Biostatistics (2015), **16**, 3, *pp.* 537–549
doi:10.1093/biostatistics/kxv001
Advance Access publication on February 5, 2015

Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models

ANDREW WEY*, JOHN CONNETT, KYLE RUDSER

University of Hawaii, Honolulu, HI 96815, USA
University of Minnesota, Minneapolis, MN 55455, USA
awey@hawaii.edu

Summary

For estimating conditional survival functions, non-parametric estimators can be preferred to parametric and semi-parametric estimators due to relaxed assumptions that enable robust estimation. Yet, even when misspecified, parametric and semi-parametric estimators can possess better operating characteristics in small sample sizes due to smaller variance than non-parametric estimators. Fundamentally, this is a bias–variance trade-off situation in that the sample size is not large enough to take advantage of the low bias of non-parametric estimation. Stacked survival models estimate an optimally weighted combination of models that can span parametric, semi-parametric, and non-parametric models by minimizing prediction error. An extensive simulation study demonstrates that stacked survival models consistently perform well across a wide range of scenarios by adaptively balancing the strengths and weaknesses of individual candidate survival models. In addition, stacked survival models perform as well as or better than the model selected through cross-validation. Finally, stacked survival models are applied to a well-known German breast cancer study.

Keywords: Bias–variance trade-off; Brier score; Cross-validation; Stacked regressions; Survival ensembles; Survival prediction.

1. Introduction

Survival function estimation has long been a major component of survival analysis. Yet estimation of conditional survival functions, i.e., survival functions that depend on covariate values, remains a challenging problem. A common semi-parametric approach combines the Cox proportional hazard model with a baseline hazard estimate, e.g., see Kalbfleisch and Prentice (2002). However, if the functional form is misspecified or the proportional hazards assumption is violated, then this approach may perform poorly. In terms of the bias–variance trade-off, the Cox model, and other parametric models, achieve low variance by making distributional and functional form assumptions. If the assumptions are approximately

*To whom correspondence should be addressed.

correct, then the bias term is small and the parametric and semi-parametric models perform well. On the other hand, if the assumptions are badly violated, then the bias term can be large and the models perform poorly.

Many non-parametric methods have been proposed to overcome the bias induced by violated assumptions. For example, Kooperberg *et al.* (1995) propose a flexible spline approach for the log-hazard that encompasses more than a proportional hazards model. Alternatively, tree-based approaches have been considered by several authors (Ishwaran *et al.*, 2008; Bou-Hamad *et al.*, 2011; Zhu and Kosorok, 2012). Despite possessing low bias in a wide variety of situations, non-parametric estimators suffer from high variance and can require a large sample size to perform well. This can lead to surprising situations where misspecified parametric models perform better than non-parametric estimators. Specifically, the effect of bias of misspecified parametric models is smaller than the effect of variance of non-parametric estimators, i.e., the bias–variance trade-off.

This article pursues a flexible estimator of a conditional survival function, i.e., an estimator that performs well when parametric assumptions are approximately correct while also maintaining robustness when parametric assumptions are violated. Traditionally, a single conditional survival function estimator is chosen from a set of candidate models, e.g., using an information criterion (Kooperberg *et al.*, 1995) or through cross-validation. Rather than select a single survival model, our goal is to estimate an optimally weighted combination of several survival models.

A variety of approaches that combine several models, often referred to as ensembles, have been explored in the uncensored setting. One approach, called "stacking," determines the optimally weighted average of several models by minimizing predicted error. Wolpert (1992) introduced stacking in the context of neural networks, while Breiman (1996) extended the idea to uncensored regression models and showed that stacking could improve prediction error. In particular, Breiman (1996) found that combining fundamentally different regression models, e.g., ridge regression and subset regression, had the largest reduction in prediction error. LeBlanc and Tibshirani (1996) found stacking with a constraint of non-negative weights to be an efficient way to combine models. Van der Laan *et al.* (2007) independently developed uncensored stacking as the 'Super Learner' algorithm, and presented results regarding the stacked estimator's rate of convergence. More recently, Boonstra *et al.* (2013) used stacking to improve prediction when incorporating different generation sequencing information in high-dimensional genome analysis.

Stacking models in a censored data setting presents additional challenges. Polley and Van der Laan (2011) mention stacking within a general censored data framework and provide an example for hazard function estimation with discrete survival times. This paper differs in two notable ways. First, we focus on estimating conditional survival functions with continuous survival times rather than a hazard function. This requires a different loss function that is tailored directly to estimating survival functions. We are particularly interested in the conditional survival function due to its role in many survival analysis methods; see the last paragraph of Section 7 for several examples. We also pursue the potential advantages of stacking parametric, semi-parametric, and non-parametric estimators. In particular, we show that stacked survival models perform well by giving weight to approximately correct parametric models, while shifting weight to non-parametric estimators when assumptions are violated. This allows stacked survival models to outperform the single model selected via cross-validation and, in some situations, outperform every individual model considered in the stacking procedure. We believe that combining parametric, semi-parametric, and non-parametric estimators is the biggest advantage of stacked survival models.

The remainder of the manuscript is organized as follows: stacked survival models are proposed in Section 2. Section 3 investigates the mean-squared error of stacked survival models with some asymptotic properties presented in Section 4. Section 5 investigates the finite sample performance through an extensive simulation study. Stacked survival models are then applied to the German breast cancer study data set in Section 6, with concluding remarks presented in Section 7.

2. Stacking survival models

Throughout the paper, random variables and observed variables are distinguished by capital and lower case letters, respectively. Our objective is to estimate the survival function of the event time random variable T that depends on p baseline covariates $\boldsymbol{x}$, i.e., $S_o(t|\boldsymbol{x}) = P(T > t|\boldsymbol{x})$. In survival analysis, T may be only partially observed due to a censoring random variable C that may also depend on $\boldsymbol{x}$. Define the conditional survival function of the censoring distribution as $G(t|\boldsymbol{x}) = P(C > t|\boldsymbol{x})$. We assume throughout that the event time and censoring random variables are conditionally independent, i.e., $T \perp C|\boldsymbol{x}$. The observed time is $y_i = \min(t_i, c_i)$, and $\delta_i = I(t_i < c_i)$ indicates whether an event was observed. Hence, a sample of right censored survival data of size n consists of triplets $\{y_i, \delta_i, \boldsymbol{x}_i\}$, $i = 1, \ldots, n$. Using the observed triplets, we can construct, for example, an estimate of the event time survival function from each of m candidate models with the kth estimate denoted as $\hat{S}_k(t|\boldsymbol{x})$.

To combine several predictors, we need a loss function that is tailored to survival functions. Our approach uses the Brier Score [BS(t)], which measures the squared error of a survival function at a given time point. In the absence of censoring, BS(t) is defined as

$$\text{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}\{Z_i(t) - \hat{S}(t|\boldsymbol{x}_i)\}^2, \tag{2.1}$$

where $Z_i(t) = I(t_i > t)$. Note that t denotes a chosen time point, and t_i (with the subscript) denotes the event time for the ith observation and may not be observed due to censoring.

Unfortunately, right-censoring implies that equation (2.1) is only partially observed; that is, $Z_i(t)$ is undefined for censored observations when $y_i > t$. To correct this issue, we use inverse probability-of-censoring weights (IPCW) to account for the probability of an observation being censored (Lostritto *et al.*, 2012). In particular, the 'inverse probability-of-censoring-weighted Brier Score' at time t [IPCW-BS(t)] can be written as

$$\text{IPCW-BS}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i(t)}{G(T_i(t)|\boldsymbol{x}_i)} \times \{Z_i(t) - \hat{S}(t|\boldsymbol{x}_i)\}^2, \tag{2.2}$$

where $T_i(t) = \min\{t_i, t\}$, $\Delta_i(t) = I(\min\{t_i, t\} < c_i)$, and $G(\cdot|\boldsymbol{x}_i)$ are the conditional survival function of the censoring distribution, which is estimated by a marginal Kaplan–Meier throughout the rest of the paper. From a technical point of view, the IPCWs ensure that the expectations of equations (2.1) and (2.2) are the same (assuming that the estimator of the censoring distribution is uniformly consistent). Thus, the true conditional survival function, $S_o(t|\boldsymbol{x})$, is the minimizer of $E\{\text{IPCW-BS}(t)\}$ (see supplementary material available at *Biostatistics* online).

There are several points that are helpful to note for understanding the calculation of the IPCW-BS(t). The values of $T_i(t)$ and $\Delta_i(t)$ depend on t and the censoring status of the ith observation. For each uncensored observation, the value of $G(T_i(t)|\boldsymbol{x}_i)$, and therefore the calculation of the weight, depends on whether the event has occurred by time t. For censored observations, there are two possible situations at a given time point t:

- If $c_i > t$, then $T_i(t) = t$ and $\Delta_i(t) = I(t < c_i) = 1$.
- If $c_i < t$, then $\Delta_i(t) = 0$ (since $c_i < t_i$ for censored observations) and hence $\Delta_i(t)/G(T_i(t)|\boldsymbol{x}_i) = 0$.

Thus, for a fixed time t, censored observations with $c_i > t$ will contribute to the Brier Score, while censored observations with $c_i < t$ will still contribute to the Brier Score but only indirectly through the estimation of the censoring distribution.

Since the goal is to estimate the entire conditional survival function, the Brier Score is minimized over a set of time points, say $t_1, \ldots, t_s$. This implies the following weighted least squares problem with the additional constraints that $\sum_{k=1}^{m} \hat{\alpha}_k = 1$, which is required for the theoretical results, and $\hat{\alpha}_k \geqslant 0$ for all $k = 1, \ldots, m$, which has been shown to improve performance in the uncensored setting (Breiman, 1996; LeBlanc and Tibshirani, 1996),

$$\hat{\boldsymbol{\alpha}} = \arg \min_{\boldsymbol{\alpha}, \alpha_k \geqslant 0} \sum_{r=1}^{s} \sum_{i=1}^{n} \frac{\Delta_i(t_r)}{\hat{G}(T_i(t_r)|\boldsymbol{x}_i)} \times \{Z_i(t_r) - \sum_{k=1}^{m} \alpha_k \hat{S}_k^{(-i)}(t_r|\boldsymbol{x}_i)\}^2, \tag{2.3}$$

where $\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)$ is the survival estimate from the kth model while leaving the ith observation out during the fitting process. This ensures that stacking does not reward model complexity (i.e., does not overfit the data). To reduce computational demands, we use 5-fold cross-validation rather than n-fold cross-validation to obtain $\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)$. In particular, the data are randomly split into five roughly equally sized sets and $\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)$ is obtained for observations in a given set by fitting the candidate survival models to the observations in the other four sets. As such, five survival models, rather than n survival models, are fit for each of the m candidate survival models.

Finally, the stacked estimate of the conditional survival function with time-independent weights is

$$\hat{S}(t|\boldsymbol{x}) = \sum_{k=1}^{m} \hat{\alpha}_k \hat{S}_k(t|\boldsymbol{x}), \tag{2.4}$$

where $\hat{S}_k(t|\boldsymbol{x})$ is the kth survival model estimated with all the data.

REMARK 1 The Brier Score measures agreement at only one particular time. As such, the value(s) of t over which it is evaluated, i.e., $t_1, \ldots, t_s$, have implications for performance. In particular, care should be taken to avoid picking only very small, or very large t values, though one could also consider unequal weighting or restricting to certain areas of support. We find that nine evenly spaced quantiles of the observed event distribution works well (see supplementary material available at *Biostatistics* online).

REMARK 2 Time-dependent stacking, i.e., allowing the weighted combination of models to depend on time, was also considered (see supplementary material available at *Biostatistics* online). Though potentially adding flexibility, a major flaw of time-dependent stacking is that the conditional survival function may, at times, increase, which violates the non-increasing property of survival functions. As such, this paper focuses on time-independent stacking.

3. MEAN-SQUARED ERROR DECOMPOSITION

We analyze the mean-squared error of the stacked survival model. We start by defining the mean-squared error for the stacked estimator as $\text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\} = E \int_0^\tau [\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]^2 \, \mathrm{d}t$, where the expectation is over the random variable for the covariate space and the sampling distribution of the stacked estimator. This definition of mean-squared error is motivated, in part, by the Brier Score. In particular, supplementary material available at *Biostatistics* online shows that $E \int_0^\tau \text{IPCW-BS}(t) \, \mathrm{d}t = \sigma^2 + \text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\}$, where σ^2

is irreducible prediction error. Similar to the analysis of Fumera and Roli (2005), we show in supplementary material available at *Biostatistics* online that the mean-squared error decomposes into

$$\begin{aligned}\mathrm{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\} = \sum_{k=1}^{m} \alpha_k^2 \,\mathrm{MSE}_\tau\{\hat{S}_k(\cdot|\boldsymbol{x})\} + E\sum_{k=1}^{m}\sum_{l\neq k} \alpha_k\alpha_l \int_0^\tau [\mathrm{Bias}\{\hat{S}_k(t|\boldsymbol{x})\} \times \mathrm{Bias}\{\hat{S}_l(t|\boldsymbol{x})\} \\ + \mathrm{Corr}\{\hat{S}_k(t|\boldsymbol{x}), \hat{S}_l(t|\boldsymbol{x})\} \times \mathrm{Var}\{\hat{S}_k(t|\boldsymbol{x})\}^{1/2} \times \mathrm{Var}\{\hat{S}_l(t|\boldsymbol{x})\}^{1/2}]\,\mathrm{d}t,\end{aligned}$$

where $\mathrm{MSE}\{\hat{S}_k(\cdot|\boldsymbol{x})\}$, $\mathrm{Bias}\{\hat{S}_k(t|\boldsymbol{x})\} = E[\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]$, and $\mathrm{Var}\{\hat{S}_k(t|\boldsymbol{x})\} = E\{[\hat{S}_k(t|\boldsymbol{x}) - E\,\hat{S}_k(t|\boldsymbol{x})]^2\}$ are, respectively, the mean-squared error, bias at time t, and variance at time t for the kth survival model in the stacking procedure, while $\mathrm{Corr}\{\hat{S}_k(t|\boldsymbol{x}), \hat{S}_l(t|\boldsymbol{x})\}$ is the correlation at time t between the kth and lth survival model.

The mean-squared error of the stacked estimator decomposes into two parts: a weighted combination of the mean-squared error of candidate survival models and the interaction between candidate survival models in terms of bias and correlation. The decomposition makes it easy to show, given a set of candidate survival models, that there exists a set of stacking weights such that the stacked estimator possess as good, or better, mean-squared error as the best performing model in the set of candidate survival models. However, this property is *not* guaranteed after estimating the stacking weights. Thus, careful selection of candidate survival models is warranted.

The MSE decomposition provides insight into how features of candidate survival models impact performance of the stacked estimator. As stated in Section 1, the motivation for stacked survival models is to obtain robustness across a wide variety of scenarios by including models from different classes, i.e., parametric, semi-parametric, and non-parametric models, and with different assumptions (e.g., proportional hazards or accelerated failure time). In this fashion, the stacked estimator may assign more weight to one model in the stack for one scenario and shift to another model, e.g., one based on different assumptions, for a different scenario. This motivates including a set of models that "represent" a variety of classes and types of models, i.e., ensuring a diverse set of candidate survival models (Breiman, 1996). This is also supported by the MSE decomposition as the correlation between diverse models will tend to be lower due to different assumptions.

The next consideration is the number of models of a given type to include in the stack, e.g., the number of Cox proportional hazards models to include. Due to numerous options regarding potential covariates and the functional form of those covariates, e.g., linear terms versus quadratic terms, there are many different Cox models that could be included. However, models with the same distributional assumptions and similar sets of covariates are expected to have similar MSE with a rather high between-model correlation. Since there is no guarantee that only one model among a set of highly correlated models will receive non-zero weight, the MSE decomposition suggests that the stack will perform better by excluding models with small differences in the set of considered covariates. Further discussion, illustrative examples, and simulations are included in supplementary material available at *Biostatistics* online.

4. Asymptotic properties

We show model selection and uniform consistency for the stacked estimate of the conditional survival function. The former refers to the idea that if the set of stacked models contains uniformly consistent models, then all weight is asymptotically given to those models in the stack. Consistent model selection implies uniform consistency as long as there is at least one uniformly consistent estimator of the conditional survival function. Our main assumption is that there exists no weighted average of misspecified models

that approaches the true survival function for every time point included in equation (2.3). Supplementary material available at *Biostatistics* online contains all of the assumptions and proofs.

Let $\Omega = (0, \tau)$ be the support of interest for estimating the conditional survival function, and consider m estimators for the stacking procedure. Then

THEOREM 4.1 Let $\hat{\boldsymbol{\alpha}}$ be estimated by equation (2.3). Assume that models $1, \ldots, l$, where $l < m$, are the only uniformly consistent estimators and conditions (A1)–(A3) in supplementary material available at *Biostatistics* online hold, then $\sum_{k=1}^{l} \hat{\alpha}_k \to 1$, in probability, as $n \to \infty$.

This ensures that uniformly consistent model(s) will asymptotically receive all of the weight for the stacked conditional survival function estimate in equation (2.4). There can be more than one uniformly consistent estimator, e.g., a correctly specified Weibull model and Cox model. In the special case, when only one model is uniformly consistent, we obtain the corollary:

COROLLARY 4.2 If $\hat{S}_1(t|\boldsymbol{x})$ is the only uniformly consistent estimator, then $\hat{\alpha}_1 \to 1$, in probability, as $n \to \infty$.

The result of Theorem 4.1 and Corollary 4.2 is required for uniform consistency of the stacked estimator with time-independent weights.

THEOREM 4.3 Let the stacked estimate of the conditional survival function be defined as $\hat{S}(t|\boldsymbol{x})$ in equation (2.4). Assume that conditions (A1)–(A3) in supplementary material available at *Biostatistics* online hold then, as $n \to \infty$,

$$\sup_{t \in \Omega} \sup_{\boldsymbol{x}} \left| \sum_{k=1}^{m} \hat{\alpha}_k \hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x}) \right| \to 0.$$

The rate of convergence of the stacked estimator is not addressed here. However, Van der Laan *et al.* (2007) showed that, in the uncensored case, the stacked estimator's risk converged at either the best rate of a correctly specified model, or slightly slower than the parametric rate. These results are not directly applicable since the Brier Score does not measure the risk of the entire conditional survival function. In addition, distributional results for the conditional survival function are complicated by the constrained estimation of $\boldsymbol{\alpha}$ (see supplementary material available at *Biostatistics* online for in-depth discussion).

5. SIMULATIONS

An extensive simulation study examines the finite sample performance of stacked survival models. In particular, two settings are investigated: a moderate number of covariates (Section 5.1) and a large number of covariates (Section 5.2).

The simulations are comprised of combinations of an event distribution ($d = 1, 2, 3$) and linear form of covariates ($q = 1, 2$). The covariate distributions are multivariate normal: $\boldsymbol{x}_p \sim \text{MVN}(\mathbf{0}, \Sigma)$, where Σ is the correlation matrix and for all $i, j = 1, \ldots, p$, $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.4$ (p is the vector dimension). Section 5.1 has an eight-dimensional covariate space (i.e., $p = 8$), while Section 5.2 has a $p = 80$ dimensional covariate space. For Section 5.1, the covariate effects are $\boldsymbol{\beta} = (1, 0, -1, 0, 0.5, 0, -0.5, 0)$, while for Section 5.2 the first 12 covariate effects are $(1, 0, -1, 0, 0.5, 0, -0.5, 0, 0.25, 0, -0.25, 0)$ with the other 68 effects set to zero. Two different linear combinations are considered: $\boldsymbol{\gamma}^1 = \boldsymbol{x}_p$ and $\boldsymbol{\gamma}^2 = \Phi(4 \times \boldsymbol{x}_p)$

which imply linear and non-linear covariate effects, respectively. The event distributions are defined as

1. $T_1^{(q)} \sim \exp\left\{\text{Normal}\left(\boldsymbol{\beta}\boldsymbol{\gamma}^q, \frac{1}{4}\right)\right\}$
2. $T_2^{(q)} \sim \text{Weibull}(\text{scale} = \exp\{\boldsymbol{\beta}\boldsymbol{\gamma}^q\},\ \text{shape} = 1.1)$
3. $T_3^{(q)} \sim \text{Gamma}(\text{scale} = \frac{1}{4}\exp\{\boldsymbol{\beta}\boldsymbol{\gamma}^q\},\ \text{shape} = 5)$

Each subsection investigates every combination of the event distribution (d) and linear form (q), i.e., there are six scenarios for both Sections 5.1 and 5.2.

We compare the performance of survival models based on an approximation to the mean squared error presented in Section 3, which we call integrated squared survival error (ISSE):

$$\text{ISSE} \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{19}(\hat{S}(t_j|\boldsymbol{x}_i) - S_o(t_j|\boldsymbol{x}_i))^2,$$

where t_j are a fixed set of 19 equally spaced quantiles of the survival time distribution given that an event occurs, and $S_o(\cdot|\boldsymbol{x}_i)$ is the true conditional survival function.

One comparison of interest for the stacked estimator is the model chosen through cross-validation. We use the integrated Brier Score (IBS) as the measure of the predicted error for selecting the individual model. In particular, the IBS for the kth model is defined as $\hat{\text{IBS}}_k = \int_0^\tau \widehat{\text{BS}}_k(t)\,\text{d}t$, where τ is the maximum observed time and $\widehat{\text{BS}}_k(t)$ is the estimated Brier Score at time t for the kth model (with an out-of-bag estimate of the conditional survival function). The cross-validated estimator is then defined as $\hat{S}_l(\cdot|\boldsymbol{x})$, where $l = \arg\min_k \hat{\text{IBS}}_k$.

All simulations were run in R version 3.0.0 (R Development Core Team, 2013). The constrained minimization problem was solved using the `alabama` package (Varadhan, 2012). The stacking weights, i.e., equation (2.3), were estimated by minimizing the Brier Score over the $0.1, 0.2, \ldots, 0.9$ quantiles of the observed event distribution.

5.1 *Modest-dimensional covariate space*

This setting has relatively few covariates ($p = 8$) with a modest censoring rate (25%) and sample size ($n = 200$). This illustrates stacked survival models in a relatively straightforward scenario.

The stacked survival models include a Weibull model and log-Normal model as parametric models, a Cox proportional hazards model with an Efron estimate of the baseline cumulative hazard function as a semi-parametric model, and random survival forests (RSFs) as a non-parametric model. The parametric and semi-parametric models include only first-order main effects and no interactions. All of the parametric and semi-parametric models are estimated using the `survival` package in R (Therneau, 2013), and all of the parametric and semi-parametric models use 5-fold cross-validation to estimate $\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)$. RSF is estimated with the `randomSurvivalForest` package in R (Ishwaran and Kogalur, 2013). The RSF is an ensemble of 250 trees grown using package defaults. For RSF, $\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)$ is estimated with the out-of-bag ensemble from the `rsf` function. The censoring distribution is a uniform distribution for all $T_d^{(q)}$: $C_{d,q} \sim \text{Unif}(0, c(d,q))$, where $c(d,q)$ is a constant that depends on (d,q) and ensures ∼25% censoring.

The log-Normal and Weibull scenarios with linear covariate effects illustrate performance when there is a correctly specified parametric or semi-parametric model in the stack. Stacking is not expected to perform better than a correctly specified parametric model, but should still perform relatively well in such situations. The Gamma scenario with linear covariate effects illustrates performance when there are approximately correct parametric models in the stack (e.g., a correct mean function). The scenarios with

Table 1. *Simulation results for Section* 5.1 ($n = 200$, $p = 8$ *covariates, and* 25% *censoring*) *presented in ISSE over the observed support*

		Models	Log-Normal	Weibull	Gamma
Linear effects	Single models	Log-Normal	**0.35**	0.82	**0.34**
		Weibull	0.61	**0.53**	0.41
		Cox	0.86	0.68	0.69
		RSF	7.26	4.88	7.36
	Flexible models	Stacking	**0.42**	**0.58**	**0.37**
		CV	0.72	0.70	0.53
Non-linear effects	Single models	Log-Normal	4.71	2.54	5.03
		Weibull	5.17	**2.27**	5.30
		Cox	5.15	2.33	5.33
		RSF	**4.29**	3.49	**4.46**
	Flexible models	Stacking	**3.49**	**2.08**	**3.69**
		CV	5.00	2.48	5.18

Each simulation is replicated 2000 times, and the error is multiplied by 10. The two top estimators are bolded for each simulation scenario. 'RSFs' stand for RSFs, 'Stacking' is stacked survival models, and 'CV' is the estimator selected through cross-validation. The standard error for the estimate ISSE for each method in each scenario is <0.01.

non-linear covariate effects were designed to have badly misspecified parametric and semi-parametric models. Due to the lack of a correctly specified parametric model, stacked survival models should perform relatively well by, in particular, assigning more weight to the non-parametric estimator: RSFs.

Table 1 presents the results in terms of ISSE. Since the goal is an estimator that performs well in a wide variety of situations, the top two estimators are bolded for each scenario. The stacked survival model, i.e., "Stacking", is a top two estimator for all six scenarios. For the scenarios with non-linear covariate effects, the stacked estimator reduces the ISSE by ∼8–15% compared with the best single model. In addition, the stacking procedure outperforms selecting a single model via cross-validation in every situation.

As an illustration, Table 2 presents the average stacking weights for the individual models. For the linear scenarios, the stacking procedure gives a majority of weight to correctly specified parametric models. The weights are more interesting for the scenarios with non-linear covariate effects. In particular, the parametric models always receive over 40% of the weight despite, at times, having 10% higher ISSE than RSF. This is a good example of stacked survival models combining misspecified parametric models and an inefficient non-parametric model to obtain an estimator that outperforms every single model considered in the stacking procedure.

REMARK 3 RSFs possess tuning parameters that influence performance, e.g., the minimum number of events in a node. While the performance of RSF could be improved by adaptively selecting tuning parameters (e.g., by cross-validation), stacked survival models are likely to also inherit any improvement in RSF since it is included in the stack.

5.2 *Large covariate space*

This setting has a large number of covariates ($p = 80$) relative to the sample size ($n = 200$). The censoring distributions are the same as Section 5.1. In general, the parametric and semi-parametric models used (and stacked) in Section 5.1 will not perform well in large covariate spaces without regularization. As such, they are not included for these scenarios. We instead stack a Cox model with an l_1 penalty (i.e., lasso), a boosted

Table 2. *Average weights for the individual models included in the stacked survival model for each of the six scenarios in Section* 5.1 ($n = 200$, $p = 8$ *covariates*, *and* 25% *censoring*)

	Stacked models	Log-Normal	Weibull	Gamma
Linear effects	Log-Normal	0.61	0.19	0.42
	Weibull	0.23	0.45	0.37
	Cox	0.14	0.31	0.19
	RSF	0.02	0.05	0.02
Non-linear effects	Log-Normal	0.34	0.12	0.27
	Weibull	0.14	0.28	0.18
	Cox	0.06	0.21	0.08
	RSF	0.46	0.39	0.47

Each simulation is replicated 2000 times. 'RSFs' stand for random survival forests.

Table 3. *Simulation results for Section* 5.2 ($n = 200$, $p = 80$ *covariates*, *and* 25% *censoring*) *presented in integrated squared survival error* (*ISSE*) *over the observed support*

		Models	Log-Normal	Weibull	Gamma
Linear effects	Single models	Cox-Lasso	**2.43**	**1.68**	**2.50**
		Cox-boosting	2.60	1.75	2.66
		RSF	2.46	1.86	**2.50**
	Flexible models	Stacking	**2.43**	**1.68**	**2.50**
		CV	**2.43**	1.69	**2.50**
Non-linear effects	Single models	Cox-Lasso	2.02	1.03	2.08
		Cox-boosting	2.01	**1.01**	2.08
		RSF	**1.89**	1.15	**1.95**
	Flexible models	Stacking	**1.97**	**1.00**	**2.04**
		CV	2.01	1.03	2.07

Each simulation is replicated 2000 times, and the error is multiplied by 1. The two top estimators are bolded for each simulation scenario. 'RSFs' stand for random survival forests, 'Stacking' is stacked survival models, and 'CV' is the estimator selected through cross-validation. The standard error for the estimate ISSE for each method in each scenario is <0.005.

version of the Cox model, and RSFs. The l_1 penalized version of the Cox model is fit using the R package `penalized` with the penalty parameter chosen via cross-validation (Goeman, 2012). The boosted Cox model is fit using the package `CoxBoost` in R with default tuning parameters (Binder, 2013). RSF is fit in the same manner as Section 5.1.

The stacked survival model is again a top two estimator in every scenario (see Table 3). Relative to Section 5.1, stacked survival models offer smaller improvements (e.g., $\sim$1–5% lower ISSE compared with the cross-validated estimator). However, the improvements in ISSE remain consistent across the scenarios. In addition, the stacking procedure still performs as good, or better, in every scenario as the model selected via cross-validation.

REMARK 4 Supplementary material available at *Biostatistics* online presents numerous extensions to the simulation study. In particular, we investigate scenarios with a larger sample size, a high censoring rate, non-monotonic covariate effects, and a misspecified censoring model. In addition, we comment on the required computational time and the influence of the out-of-bag estimator for RSF.

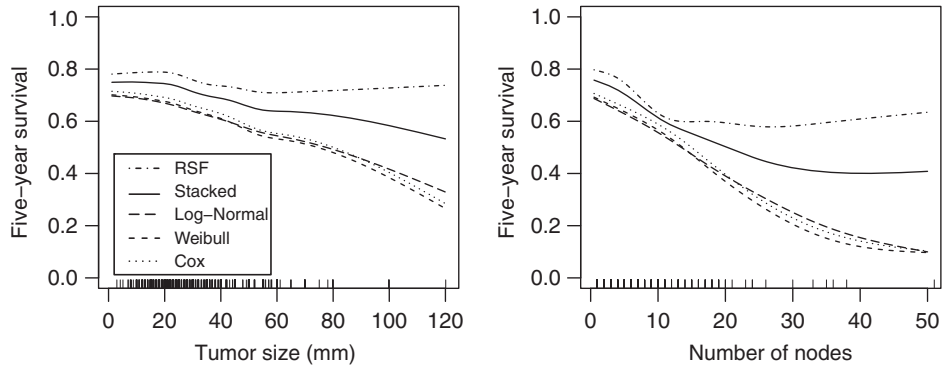

Fig. 1. The association of tumor size (mm) and the number of nodes with 5-year survival for the GBCS data set with the other covariates to their median value. The tick marks at the bottom of the plots indicate the skewness of both covariates.

6. GERMAN BREAST CANCER STUDY

Stacked survival models are illustrated on a well-known survival benchmark data set: the German breast cancer study (GBCS) described by Hosmer *et al.* (2008), and accessible at the University of Massachusetts website for statistical software information. There are eight covariates included in the analysis: age at diagnosis, tumor size, tumor grade, number of nodes, menopausal status, the number of progesterone receptors, the number of estrogen receptors, and hormone therapy status. The outcome of interest is the time till death, and there is complete data on 686 patients with ∼75% censoring. The stacking procedure uses the same models as Section 5.1. That is, the Weibull and log-Normal models are the parametric models, the Cox proportional hazards model is the semi-parametric model, and an RSF is the non-parametric model. The minimum number of deaths (for RSF) is set at 12, which was selected by minimizing predicted error among five potential values: 3, 6, 12, 24, and 48.

We are particularly interested in the association of tumor size and the number of nodes with 5-year survival. In order to evaluate the association, the stacked survival model and each model included in the stacking procedure predicts the 5-year survival rate for each patient in the study. After predicting 5-year survival, a generalized additive model with penalized B-splines for the continuous covariates (i.e., the `gam` function from the `mgcv` package; Wood, 2006) estimates the association of tumor size and the number of nodes with 5-year survival while adjusting for the other covariates.

Figure 1 presents the estimated 5-year survival as a function of tumor size and the number of nodes at the median of the other covariates. The parametric/semi-parametric models suggest worse 5-year survival with increasing tumor size and number of nodes. In contrast, RSF suggests that 5-year survival dips slightly ∼40 mm for tumor size, while 5-year survival for the number of nodes has a sharp early decrease but plateaus after ∼ 10 nodes. The stacked survival model— which gives weight to the Weibull model (0.06), the Cox model (0.38), and RSF (0.56)—is a compromise between the parametric/semi-parametric models and RSF.

The GBCS data set has a marginal 5-year survival rate of 70% due, in part, to a censoring rate of 75%. As such, predicted 5-year survival rates $<20\%$ are surprising (i.e., the parametric/semi-parametric models for the number of nodes). Due to the sparsity of patients with >20 nodes, the low model-based predicted probabilities are likely due to parametric/semi-parametric models being heavily influenced by a strong negative association with survival for patients with <20 nodes (98% of patients have <20 nodes) through the first-order linear effect (note that the patient with over 50 nodes was censored after 2 years). In contrast, RSF does not require any linearity assumptions and is more influenced by local observations

in predicting 5-year survival (Ishwaran *et al.*, 2008). From this perspective, the stacked survival model is balancing model-based predictions that require assumptions of linearity with locally based predictions.

REMARK 5 In this example, the model weights provide insight into how candidate survival models were combined to form the stacked estimate of the conditional survival function. However, we caution against interpreting model weights as an indication of a "correct model." As noted by a referee, this is particularly dangerous when two models possess similar survival functions due to potential instability in the minimization procedure.

7. CONCLUSION AND FUTURE DIRECTIONS

We propose stacking survival models to flexibly estimate conditional survival functions. Stacked survival models can combine several models, spanning the full range of parametric, semi-parametric, and non-parametric estimators. This allows stacking to exploit the low variance of approximately correct parametric models, while maintaining the robustness of non-parametric estimators. As illustrated in the simulation study, stacked survival models give more weight to parametric and semi-parametric models when assumptions are approximately correct, but shift weight to non-parametric estimators when assumptions are badly violated. In this manner, stacked survival models perform well across a wide range of scenarios. In particular, for a given scenario, stacked survival models were found to perform better than the single model chosen through cross-validation and, at times, perform better than any single model considered in the stacking procedure.

In practice, the true underlying data generation process is never known, i.e., one does not choose the true event distribution or functional form of the covariates. This motivates an adaptive approach that can perform well in a wide variety of situations. Cross-validation is currently the most common adaptive approach. Yet, the set of simulations illustrate that stacked survival models perform as good, or better, than the model selected through cross-validation, which picks a single model to receive all the weight (i.e., $\alpha_k = 1$ for some k). As such, stacked survival models warrant consideration whenever cross-validated models are used. Other predictive models could also have been considered, though stacked survival models could inherit any particular advantages of such models through inclusion in the stack.

As shown in Section 3, the MSE decomposition of the stacked survival model depends on the MSE for each candidate model, the pairwise correlation between candidate models, and the weight (i.e., α_k) given to each model. The correlation term suggests that including an additional model that is very similar, e.g., a model of the same type as one already in the stack but with small differences in the covariates included, may not improve and could harm performance (see supplementary material available at *Biostatistics* online). While the estimated weights can theoretically be zero, there is no guarantee that this will occur for a highly correlated model. This could motivate a preliminary screen procedure for candidate survival models that is based on pairwise correlations. For example, a procedure to determine the covariate combination and functional form for parametric or semi-parametric models, or determine values of tuning parameters for non-parametric estimators. However, as noted by an anonymous referee, any screening procedure needs to be careful to avoid overfitting by using, for example, nested cross-validation or independent, external data. The advantages of a potential screening procedure deserve further research.

Covariate-dependent stacking (or, allowing the α_k to depend upon $\boldsymbol{x}$) is a potential avenue for improving stacked survival models. LeBlanc and Tibshirani (1996) mention this approach for uncensored stacking, and a collaborative group using covariate-dependent stacking won the Netflix Prize competition to improve movie recommendations (Sill *et al.*, 2009). However, extending the stacking procedure to include covariate-dependent weights with the constraints introduced here is not straightforward. For example, Sill *et al.* (2009) do not constrain their covariate-dependent weights despite prior experiences suggesting that regularization improves performance (Breiman, 1996; LeBlanc and Tibshirani, 1996). Investigation

of covariate-dependent stacking and different approaches to constraining the covariate-dependent weights deserves further investigation.

The Brier Score, used to estimate the weighted combination of survival models, is essentially an inverse probability-of-censoring-weighted (IPCW) estimate of prediction error. The IPCW estimate requires estimating the (possibly conditional) censoring distribution. The simulation scenarios introduced in Section 5 use a Kaplan–Meier estimator for the censoring distribution that is correctly specified. In our experience, the stacking procedure maintains good operating characteristics when the censoring model is misspecified. However, if there is strong evidence of differential censoring among the covariates, then a conditional estimator may be warranted.

The importance of efficient, yet robust, estimators of conditional survival functions (or, equivalently, conditional distribution functions) continues to grow. Methods in a wide range of areas require estimating a conditional survival function as a nuisance parameter, for example, censored quantile regression (Wey *et al.*, 2014), time-dependent ROC curves (Zheng and Heagerty, 2004), inverse probability-of-censoring-weighted estimators, e.g., Fine and Gray (1999), model-free contrast approaches (Rudser *et al.*, 2012), and dynamic treatment regime methods (Zhao *et al.*, 2011). The simulations presented here suggest that stacking parametric, semi-parametric, and non-parametric models for the nuisance parameter will likely result in better estimation of regression parameters of interest, though these topics warrant further investigation.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://biostatistics.oxfordjournals.org

Conflict of Interest: None declared.

FUNDING

This work was supported by grants UL1TR000114 of the National Center for Advancing Translational Sciences, U54MD007584 of the National Institute on Minority Health and Health Disparities, and G12MD007601 of the National Institute on Minority Health and Health Disparities.

REFERENCES

BINDER, H. (2013). *CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks*. R package version 1.4. http://cran.r-project.org/web/packages/CoxBoost/index.html.

BOONSTRA, P. S., TAYLOR AND, J. M. G. MUKHERJEE, B. (2013). Incorporating auxiliary information for improved prediction in high-dimensional datasets: an ensemble of shrinkage approaches. *Biostatistics* **14**, 259–272.

BOU-HAMAD, I., LAROCQUE AND, D. BEN-AMEUR, H. (2011). A review of survival trees. *Statistics Surveys* **5**, 44–71.

BREIMAN, L. (1996). Stacked regressions. *Machine Learning* **24**, 49–64.

FINE AND, J. P. GRAY, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**, 496–509.

FUMERA AND, G. ROLI, F. (2005). A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, 942–956.

GOEMAN, J. J. (2012). *Penalized R Package*. R package version 0.9-42. http://cran.r-project.org/web/packages/penalized/index.html.

HOSMER, D. W., LEMESHOW AND, S. MAY, S. (2008) *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Hoboken, New Jersey: Wiley.

Ishwaran and, H. Kogalur, U. B. (2013). *Random Survival Forests*. R package version 3.6.4. http://cran.r-project.org/web/packages/randomSurvivalForest/index.html.

Ishwaran, H., Kogalur, U. B., Blackstone and, E. H. Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics* **2**, 841–860.

Kalbfleisch and, J. D. Prentice, R. L. (2002) *The Statistical Analysis of Failure Time Data*. Hoboken, New Jersey: Wiley.

Kooperberg, C., Stone and, C. J. Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association* **90**, 78–94.

LeBlanc and, M. Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association* **91**, 1641–1650.

Lostritto, K., Strawderman and, R. L. Molinaro, A. M. (2012). A partitioning deletion/substitution/addition algorithm for creating survival risk groups. *Biometrics* **68**, 1146-–1156.

Polley and, E. C. Van der Laan, M. (2011). Super learning for right-censored data. In: *Targeted Learning: Causal Inference for Observational and Experimental Data*. Mark van der Laan and Sherri Rose: Springer.

R Development Core Team. (2013). *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/.

Rudser, K. D., LeBlanc and, M. L. Emerson, S. S. (2012). Distribution-free inference on contrasts of arbitrary summary measures of survival. *Statistics in Medicine* **31**, 1722–1737.

Sill, J., Takacs, G., Mackey and, L. Lin, D. (2009). Feature-weighted linear stacking. *Arxiv*.

Therneau, T. (2013). *Survival Analysis, Including Penalized Likelihood*. R package v 2.37-4.

Van der Laan, M. J., Polley and, E. C. Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology* **6**, 00–00.

Varadhan, R. (2012). *alabama: Constrained Nonlinear Optimization*. R package v 2011.9-1. http://cran.r-project.org/web/packages/alabama/index.html.

Wey, A., Wang and, L. Rudser, K. (2014). Censored quantile regression with recursive partitioning based weights. *Biostatistics* **15**, 170–181.

Wolpert, D. H. (1992). Stacked generalization. *Neural Network* **5**, 241–259.

Wood, S. N. (2006) *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman and Hall.

Zhao, Y., Zeng, D., Socinski and, M. A. Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* **67**, 1422–1433.

Zheng and, Y. Heagerty, P. J. (2004). Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* **5**, 615–632.

Zhu and, R. Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association* **107**, 331–340.

[*Received May 18, 2014; revised December 19, 2014; accepted for publication January 5, 2015*]