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

ANDREW WEY*, JOHN CONNETT, KYLE RUDSER

Division of Biostatistics, School of Public Health, University of Minnesota

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 tradeoff 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. As such, stacking can exploit the low variance of approximately correct parametric models, while maintaining the robustness of non-parametric models. This is demonstrated through an extensive simulation study wherein stacked survival models consistently perform well across a wide range of scenarios. In addition, stacked survival models perform as good, or better, than the model selected through cross-validation. Finally, stacked survival models are applied to the well-known German breast

*To whom correspondence should be addressed.
A. Wey, J. Connett, and K. Rudser

cancer study.

Key words: Bias-variance tradeoff, Brier Score, Cross-validation, Survival Ensembles, Stacked Regressions.

1. Introduction

Survival function estimation has long been a major component of survival analysis (Kaplan and Meier, 1958). 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 tradeoff, the Cox model, and other parametric models, achieve low variance by making distributional and functional form assumptions. If the assumptions are approximately 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) proposes 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 bias of misspecified parametric models is smaller than the 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 is both efficient when parametric assumptions are approximately correct and robust 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 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 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 one example for hazard function estimation. This paper differs in two significant ways. First, we focus on estimating conditional survival functions rather than a hazard function, which requires a different loss function that is tailored to directly estimating survival functions. We also pursue the potential advantages of stacking parametric, semi-parametric, and non-parametric estimators. In particular, we show that stacked survival models gain efficiency by giving weight to approximately
correct parametric models, but remains robust by 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: uncensored stacking and the extension to censored data are introduced in Section 2. Some asymptotic properties of stacked survival models are discussed in Section 3. Section 4 investigates the finite sample performance of stacked survival models through an extensive simulation study. Stacked survival models are then applied to the German breast cancer study data set in Section 5, with concluding remarks are presented in Section 6.

2. Stacking

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 $x$, i.e., $S(t|x) = P(T > t|x)$. In survival analysis, $T$ may only be partially observed due to a censoring random variable $C$ that may also depend on $x$. Define the conditional survival function of the censoring distribution as $G(t|x) = P(C > t|x)$. We assume throughout that the event time and censoring random variables are conditionally independent, i.e., $T \perp C|x$. 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 $\{y_i, \delta_i, x_i\}, i = 1, ..., n$. 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 $k^{th}$ estimate denoted as $\hat{S}_k(t|x)$. 
2.1 Uncensored Stacking

Stacking requires predicting outcomes with each model, then finding the combination that minimizes predicted error. In the uncensored case, the event time \( t_i \) is observed for all \( i = 1, \ldots, n \), so the predicted values are \( \hat{t}_{i,k} \) for the \( k^{th} \) model \( (k = 1, \ldots, m) \). Since increasing model complexity will improve training set prediction but not necessarily true error, the predicted values are commonly estimated via \( n \)-fold cross-validation. That is, \( \hat{t}_{i,k}^{(-i)} \) is calculated by fitting the \( k^{th} \) model without the \( i^{th} \) observation. Then minimizing squared predicted error implies

\[
\hat{\alpha} = \arg \min_{\alpha, \alpha_k \geq 0} \sum_{i=1}^{n} \left( t_i - \left( \sum_{k=1}^{m} \alpha_k \hat{t}_{i,k}^{(-i)} \right) \right)^2,
\]

where the final model predictions are \( \hat{t}_i = \sum_{k=1}^{m} \hat{\alpha}_k \hat{t}_{i,k} \). The non-negativity constraint, \( \hat{\alpha}_k \geq 0 \) for all \( k = 1, \ldots, m \), is not required, but has been shown to perform well (Breiman, 1996; LeBlanc and Tibshirani, 1996).

2.2 Censored Stacking

Stacking survival models is not immediately straightforward. In particular, equation (2.1) is effectively an unknown quantity in the presence of censoring (e.g., \( t_i \) is not always observed). We are also interested in the entire survival curve for a given set of covariates and not just a single quantity. Adjustments are therefore required. While optimal measures of predictive error for survival models are not well established, we use the Brier Score (Graf et al., 1999), which is commonly used and has a connection with squared error. Following Lostritto et al. (2012), the Brier Score at time \( t \) can be written as

\[
BS(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i(t)}{G(T_i(t)|x_i)} \times \left( Z_i(t) - \hat{S}(t|x_i) \right)^2,
\]

where \( Z_i(t) = I(t_i > t), T_i(t) = \min\{t_i, t\}, \Delta_i(t) = I(\min\{t_i, t\} \leq c_i), \) and \( G(\cdot|x_i) \) is the conditional survival function of the censoring distribution. For a fixed time \( t \), censored observations with \( c_i > t \) will contribute to the Brier Score, but when \( c_i < t \) the censored observations will
only contribute to the Brier Score indirectly through the estimation of $G(\cdot | x_i)$. Using double expectation arguments, it is possible to show that $BS(t)$ estimates the expected squared error of the survival distribution at time $t$. Thus the true conditional survival function, $S_o(t | x)$, is the minimizer of $E\{BS(t)\}$.

Since the goal is to estimate the entire conditional survival function, the Brier Score is minimized over a set of time points, say $t_1, ..., t_s$. These modifications imply the following weighted least squares problem with the additional constraint that $\sum_{k=1}^{m} \hat{\alpha}_k = 1$

$$\hat{\alpha} = \arg \min_{\alpha, \alpha_k > 0} \sum_{r=1}^{s} \sum_{i=1}^{n} \frac{\Delta(t_r)}{G(\hat{T}_i(t_r) | x_i)} \times \{ Z_i(t_r) - \sum_{k=1}^{m} \alpha_k \hat{S}_k^{(-i)}(t_r | x_i) \}^2, \quad (2.3)$$

where $\hat{S}_k^{(-i)}(t | x_i)$ is the survival estimate from the $k^{th}$ model that does not include the $i^{th}$ observation in the fitting process. Finally, the stacked estimate of the conditional survival function with time-independent weights is

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

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

The computational requirements of non-parametric estimators generally prevent estimating $\hat{S}_k^{(-i)}(t | x_i)$ with $n$-fold cross-validation. Instead, $v$-fold cross-validation can be used ($v < n$) as the important point is that the $i^{th}$ observation is not used to estimate $\hat{S}_k^{(-i)}(t | x_i)$. For example, Breiman (1996) found that 10-fold cross-validation performed similarly to $n$-fold cross-validation for stacking uncensored regression models. In addition, the simulation studies in Section 4 illustrate that stacked survival models perform well even when $\hat{S}_k^{(-i)}(t | x_i)$ is estimated via five-fold cross-validation.

**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., $t_1, ..., t_s$, have implications for performance. In particular, care should be taken to avoid picking only very small, or very large $t$ values. Preliminary investigations
suggest that using at least four evenly spaced quantiles of the observed event distribution ensures good performance.

**Remark 2.** Time-dependent stacking, i.e., allowing the weighted combination of models to depend on time, was also considered. The results are presented in the Supplementary Materials. Though potentially adding flexibility, a major flaw of time-dependent stacking is that the conditional survival function may, at times, increase, which violates a fundamental property of survival functions. As such, this paper focuses on time-independent stacking.

### 3. Asymptotic Properties

We show that stacked survival models ensure consistent model selection and uniform consistency of the conditional survival function estimate. The former refers to the idea that if the set of stacked models contains asymptotically correct models, then all weight is asymptotically given to the correct models in the stack. Consistent model selection implies uniform consistency as long as the correctly specified model is a uniformly consistent estimator of the conditional survival function. Our main assumption for consistent model selection 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). The Supplementary Materials contain the specific assumptions and proofs.

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

**THEOREM 3.1** Let $\hat{α}$ be estimated by equation (2.3). Assume that models $1, ..., l$, where $l < m$, are the only uniformly consistent estimators and conditions (A1)-(A3) in the Supplementary Materials hold, then $\sum_{k=1}^{l} \hat{α}_k \rightarrow 1$, in probability, as $n \rightarrow \infty$.

This ensures that, for the time-independent weights, the correct 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 when considering different tuning parameters for non-parametric estimators. Another example is a correctly specified Weibull model and Cox model. In the special case, when only one model is correctly specified, we obtain the corollary:

**Corollary 3.2** If \( \hat{S}_1(t|x) \) is the estimate for the only correctly specified conditional survival model, then \( \hat{\alpha}_1 \to 1 \), in probability, as \( n \to \infty \).

The result of Theorem 3.1 and Corollary 3.2 is required for uniform consistency of the stacked estimator with time-independent weights.

**Theorem 3.3** Let the stacked estimate of the conditional survival function be defined as \( \hat{S}(t|x) \) in equation (2.4). Assume that conditions (A1)-(A3) in the Supplementary Materials hold then, as \( n \to \infty \),

\[
\sup_{t \in \Omega} \sup_{x} \left| \sum_{k=1}^{m} \hat{\alpha}_k \hat{S}_k(t|x) - S_0(t|x) \right| \to 0.
\]

Both theorems are proved in Supplementary Materials. The effect of stacking on the rate of convergence of the conditional survival model remains beyond the scope of this paper. 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.

4. Simulations

An extensive simulation study illustrates the finite sample performance of stacked survival models for several commonly encountered situations. In particular, three settings are investigated: a
moderate number of covariates with a modest censoring rate (Section 4.1) and a high censoring rate (Section 4.2), and then a large number of covariates with a modest censoring rate (Section 4.3).

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: $x_p \sim \text{MVN}(\mathbf{0}, \Sigma)$, where $\Sigma$ 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). Sections 4.1 and 4.2 have a four-dimensional covariate space (i.e., $p = 4$), while Section 4.3 has a $p = 40$ dimensional covariate space. For all simulations, only the first and third covariate have a non-zero effect. Specifically, the first three covariate effects are $(1, 0, -1)$. Two different linear combinations are considered: $\gamma^1 = x_p$ and $\gamma^2 = \Phi(4 \times x_p)$ which imply linear and non-linear covariate effects, respectively. The event distributions are defined as

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

Each subsection investigates every combination of the event distribution ($d$) and linear form ($q$), i.e., there are six scenarios for each of the three subsections.

Survival models are compared on the basis of integrated squared survival error (ISSE),

$$\int \int (\hat{S}(u|x) - S_o(u|x))^2 du \ dx,$$

which is approximated by

$$\text{ISSE} \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} (t_{(j)} - t_{(j-1)}) \times (\hat{S}(t_{(j)}|x_i) - S_o(t_{(j)}|x_i))^2,$$

where $t_{(j)}$ are the ordered event times, i.e., $t_{(j)} - t_{(j-1)} > 0$ for all $j$, $N$ is the number of observed event times, and $S_o(\cdot|x_i)$ is the true conditional survival function. Therefore, ISSE measures the squared distance between the estimated and true conditional survival functions.

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

4.1 Modest Censoring

This setting has relatively few covariates ($p = 4$) with a modest censoring rate (25%) and sample size ($n = 200$). This illustrates stacked survival models in a relatively straightforward and simple scenario.

The stacked survival models include a Weibull model and log-Normal model as parametric models, a Cox proportional hazards model as a semi-parametric model, and random survival forests (RSF) as a non-parametric model. The parametric and semi-parametric models only include 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 five-fold cross-validation to estimate $\hat{S}^{(i)}(t|x_i)$. The RSF is estimated using the `randomSurvivalForest` package in R (Ishwaran and Kogalur, 2013). The RSF is an ensemble of 250 trees grown using package defaults. For RSF, $\hat{S}^{(i)}(t|x_i)$ is estimated using the out-of-bag ensemble from the `rsf` function. The censoring distribution is a uniform distribution for all $T^{(q)}; C_{d,q} \sim \text{Unif}(0, c(d, q))$, where $c(d, q)$ is a constant that depends on $(d, q)$ and ensures approximately 25% censoring. A Kaplan-Meier estimates the survival function of the censoring distribution required for the Brier Score, i.e., $G(·|x)$ in equation (2.3).

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 specific parametric model, but should remain reasonably efficient 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 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: random survival forests (RSF).

Table 1 presents the results in terms of integrated squared survival error (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 five of the six scenarios. Stacked survival models reduce the ISSE by approximately 20% compared to the best single model for the log-Normal and gamma distributions with non-linear covariate effects. In addition, the stacking procedure outperforms the approach of selecting the final model via cross-validation in every situation.

As an illustration, Table 2 presents the stacking weights (averaged over all simulations) for the individual models. For the linear scenarios, the stacking procedure gives a majority of the weight to correctly specified parametric models. The weights are even more interesting for the scenarios with non-linear covariate effects. In particular, the random survival forests (RSF) receive the most weight for the stacking procedure despite having the largest ISSE. This is a good example of stacked survival models combining misspecified parametric models and an inefficient non-parametric model to obtain a new estimator that outperforms every single model considered in the stacking procedure.

**Remark 3.** Random survival forests (RSF) 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.
4.2 High Censoring

This setting is similar to Section 4.1 except that the censoring rate is approximately 75% and the sample size is $n = 1000$, which are designed to mimic large observational trials that experience substantial administrative censoring at the end of the observed support. To simulate administrative censoring, the censoring is uniformly distributed: 

$$C_{d,q} \sim \text{Unif}(c(d,q), c(d,q) + 0.5),$$

where $c(d,q)$ is a constant that depends on $(d,q)$ and ensures approximately 75% censoring.

Table 3 presents the results in terms of integrated squared survival error (ISSE). Again, the top two estimators are bolded to highlight flexibility in a wide range scenarios. Stacked survival models is a top two estimator for five of the six scenarios, while none of the alternatives are a top two estimator for more than two scenarios. Additionally, stacking possesses approximately 30% higher ISSE than correctly specified parametric models (i.e., log-Normal and Weibull distributions with linear effects), and as good or better ISSE when the parametric models are slightly misspecified (i.e., Gamma distribution with linear effects). The stacking procedure also outperforms the model selected via cross-validation in every situation.

An interesting point is that the parametric and semi-parametric models perform very poorly for the log-Normal and Gamma scenarios with non-linear covariate effects, where random survival forests (RSF) are the best performing single survival model. Despite the poor performance of the parametric models, stacking is able to improve on the performance of RSF to achieve more than a 10% reduction in integrated squared survival error. In contrast, cross-validation, which always selects RSF as the best performing model, is naturally unable to improve upon the performance of RSF. This again illustrates the ability of stacked survival models to perform well in a wide range of situations and, at times, improve the estimation of the conditional survival function above any model in the stack.
4.3 High Dimensional Covariate Space

This setting has $p = 40$ covariates with a relatively small sample size ($n = 200$). The censoring distributions are the same as Section 4.1. This represents the situation where the number of covariates is large relative to the sample size. In general, the parametric and semi-parametric models used (and stacked) in Sections 4.1 and 4.2 will not perform well in high dimensional settings without regularization. As such, these models are not included for these high dimensional scenarios. We instead stack a Cox model with an $l_1$ penalty (i.e., lasso), a boosted version of the Cox model, and random survival forests (RSF). 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 Sections 4.1 and 4.2.

The stacked survival model is the best performing model for every non-proportional hazards scenario (see Table 4). Stacked survival models are one of the top two estimators for five of the six scenarios and, relative to Sections 4.1 and 4.2 stacked survival models offer smaller improvements (approximately $5\% - 15\%$ reductions in ISSE). However, the improvements in ISSE are more consistent across the scenarios. In addition, the stacking procedure outperforms the model selected via cross-validation in every situation.

5. 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 approximately 75% censoring, which is similar to the simulation scenarios in Section 4.2.

The stacking procedure uses the same models as Sections 4.1 and 4.2. That is, the Weibull and log-Normal model are the parametric models, the Cox proportional hazards model is the semi-parametric model, and a random survival forest 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, 48.

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

Figure 1 presents the estimated five-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 five-year survival with increasing tumor size and number of nodes. In contrast, RSF suggests that five-year survival dips slightly around 40 mm for tumor size, while five-year survival for the number of nodes has a sharp early decrease but plateaus after about 10 nodes. The stacked survival model - which gives weight to the Weibull model (0.36), the Cox model (0.04), and RSF (0.58) - is a compromise between the parametric/semi-parametric models and RSF.

The GBCS data set has a marginal five-year survival rate of 70% due, in part, to a censoring rate of 75%. As such, predicted five-year survival rates less than 20% are surprising (i.e., the parametric/semi-parametric models for the number of nodes). Due to the sparsity of patients with more than 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 less than 20 nodes (98% of patients have less than 20 nodes) through the first order linear effect (note that the patient with over 50 nodes was censored after two years). In contrast, RSF does not require any linearity assumptions and is more influenced by local observations in predicting five-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.

6. 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, in the hope of potentially gaining efficiency from the parametric estimators 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 efficiently estimate conditional survival functions across a wide range of scenarios. In particular, for a given scenario, stacked survival models are consistently one of the two best estimators 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 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 simulation study illustrates 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.

The Supplementary Materials include an investigation of time-dependent stacking, but the resulting survival function is not guaranteed to be non-increasing. This is a major flaw that compromises the conceptual cohesion of time-dependent stacking. As an alternative, time-dependent stacking on the conditional hazard function would ensure a non-increasing survival function. However, this requires a different loss function than the Brier Score and remains beyond the scope of this paper.

Similar to time-dependent stacking, covariate dependent stacking (or, allowing the \( \alpha_k \) to depend upon \( x \)) 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 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 set of \( m \) survival models included in the stacking procedure will influence the performance of the final stacked survival model. Breiman (1996) found that uncensored stacking performed best when the models were fundamentally different. As such, models based on different classes are recommended, e.g., parametric, semi-parametric, and non-parametric estimators. However, the number of models from each class of models to include in the stacking procedure is an open question. For example, an information criterion, e.g., AIC, could select a single model for each parametric and semi-parametric method, but the corresponding impact on performance is
unknown. As such, further research on this topic is warranted.

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 4 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 using a conditional estimator may be warranted (e.g., a Cox proportional hazards model).

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 Materials

The German breast cancer study data is available at http://www.umass.edu/statdata/statdata/index.html.

Acknowledgements

Research reported in this publication was supported in part by grant UL1TR000114 of the National Center for Advancing Translational Sciences.
BINDER, HARALD. (2013). *CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks*. R package version 1.4.

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

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

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

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

GOEMAN, JELLE J. (2012). *Penalized R Package*. R package version 0.9-42.

GRAF, ERIKA, SCHMOOR, CLAUDIA, SAUERBREI, WILLI AND SCHUMACHER, MARTIN. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**, 2529–2545.

HOSMER, DAVID W., LEMESHOW, STANLEY AND MAY, SUSANNE. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley.

ISHWARAN, H. AND KOGALUR, U.B. (2013). *Random Survival Forests*. R package version 3.6.4.

ISHWARAN, HEMANT, KOGALUR, UDAYA B., BLACKSTONE, EUGENE H. AND LAUER, MICHAEL S. (2008). Random survival forests. *Annals of Applied Statistics* **2**, 841–860.

KALBFLEISCH, JOHN D. AND PRENTICE, ROSS L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley.
KAPLAN, E. L. AND MEIER, PAUL. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

KOOPERBERG, CHARLES, STONE, CHARLES J. AND TRUONG, YOUNG K. (1995). Hazard regression. *Journal of the American Statistical Association* **90**, 78–94.

LEBLANC, MICHAEL AND TIBSHIRANI, ROBERT. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association* **91**, 1641–1650.

LOSTRITTO, KAREN, STRAWDERMAN, ROBERT L. AND MOLINARO, ANNETTE M. (2012). A partitioning deletion/substitution/addition algorithm for creating survival risk groups. *Biometrics* **68**, 1146–1156.

POLLEY, ERIC C. AND VAN DER LAAN, MARK. (2011). Super learning for right-censored data. In: Targeted Learning: Causal Inference for Observational and Experimental Data. 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.

RUDSER, KYLE D., LEBLANC, MICHAEL L. AND EMERSON, SCOTT S. (2012). Distribution-free inference on contrasts of arbitrary summary measures of survival. *Statistics in Medicine* **31**, 1722–1737.

SILL, JOSEPH, TAKACS, GABOR, MACKEY, LESTER AND LIN, DAVID. (2009). Feature-weighted linear stacking. *Arxiv*.

THERNEAU, TERRY. (2013). *Survival analysis, including penalized likelihood*. R package version 2.37-4.

VAN DER LAAN, MARK J., POLLEY, ERIC C. AND HUBBARD, ALAN E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology* **6**.
REFERENCES

VARADHAN, RAVI. (2012). *alabama: Constrained nonlinear optimization*. R package version 2011.9-1.

WEY, ANDREW, WANG, LAN AND RUDSER, KYLE. (2014). Censored quantile regression with recursive partitioning based weights. *Biostatistics* **15**, 170–181.

WOLPERT, DAVID H. (1992). Stacked generalization. *Neural Network* **5**, 241–259.

WOOD, SIMON N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall.

ZHAO, YUFAN, ZENG, DONGLIN, SOCINSKI, MARK A. AND KOSOROK, MICHAEL R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* **67**, 1422–1433.

ZHENG, Y AND HEAGERTY, P.J. (2004). Semiparametric estimation of time-dependent roc curves for longitudinal marker data. *Biostatistics* **5**, 615–632.

ZHU, RUOQING AND KOSOROK, MICHAEL R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association* **107**, 331–340.
Table 1. Simulation results for Section 4.1 ($n = 200, \ p = 4$ covariates, and 25% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 2000 times, and the error is multiplied by 100. The two top estimators are bolded for each simulation scenario. ‘RSF’ stands for random survival forests, ‘Stacking’ is stacked survival models, and ‘CV’ is the cross-validated selected estimator.

| Models          | log-Normal | Weibull | Gamma |
|-----------------|------------|---------|-------|
| Linear Effects  |            |         |       |
| Single Models   | log-Normal | 0.92    | 2.48  | 1.17  |
| Models Cox      |            | 3.76    | 2.03  | 3.49  |
| Models CV       |            | 3.97    | 1.99  | 2.38  |
| Flexible Models | log-Normal | 7.59    | 4.80  | 9.02  |
| Models CV       |            | 10.3    | 5.62  | 12.1  |

Table 2. Average weights for the individual models included in the stacked survival model for each of the six scenarios in Section 4.1 ($n = 200, \ p = 4$ covariates, and 25% censoring). Each simulation is replicated 2000 times. ‘RSF’ stands for random survival forests.

| Stacked Models | log-Normal | Weibull | Gamma |
|----------------|------------|---------|-------|
| Linear Effects |            |         |       |
| log-Normal     | 0.64       | 0.17    | 0.43  |
| Weibull        | 0.21       | 0.49    | 0.39  |
| Cox            | 0.13       | 0.29    | 0.16  |
| RSF            | 0.03       | 0.06    | 0.02  |
| Non-Linear Effects |    |         |       |
| log-Normal     | 0.21       | 0.12    | 0.14  |
| Weibull        | 0.10       | 0.31    | 0.14  |
| Cox            | 0.03       | 0.22    | 0.04  |
| RSF            | 0.67       | 0.35    | 0.68  |
Table 3. Simulation results for Section 4.2 \((n = 1000, p = 4\) covariates, and 75\% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 2000 times, and the error is multiplied by 1000. The two top estimators are bolded for each simulation scenario. ‘RSF’ stands for random survival forests, ‘Stacking’ is stacked survival models, and ‘CV’ is the cross-validated selected estimator.

| Models         | log-Normal | Weibull | Gamma |
|----------------|------------|---------|-------|
|                | Single     |         |       |
| Single Models  | Log-Normal | 0.26    | 1.02  | 0.38  |
| Weibull        | RSF        | 0.80    | 0.33  | 0.45  |
| Cox            | CV         | 0.99    | 0.41  | 0.69  |
|                | Flexible   | 7.64    | 4.84  | 8.16  |
| Stacking       |            | 0.34    | 0.42  | 0.37  |
| Models         | CV         | 0.93    | 0.71  | 0.55  |

Table 4. Simulation results for Section 4.3 \((n = 200, p = 40\) covariates, and 25\% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 2000 times, and the error is multiplied by 100. The two top estimators are bolded for each simulation scenario. ‘RSF’ stands for random survival forests, ‘Stacking’ is stacked survival models, and ‘CV’ is the cross-validated selected estimator.

| Models         | log-Normal | Weibull | Gamma |
|----------------|------------|---------|-------|
|                | Single     |         |       |
| Single Models  | Log-Normal | 13.2    | 2.96  | 14.9  |
| Cox - Lasso    | RSF        | 14.1    | 2.84  | 15.6  |
| Cox - Boosting | CV         | 14.0    | 2.90  | 15.6  |
|                | Flexible   | 5.20    | 3.47  | 5.63  |
| Stacking       |            | 4.56    | 2.29  | 5.01  |
| Models         | CV         | 5.20    | 3.47  | 5.63  |
Fig. 1. The association of tumor size (mm) and the number of nodes with five-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.