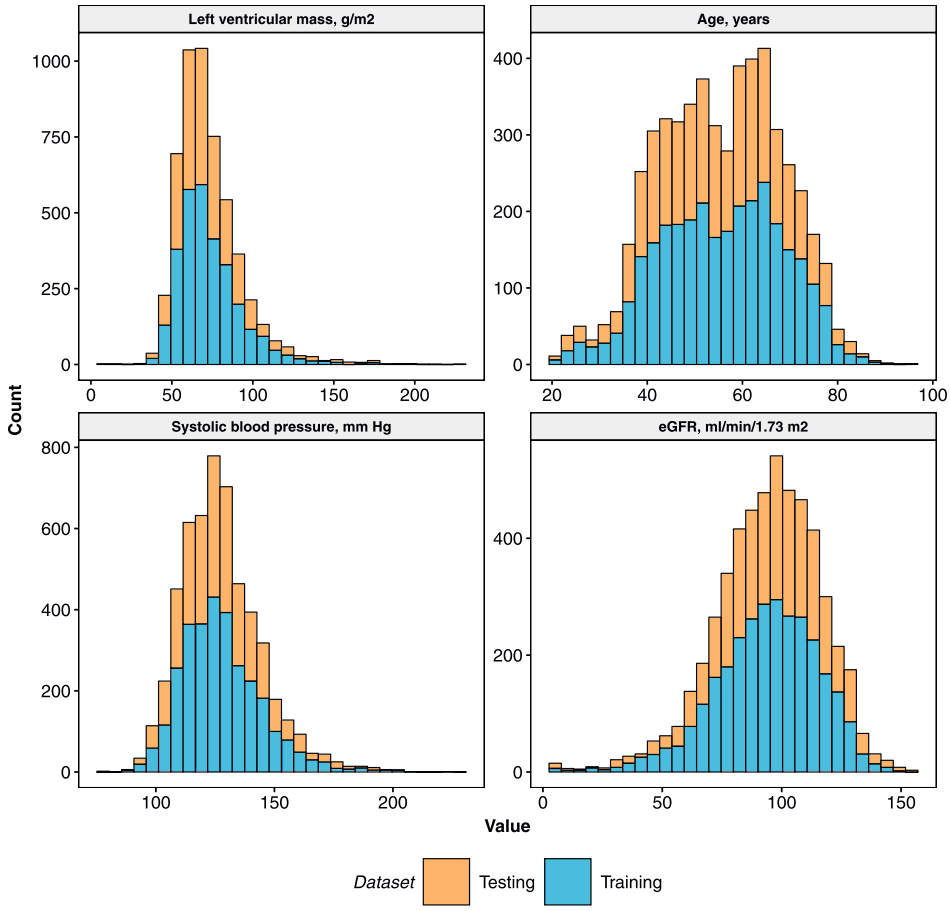

FIG. 6. *Descriptive summary of left ventricular mass, age, systolic blood pressure and estimated glomerular filtration rate. Data are from the Jackson Heart Study.*

showed that, according to the four variables we analyzed, the ORSF's predicted risk function had stronger alignment with the Pooled Cohort Risk equations compared to the CIF.

8.2. *Why the ORSF works* (*and when it does not*). In our application to real data, we found that the ORSF's predicted risk function may substantially lower concordance error compared to the RSF and Brier score compared to the CIF. Given the similarity in these three algorithms, the most likely explanation for the differences we observed is the use of LCIVs, which has been shown to result in additional accuracy and diversity of individual trees in the ensemble (Breiman (2001), Menze et al. (2011), Rainforth and Wood (2015)). As an informal demonstration (not a comprehensive comparison), we computed predicted survival curves

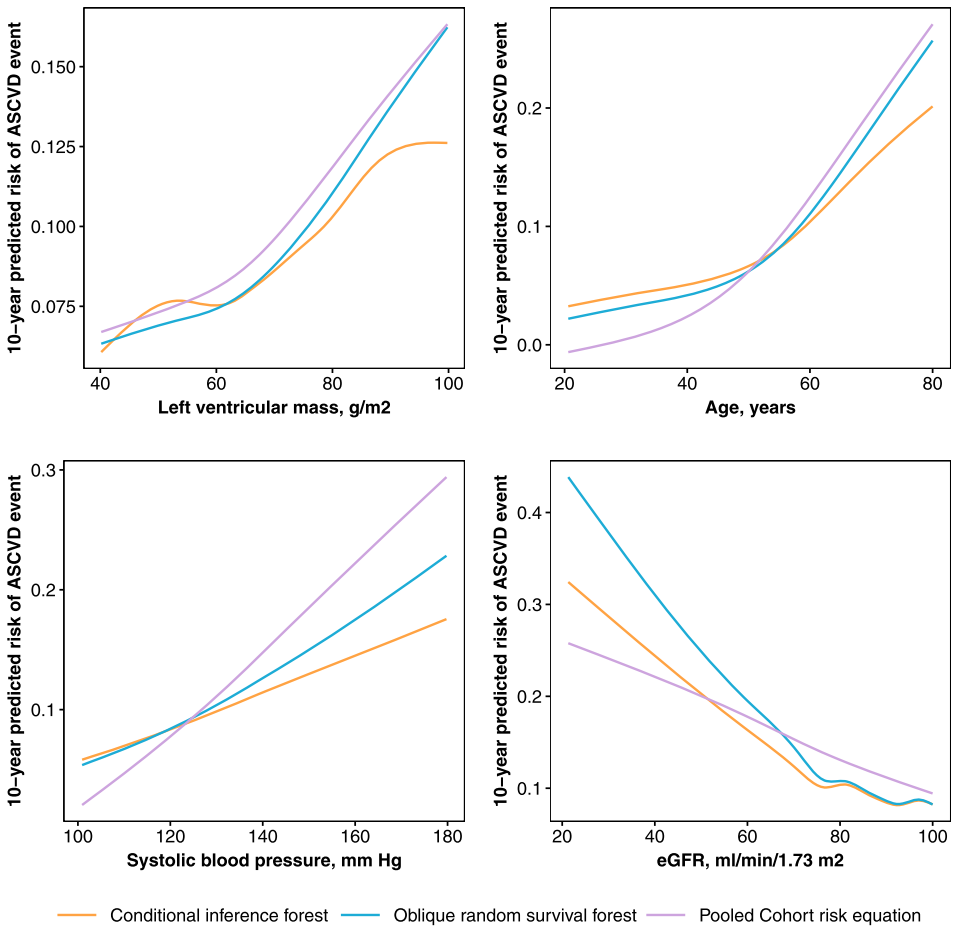

FIG. 7. *Variable dependence plots for ten-year risk of stroke or coronary heart disease as a function of left ventricular mass, age, systolic blood pressure, and estimated glomerular filtration rate, separately, according to data from participants in the Jackson Heart Study.*

for one observation generated from Scenario C in our simulation study using the ORSF and RSF. Figure 10 illustrates predicted survival curves for this observation from each survival tree in the fitted ORSF and RSF ensembles, separately. In the RSF, many trees give nearly identical survival curves for this participant. On the other hand, there are many more types of survival curves in the ORSF ensemble (diversity). Additionally, survival curves in the ORSF tend to be more aligned with the true survival curve (accuracy). This figure shows one instance where predictions from trees in the ORSF exhibit a clear increase in diversity and accuracy compared to survival trees in the RSF ensemble. However, this is clearly not a claim that this pattern holds in general.

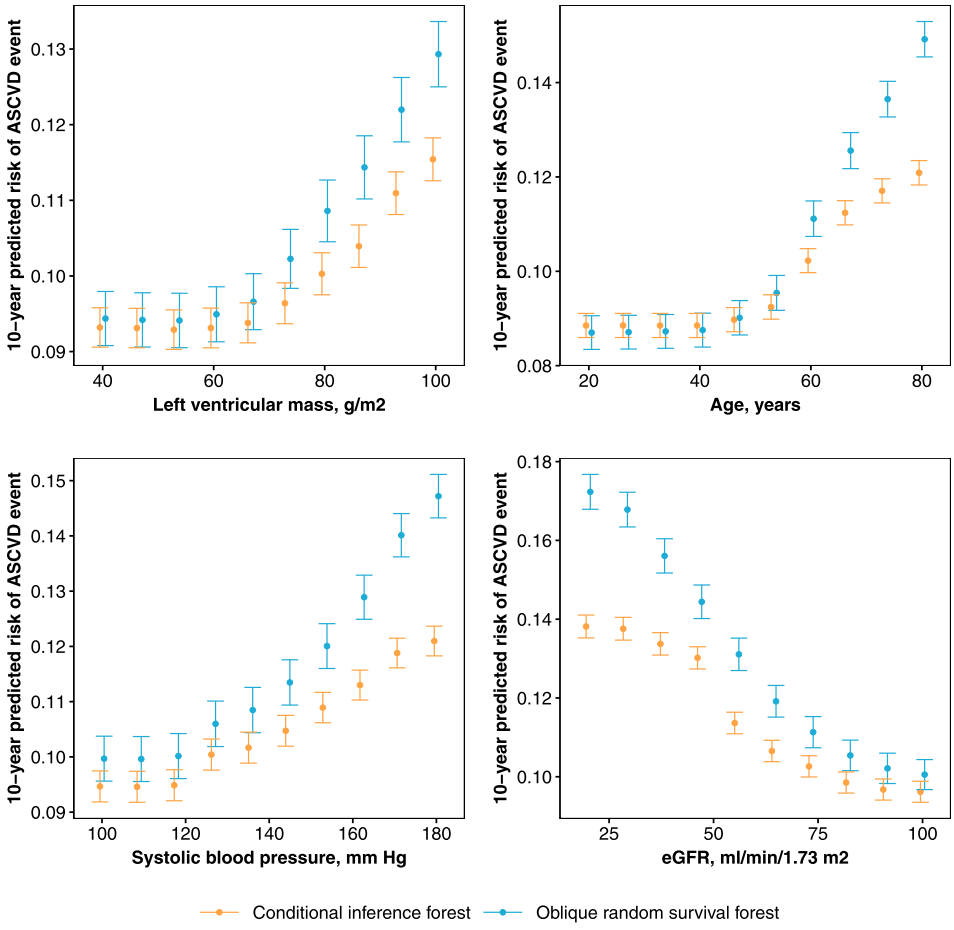

FIG. 8. *Partial dependence plots for ten-year risk of stroke or coronary heart disease as a function of left ventricular mass, age, systolic blood pressure, and estimated glomerular filtration rate, separately, according to data from participants in the Jackson Heart Study.*

In our simulated trials, the ORSF recorded higher Brier scores than gradient boosted decision trees. However, this pattern was reversed in the analysis of real data. This is likely due to the invalidity of the proportional hazards assumption in many of the real data problems we considered. These results suggest that in analyses where the proportional hazards assumption is entirely valid, the ORSF and other implementations of RFs for survival analyses may not be able to compete with gradient boosted decision trees or penalized Cox PH models.

In addition to LCIVs, early stopping and class-specific shrinkage of categorical predictors are important components of the ORSF that are based on the CIF (Nasejje et al. (2017)). The protocol described by Breiman (1984) is known to result in overfitting and a selection bias towards covariates with many possible splits

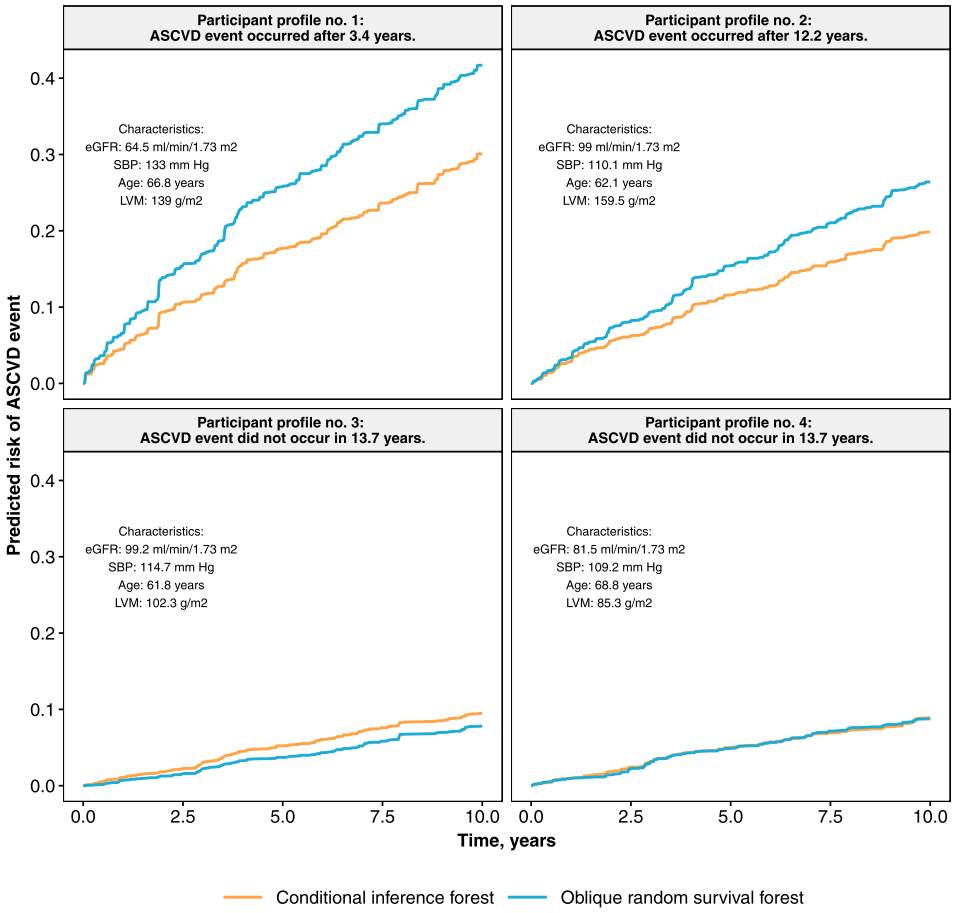

FIG. 9. *Predicted survival curves for four JHS participants in the testing data.*

(i.e., unordered categorical variables) (Hothorn, Hornik and Zeileis (2006)). The ORSF and CIF each apply a mechanism for early stopping if a certain level of statistical association is not measured, and this may explain why both the ORSF and the CIF provided relatively low concordance error. We speculate that early stopping prevents these ensembles from growing decision trees that overfit the training data. However, the ORSF and CIF also provided relatively unimpressive Brier scores in our simulation study and real data analysis compared to gradient boosted decision trees. Therefore, we speculate that early stopping may prevent decision trees in the ORSF and CIF from partitioning a training set with the same level of granularity and accuracy as gradient boosting, especially when the proportional hazard assumption is valid (i.e., scenarios A and B in our simulation study).

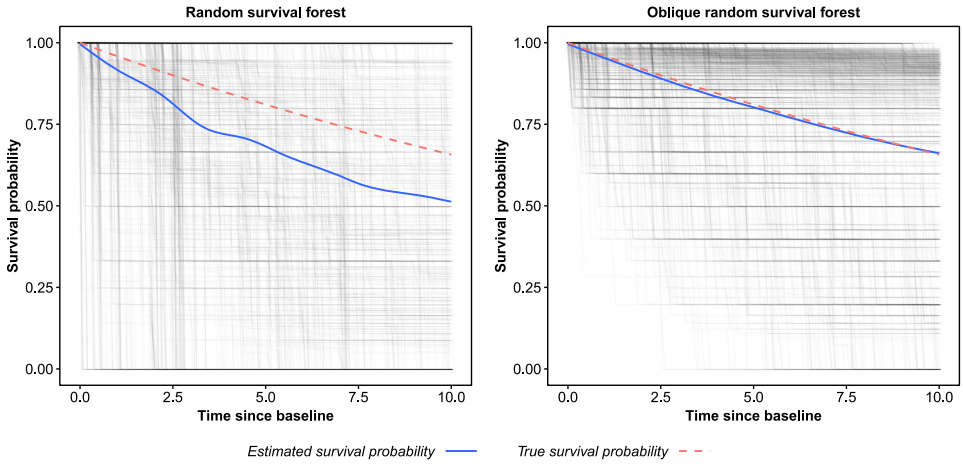

FIG. 10. *Tree and ensemble predicted survival functions for a single observation from simulated data* (*Scenario C with* $p = 50$; *see Section* 4.1). *Estimates are from the random survival forest* (*left*) *and oblique random survival forest* (*right*).

8.3. *Limitations and future research.* The additional computational resources required by the ORSF are a clear limitation. While future development and optimization of the `C++` routines in the `obliqueRSF` package may alleviate this limitation, there are several additional topics of interest. For example, the ORSF could be extended to assess the importance of individual variables or clusters of variables. Multiple imputation could be incorporated into the ORSF by imputing missing values separately for each tree in the ensemble using a shared imputation model or a shared set of imputation models. To generate additional diversity in the ensemble, a stochastic error term could be added to the imputation model's predictions. Predictions from the the ORSF could also be linked to programs that operate on the predicted values of a tree-based ensemble learning algorithm, such as the method of conducting statistical inference introduced by Mentch and Hooker (2016) or the efficient method to compute Shapley additive explanatory values discussed in Lundberg, Erion and Lee (2018). Last, a hybrid algorithm for ORSF that applies cross-validation conditional on sufficient sample size in the current node could be developed. In summary, the `obliqueRSF` R package offers a powerful learning algorithm, the ORSF (and optionally, the ORSFCV), which may exceed the performance of state-of-the-art learning methods for right-censored time-to-event analyses.

Acknowledgements. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Neurological Disorders and Stroke or the National Institutes of Health. Representatives of the funding agency have been involved in the review of the manuscript but not directly involved in the collection, management, analysis or

interpretation of the data. The authors thank the other investigators, the staff and the participants of the REGARDS study for their valuable contributions. A full list of participating REGARDS investigators and institutions can be found at http://www.regardsstudy.org.

The authors also wish to thank the staffs and participants of the JHS. More information about the JHS can be found at https://www.jacksonheartstudy.org/.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Neurological Disorders and Stroke, the National Institutes of Health, The National Heart, Lung, and Blood Institute, The National Institute on Minority Health and Health Disparities, or the Department of Health and Human Services.

SUPPLEMENTARY MATERIAL

Source code for analyses presented in the oblique random survival forest manuscript (DOI: 10.1214/19-AOAS1261SUPP; .zip). Provides scripts written in R that were applied to generate the results presented in the manuscript. In particular, the scripts were applied to conduct the simulation/resampling study and the application of oblique random survival forests to the Jackon Heart Study.

REFERENCES

ANDERSEN, P. K., BORGAN, O., GILL, R. D. and KEIDING, N. (2012). *Statistical Models Based on Counting Processes*. Springer, Berlin

BIEN, J. and TIBSHIRANI, R. (2019). protoclust: Hierarchical clustering with prototypes. R package version 1.6.3.

BINDER, H. (2013). CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks. R package version 1.4, available at https://CRAN.R-project.org/package=CoxBoost.

BLANCHE, P., KATTAN, M. W. and GERDS, T. A. (2019). The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics* **20** 347–357. MR3922138

BOU-HAMAD, I., LAROCQUE, D. and BEN-AMEUR, H. (2011). A review of survival trees. *Stat. Surv.* **5** 44–71. MR3018509

BREIMAN, L. (1984). *Classification and Regression Trees*. Routledge, Abingdon.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BREIMAN, L. and CUTLER, A. (2003). Setting up, using, and understanding random forests V4.0. Dept. Statistics, Univ. California, Berkeley.

BRILLEMAN, S. (2018). simsurv: Simulate survival data. R package version 0.2.2, available at https://CRAN.R-project.org/package=simsurv.

BURNHAM, K. P. and ANDERSON, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* **33** 261–304. MR2086350

CHEN, T. and GUESTRIN, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining* 785–794. ACM.

CHEN, T., HE, T., BENESTY, M., KHOTILOVICH, V., TANG, Y., CHO, H., CHEN, K., MITCHELL, R., CANO, I. et al. (2019). xgboost: Extreme gradient boosting. R package version 0.81.0.1, available at https://CRAN.R-project.org/package=xgboost.

COX, D. R. (1992). Regression models and life-tables. In *Breakthroughs in Statistics* 527–541. Springer, Berlin.

DEMŠAR, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7** 1–30. MR2274360

DESMEDT, C., DI LEO, A., DE AZAMBUJA, E., LARSIMONT, D., HAIBE-KAINS, B., SELLESLAGS, J., DELALOGE, S., DUHEM, C., KAINS, J.-P. et al. (2011). Multifactorial approach to predicting resistance to anthracyclines. *J. Clin. Oncol.* **29** 1578–1586.

DHEERU, D. and KARRA TANISKIDOU, E. (2017). UCI Machine learning repository. Univ. California, Irvine.

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

FLEMING, T. R. and HARRINGTON, D. P. (2011). *Counting Processes and Survival Analysis* **169**. Wiley, New York.

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22. Available at http://www.jstatsoft.org/v33/i01/.

FRIEDMAN, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* **32** 675–701.

GERDS, T. A., KATTAN, M. W., SCHUMACHER, M. and YU, C. (2013). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Stat. Med.* **32** 2173–2184. MR3067376

GEURTS, P., ERNST, D. and WEHENKEL, L. (2006). Extremely randomized trees. *Mach. Learn.* **63** 3–42.

GRAF, E., SCHMOOR, C., SAUERBREI, W. and SCHUMACHER, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18** 2529–2545.

HARRELL, F. E., CALIFF, R. M., PRYOR, D. B., LEE, K. L. and ROSATI, R. A. (1982). Evaluating the yield of medical tests. *JAMA* **247** 2543–2546.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer Series in Statistics*. Springer, New York. MR1851606

HATZIS, C., PUSZTAI, L., VALERO, V., BOOSER, D. J., ESSERMAN, L., LLUCH, A., VIDAURRE, T., HOLMES, F., SOUCHON, E. et al. (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* **305** 1873–1881.

HEAGERTY, P. J., LUMLEY, T. and PEPE, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56** 337–344.

HEAGERTY, P. J. and ZHENG, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61** 92–105. MR2135849

HOTHORN, T., HORNIK, K., STROBL, C. and ZEILEIS, A. (2019). party: A laboratory for recursive partytioning. R package version 1.3.3, available at https://CRAN.R-project.org/package=party.

HOTHORN, T., HORNIK, K. and ZEILEIS, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Statist.* **15** 651–674. MR2291267

HOTHORN, T. and LAUSEN, B. (2003). Double-bagging: Combining classifiers by bootstrap aggregation. *Pattern Recognit.* **36** 1303–1309.

HOTHORN, T., LAUSEN, B., BENNER, A. and RADESPIEL-TRÖGER, M. (2004). Bagging survival trees. *Stat. Med.* **23** 77–91.

HOWARD, V. J., CUSHMAN, M., PULLEY, L., GOMEZ, C. R., GO, R. C., PRINEAS, R. J., GRAHAM, A., MOY, C. S. and HOWARD, G. (2005). The reasons for geographic and racial differences in stroke study: Objectives and design. *Neuroepidemiology* **25** 135–143.

IMAN, R. L. and DAVENPORT, J. M. (1980). Approximations of the critical region of the Fbietkan statistic. *Comm. Statist. Theory Methods* **9** 571–595.

ISHWARAN, H. and KOGALUR, U. B. (2019). Random forests for survival, regression, and classification (RF-SRC). R package version 2.8.0, available at https://cran.r-project.org/package=randomForestSRC.

ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* **2** 841–860. MR2516796

JAEGER, B. (2018). obliqueRSF: Oblique random forests for right-censored time-to-event data. R package version 0.1.0, available at https://CRAN.R-project.org/package=obliqueRSF.

JAEGER, B. C., LONG, L. D., LONG, D. M., SIMS, M., SZYCHOWSKI, J. M., MIN, Y.-I., MCCLURE, L. A., HOWARD, G. and SIMON, N. (2019). Supplement to "Oblique random survival forests." DOI:10.1214/19-AOAS1261SUPP.

KOWARIK, A. and TEMPL, M. (2016). Imputation with the R package VIM. *J. Stat. Softw.* **74** 1–16.

LEVEY, A. S., STEVENS, L. A., SCHMID, C. H., ZHANG, Y. L., CASTRO, A. F., FELDMAN, H. I., KUSEK, J. W., EGGERS, P., VAN LENTE, F. et al. (2009). A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* **150** 604–612.

LUNDBERG, S. M., ERION, G. G. and LEE, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

MCCALL, M. N., BOLSTAD, B. M. and IRIZARRY, R. A. (2010). Frozen robust multiarray analysis (fRMA). *Biostatistics* **11** 242–253.

MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17** Paper No. 26, 41. MR3491120

MENZE, B. H., KELM, B. M., SPLITTHOFF, D. N., KOETHE, U. and HAMPRECHT, F. A. (2011). On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 453–469. Springer, Berlin.

MOGENSEN, U. B., ISHWARAN, H. and GERDS, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *J. Stat. Softw.* **50** 1.

MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38** 2074–2102. MR3937487

NASEJJE, J. B., MWAMBI, H., DHEDA, K. and LESOSKY, M. (2017). A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Med. Res. Methodol.* **17** 115.

RAINFORTH, T. and WOOD, F. (2015). Canonical correlation forests. arXiv preprint arXiv:1507.05444.

SAFFORD, M. M., BROWN, T. M., MUNTNER, P. M., DURANT, R. W., GLASSER, S., HALANYCH, J. H., SHIKANY, J. M., PRINEAS, R. J., SAMDARSHI, T. et al. (2012). Association of race and sex with risk of incident acute coronary heart disease events. *JAMA* **308** 1768–1774.

SCHUMACHER, M., BASTERT, G., BOJAR, H., HUEBNER, K., OLSCHEWSKI, M., SAUERBREI, W., SCHMOOR, C., BEYERLE, C., NEUMANN, R. et al. (1994). Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *J. Clin. Oncol.* **12** 2086–2093.

SEGAL, M. R. (1988). Regression trees for censored data. *Biometrics* **44** 35–47.

SHABALIN, A. A., TJELMELAND, H., FAN, C., PEROU, C. M. and NOBEL, A. B. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* **24** 1154–1160.

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39** 1–13.

STRASSER, H. and WEBER, C. (1999). The asymptotic theory of permutation statistics. *Math. Methods Statist.* **8** 220–250. Johann Pfanzagl—on the occasion of his 70th birthday. MR1722622

STROBL, C., MALLEY, J. and TUTZ, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* **14** 323–348.

STROBL, C., BOULESTEIX, A.-L., ZEILEIS, A. and HOTHORN, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **8** 25.

TAYLOR JR., H. A., WILSON, J. G., JONES, D. W., SARPONG, D. F., SRINIVASAN, A., GARRISON, R. J., NELSON, C. and WYATT, S. B. (2005). Toward resolution of cardiovascular health disparities in African americans: Design and methods of the Jackson heart study. *Ethn. Dis.* **15** S6–4.

TERNÈS, N., ROTOLO, F., HEINZE, G. and MICHIELS, S. (2017). Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. *Biom. J.* **59** 685–701. MR3672690

THERNEAU, T. M. (2015). A package for survival analysis in S. R package version 2.38, available at https://CRAN.R-project.org/package=survival.

TUTZ, G. and BINDER, H. (2007). Boosting ridge regression. *Comput. Statist. Data Anal.* **51** 6044–6059. MR2407697

VAN HOUWELINGEN, H. C., BRUINSMA, T., HART, A. A. M., VAN'T VEER, L. J. and WESSELS, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Stat. Med.* **25** 3201–3216. MR2252292

VAN'T VEER, L. J., DAI, H., VAN DE VIJVER, M. J., HE, Y. D., HART, A. A., MAO, M., PETERSE, H. L., VAN DER KOOY, K., MARTON, M. J. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415** 530.

VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. Springer, New York.

WHELTON, P. K., CAREY, R. M., ARONOW, W. S., CASEY, D. E., COLLINS, K. J., HIMMELFARB, C. D., DEPALMA, S. M., GIDDING, S., JAMERSON, K. A. et al. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A report of the American college of cardiology/American heart association task force on clinical practice guidelines. *J. Am. Coll. Cardiol.* **71** e127–e248.

ZHU, R. (2013). *Tree-Based Methods for Survival Analysis and High-Dimensional Data*. Thesis (Ph.D.)–Univ. North Carolina at Chapel Hill. ProQuest LLC, Ann Arbor, MI. MR3211552

ZHU, R., ZENG, D. and KOSOROK, M. R. (2015). Reinforcement learning trees. *J. Amer. Statist. Assoc.* **110** 1770–1784. MR3449072

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

B. C. JAEGER
D. L. LONG
D. M. LONG
J. M. SZYCHOWSKI
G. HOWARD
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF ALABAMA AT BIRMINGHAM
327K RYALS PUBLIC HEALTH BUILDING
1665 UNIVERSITY BLVD
BIRMINGHAM, ALABAMA 35294-0022
USA
E-MAIL: bcjaeger@uab.edu
leannl@uab.edu
dmlong@uab.edu
jszychow@uab.edu
ghoward@uab.edu

M. SIMS
Y.-I. MIN
DEPARTMENT OF MEDICINE
UNIVERSITY OF MISSISSIPPI MEDICAL CENTER
JACKSON, MISSISSIPPI 39216
USA
E-MAIL: msims2@umc.edu
ymin@umc.edu

L. A. MCCLURE
DORNSIFE SCHOOL OF PUBLIC HEALTH
DREXEL UNIVERSITY
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: lam439@drexel.edu

N. SIMON
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195
USA
E-MAIL: nrsimon@u.washington.edu