*Biostatistics* (2014), **0**, 0, *pp.* 1–23
doi:10.1093/biostatistics/stacking˙supplementary˙materials˙v7

# Supplementary Materials to "Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models"

ANDREW WEY*, JOHN CONNETT, KYLE RUDSER

*University of Hawaii and University of Minnesota*

Section 1 demonstrates the connection between the Brier Score in the absence of censoring versus the inverse probability-of-censoring weighted Brier Score. Section 2 derives the mean-squared error decomposition presented in Section 3 of the main paper, and presents illustrations and examples regarding the impact of candidate survival models on performance. In addition, a simple example illustrates the situation of stacking a "parametric" and "non-parametric" survival model. Section 3 proves the asymptotic properties of stacked survival models presented in Section 4 of the main paper. In addition, Section 3.3 discusses the potential of distributional results for the conditional survival function. Section 4 discusses time-dependent stacking and compares the performance to time-independent stacking [equation (2.4) in the main paper]. Section 5 considers several extensions to the simulation study in the main paper.

## 1. $BS(t)$ versus $IPCW$-$BS(t)$

We demonstrate the difference between the Brier Score, $BS(t)$ [see equation (2.1) in the main paper], and the inverse probability-of-censoring weighted Brier Score for censored data, $IPCW$-$BS(t)$

*To whom correspondence should be addressed.

[see equation (2.2) in the main paper]. In particular, we consider a simple simulation scenario where the survival time distribution is an exponential distribution with a rate parameter of one. We estimate $BS(t)$ with the complete data and then estimate the $IPCW\text{-}BS(t)$ using data subject to two different levels of censoring. The censoring is generated by an exponential censoring distribution with a rate parameter that ensures approximately 25% and 50% censoring. We note that the estimated survival function is the same for each estimate of the Brier Score, thus the differences are entirely due to inverse probability-of-censoring weights, which adjust the observed Brier Score to account for censoring. Figure 1 illustrates that the $BS(t)$ and $IPCW\text{-}BS(t)$ are very similar early in the support, and experience progressively more variation as the relative amount of censoring increases.

## 2. Mean-Squared Error Decomposition

We want to define certain quantities and make a connection to the Brier Score before deriving the mean-squared error decomposition presented in Section 3 of the main paper. We define the mean-squared error for a conditional survival function estimator as the integral of the squared error at time $t$ over $\Omega = (0, \tau)$: $\text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\} = E\int_0^\tau [\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]^2 dt$, where the expectation is over the random variable for the covariate space and the sampling distribution of the stacked estimator. As mentioned in Section 3 of the main paper, the mean-squared error has a direct connection to the Brier Score. In particular,

$$
\begin{aligned}
E\int_0^\tau IPCW\text{-}BS(t)dt &= \int_0^\tau E\left[\frac{\Delta(t)}{G(T(t)|\boldsymbol{x})} \times \{Z(t) - \hat{S}(t|\boldsymbol{x})\}^2\right] \\
&= \int_0^\tau E\left[\{Z(t) - \hat{S}(t|\boldsymbol{x})\}^2 dt\right], \text{ by iterative expectation} \\
&= E\int_0^\tau \{Z(t) - S_o(t|\boldsymbol{x}) + S_o(t|\boldsymbol{x}) - \hat{S}(t|\boldsymbol{x})\}^2 dt \\
&= E\int_0^\tau \left\{[Z(t) - S_o(t|\boldsymbol{x})]^2 + [S_o(t|\boldsymbol{x}) - \hat{S}(t|\boldsymbol{x})]^2\right\} dt \\
&= \sigma^2 + \text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\},
\end{aligned}
$$

where $\sigma^2$ is irreducible prediction error.

We define the bias and variance of a conditional survival function estimator at time $t$ as, respectively, $\text{Bias}\{\hat{S}(t|\boldsymbol{x})\} = E[\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]$ and $\text{Var}\{\hat{S}(t|\boldsymbol{x})\} = E[\hat{S}(t|\boldsymbol{x}) - E\hat{S}(t|\boldsymbol{x})]^2$. The mean squared error of the stacked estimator is then decomposed as

$$
\begin{aligned}
\text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\} &= E\int_0^\tau [\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]^2 dt \\
&= E\int_0^\tau [\hat{S}(t|\boldsymbol{x}) - E\hat{S}(t|\boldsymbol{x}) + E\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]^2 dt \\
&= \int_0^\tau E\{[\hat{S}(t|\boldsymbol{x}) - E\hat{S}(t|\boldsymbol{x})]^2 + [E\hat{S}(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]^2\} dt \\
&= \int_0^\tau E\left\{[\sum_{k=1}^m \alpha_k\{\hat{S}_k(t|\boldsymbol{x}) - E\hat{S}_k(t|\boldsymbol{x})\}]^2 + [\sum_{k=1}^m \alpha_k\{E\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})\}]^2\right\} dt \quad (2.1) \\
&= \int_0^\tau \sum_{k=1}^m \sum_{l=1}^m \alpha_k\alpha_l \Big\{E\{[\hat{S}_k(t|\boldsymbol{x}) - E\hat{S}_k(t|\boldsymbol{x})][\hat{S}_l(t|\boldsymbol{x}) - E\hat{S}_l(t|\boldsymbol{x})]\} + \\
&\qquad E\{[E\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})][E\hat{S}_l(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})]\}\Big\} \\
&= \sum_{k=1}^m \alpha_k^2 E\int_0^\tau \Big[\text{Bias}^2\{\hat{S}_k(t|\boldsymbol{x})\} + \text{Var}\{\hat{S}_k(t|\boldsymbol{x})\}\Big] dt + \\
&\qquad E\sum_{k=1}^m \sum_{l\neq k} \alpha_k\alpha_l \int_0^\tau \Big[\text{Bias}\{\hat{S}_k(t|\boldsymbol{x})\} \times \text{Bias}\{\hat{S}_l(t|\boldsymbol{x})\} + \text{Cov}\{\hat{S}_k(t|\boldsymbol{x}), \hat{S}_l(t|\boldsymbol{x})\}\Big] dt \\
&= \sum_{k=1}^m \alpha_k^2 \text{MSE}_\tau\{\hat{S}_k(\cdot|\boldsymbol{x})\} + E\sum_{k=1}^m \sum_{l\neq k} \alpha_k\alpha_l \int_0^\tau \Big[\text{Bias}\{\hat{S}_k(t|\boldsymbol{x})\} \times \text{Bias}\{\hat{S}_l(t|\boldsymbol{x})\} + \\
&\qquad \text{Corr}\{\hat{S}_k(t|\boldsymbol{x}), \hat{S}_l(t|\boldsymbol{x})\} \times \text{Var}\{\hat{S}_k(t|\boldsymbol{x})\}^{\frac{1}{2}} \times \text{Var}\{\hat{S}_l(t|\boldsymbol{x})\}^{\frac{1}{2}}\Big] dt
\end{aligned}
$$

where line (2.1) holds since $\sum_{k=1}^m \alpha_k S_o(t|\boldsymbol{x}) = S_o(t|\boldsymbol{x})$ by the sum-to-one constraint. As outlined in the main paper, this decomposition motivates stacking a diverse set of survival models in order to lower the correlation between predicted survival curves.

We now consider a simple example to illustrate the potential of stacking a survival model with low bias but high variance and a badly misspecified parametric model (i.e., high bias; low variance). For example, consider a set of two independent candidate survival models where both survival models possess the same mean-squared error, say $\text{MSE}_\tau\{\hat{S}_k(\cdot|\boldsymbol{x})\} = 10$ for $k = 1, 2$. For the sake of simplicity, the bias and variance of both survival models are constant across time and

the covariate space. We set the bias of the first model to $\int_0^\tau \text{Bias}^2\{\hat{S}_1(t|\boldsymbol{x})\}dt = 9$ and consider two different values of bias for the second model: $\text{Bias}\{\hat{S}_2(t|\boldsymbol{x})\} = 0$ and $\int_0^\tau \text{Bias}^2\{\hat{S}_2(t|\boldsymbol{x})\}dt = 1$. These fully define the operating characteristics of both survival models. The first model corresponds to a badly misspecified parametric model, while the second model corresponds to a non-parametric estimator (first with no bias, then with some small sample bias). For this simple scenario, Figure 2 illustrates that the stacked estimator always achieves lower MSE than the individual models when the non-parametric estimator possesses no bias (left plot). However, this advantage can, depending on the degree of correlation, decrease substantially when the non-parametric estimator is slightly biased (right plot). In addition, the effectiveness of the stacked estimator in both situations decreases when the correlation increases between the survival models.

Based on the MSE decomposition, we can see that, everything else being equal, smaller pairwise correlations will improve the performance of the stacked estimator. In particular, suppose there are two potential survival models to add to the current set of candidate survival models that possess the same bias and variance terms, then adding the model with the lower correlation with the current set of candidate models in the stack would be expected to perform better. It is noteworthy that models that span parametric, semi-parametric and non-parametric classes may still be correlated. Thus, as recommended by an anonymous reviewer, a direction of future research may be the development of a method that incorporates correlation information into the estimation of weights (see Section 7 of the main paper for further discussion).

To further illustrate the impact of correlated candidate survival models on performance, we present the results of a small simulation study. In particular, we consider the gamma and Weibull distributed scenarios with linear covariate effects (i.e., with respect to the specifications outlined in the main paper: $\{d, q\} = \{3, 1\}$ and $\{d, q\} = \{2, 1\}$ for the gamma and Weibull scenarios, respectively). In this example, suppose we want to use the Cox proportional hazards model with a baseline hazard estimate, but are unsure about the covariates to include in the Cox model and,

thus, consider up to six different models. The first model is the same model that is fit in the main paper (eight covariates), while each successive model fits a Cox model with two additional noise covariates until there are a total of eighteen covariates. Figure 3 illustrates that the ISSE of the stacked estimator increases as more and more correlated Cox models are added to the current set of models to the stack. We also find that the average number of models receiving non-zero weight increases as additional correlated models are added to the current set of candidate models in the stack, and the percentage of times that only one model receives non-zero weight decreases as additional correlated models are added (Table 1).

## 3. Asymptotic Properties

For both proofs, we consider the general case where $l$ of the $m$ estimators for the stacking procedure are uniformly consistent. Recall that the conditional survival function is estimated on $\Omega = (0, \tau)$. Now assume the following conditions throughout the Supplementary Materials.

**A1.** There exists $l$ estimators that are uniformly consistent within the set of estimators used for the stacking procedure. Without loss of generality, let these estimators be $\hat{S}_k(t|\boldsymbol{x})$ for $k = 1, ..., l$ and, by uniform consistency, $\sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})| \to 0$ for $k = 1, ..., l$. Additionally, for $k = l+1, ..., m$, $\sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})| \to c_k$ where $c_k > 0$.

**A2.** The estimator for the censoring distribution is uniformly consistent: $\sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\hat{G}(t|\boldsymbol{x}) - G(t|\boldsymbol{x})| \to 0$, and there exists a $\delta > 0$ such that $G(\tau|\boldsymbol{x}) > \delta$ for all $\boldsymbol{x}$.

**A3.** Let $\Gamma = \{t_1, t_2, ..., t_s\}$ where $t_r \epsilon \Omega$ for $r = 1, ..., s$, i.e., the set of time points used to minimize the Brier Score for the stacking procedure. Define the sum of misspecified model weights as $\widetilde{\alpha} = \sum_{k=l+1}^{m} \alpha_k$. Then for all $\widetilde{\alpha} > 0$ there exists at least one $t_r \epsilon \Gamma$ such that,

$$\sup_{\boldsymbol{x}} | \sum_{k=l+1}^{m} \frac{\alpha_k}{\widetilde{\alpha}} \hat{S}_k(t_r|\boldsymbol{x}) - S_o(t_r|\boldsymbol{x})| \to c,$$

where $c > 0$.

### 3.1 *Proof of Theorem 4.1*

Lets start by considering the supremum of the difference between the observed and expected inverse probability-of-censoring weighted Brier Score:

$$\begin{aligned}
\sup_{\alpha}|\sum_{i=1}^{n}\frac{\Delta_i(t)}{\hat{G}(T_i(t)|\boldsymbol{x}_i)}\times\{Z_i(t)-\sum_{k=1}^{m}\alpha_k\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)\}^2-E\{IPCW\text{-}BS(t)\}| &= \sup_{\alpha}|\sum_{i=1}^{n}\Delta_i(t)\left[\frac{1}{\hat{G}(T_i(t)|\boldsymbol{x}_i)}-\right.\\
&\left.\frac{1}{G(T_i(t)|\boldsymbol{x}_i)}+\frac{1}{G(T_i(t)|\boldsymbol{x}_i)}\right]\times\{Z_i(t)-\sum_{k=1}^{m}\alpha_k\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)\}^2-E\{IPCW\text{-}BS(t)\}|\\
&\leqslant \sup_{\alpha}|\sum_{i=1}^{n}\left[\frac{\Delta_i(t)}{\hat{G}(T_i(t)|\boldsymbol{x}_i)}-\frac{\Delta_i(t)}{G(T_i(t)|\boldsymbol{x}_i)}\right]\times\{Z_i(t)-\sum_{k=1}^{m}\alpha_k\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)\}^2|+\\
&\sup_{\alpha}|\sum_{i=1}^{n}\frac{\Delta_i(t)}{G(T_i(t)|\boldsymbol{x}_i)}\times\{Z_i(t)-\sum_{k=1}^{m}\alpha_k\hat{S}_k^{(-i)}(t|\boldsymbol{x}_i)\}^2-E\{IPCW\text{-}BS(t)\}|
\end{aligned}$$

The first supremum, which is the difference between estimated and true censoring distributions, approaches zero due to assumption A2. The second supremum approaches zero if the expected value of the Brier Score is bounded by some function that has finite expectation [see Lemma 2.4 of Newey and McFadden (1994)]. By assumption A2,

$$E[\frac{\Delta_i(t_r)}{G(T_i(t_r)|\boldsymbol{x}_i)}\{Z_i(t_r)-\sum_{k=1}^{m}\alpha_k\hat{S}_k^{(-i)}(t_r|\boldsymbol{x}_i)\}^2]<E[\frac{1}{\delta}\{Z_i(t_r)-\sum_{k=1}^{m}\alpha_k\hat{S}_k^{(-i)}(t_r|\boldsymbol{x}_i)\}^2]<\infty.$$

This implies that our minimization procedure, i.e., the Brier Score, asymptotically approaches its expectation. We can therefore determine the asymptotic minimizer by considering the expectation of the Brier Score. Since assumption A2 implies $E[\frac{\Delta(t_r)}{G(T(t_r)|\boldsymbol{x})}|T,\boldsymbol{x}]=\frac{1}{G(T(t_r)|\boldsymbol{x})}\times E[\Delta(t_r)|T,\boldsymbol{x}]=1$, we obtain by double expectation that

$$\begin{aligned}
E[IPCW\text{-}BS(t_r)|\boldsymbol{x}] &= E[\{Z(t_r)-\sum_{k=1}^{m}\alpha_k S_k(t_r|\boldsymbol{x})\}^2|\boldsymbol{x}]\\
&= E[\{Z(t_r)-S_o(t_r|\boldsymbol{x})\}^2|\boldsymbol{x}]+\{S_o(t_r|\boldsymbol{x})-\sum_{k=1}^{m}\alpha_k S_k(t_r|\boldsymbol{x})\}^2\\
&= S_o(t_r|\boldsymbol{x})\{1-S_o(t_r|\boldsymbol{x})\}+\{S_o(t_r|\boldsymbol{x})-\sum_{k=1}^{m}\alpha_k S_k(t_r|\boldsymbol{x})\}^2
\end{aligned}$$

The asymptotic minimization problem becomes

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \arg\min_{\boldsymbol{\alpha}} \sum_{r=1}^{s}[S_o(t_r|\boldsymbol{x})(1-S_o(t_r|\boldsymbol{x})) + \{S_o(t_r|\boldsymbol{x}) - [\sum_{k=1}^{l}\alpha_k S_o(t_r|\boldsymbol{x}) + \sum_{k=l+1}^{m}\alpha_k S_k(t_r|\boldsymbol{x})]\}^2] \\ &= \arg\min_{\boldsymbol{\alpha}} \sum_{r=1}^{s}[S_o(t_r|\boldsymbol{x}) - \{S_o(t_r|\boldsymbol{x})\sum_{k=1}^{l}\alpha_k + \sum_{k=l+1}^{m}\alpha_k S_k(t_r|\boldsymbol{x})\}]^2.\end{aligned}$$

At this point, we know there exists $\boldsymbol{\alpha}$ such that $\sum_{k=1}^{m}\alpha_k S_k(t_r|\boldsymbol{x}) = S_o(t_r|\boldsymbol{x})$, e.g., $\alpha_1 = 1$. However, we need to show that the sum of correctly specified model weights equals one, i.e., $\sum_{k=1}^{l}\alpha_k = 1$. Suppose the sum of misspecified model weights is greater than zero, i.e., $\widetilde{\alpha} = \sum_{k=l+1}^{m}\alpha_k > 0$, then

$$\begin{aligned}&\sum_{k=1}^{m}\alpha_k S_k(t_r|\boldsymbol{x}) = S_o(t_r|\boldsymbol{x}) \\ \Rightarrow\ & S_o(t_r|\boldsymbol{x})\sum_{k=1}^{l}\alpha_k + \sum_{k=l+1}^{m}\alpha_k S_k(t_r|\boldsymbol{x}) = S_o(t_r|\boldsymbol{x}).\end{aligned}$$

Subtracting the correctly specified models from each side, we obtain by the sum-to-one constraint that $\sum_{k=l+1}^{m}\alpha_k S_k(t_r|\boldsymbol{x}) = S_o(t_r|\boldsymbol{x})\sum_{k=l+1}^{m}\alpha_k$. This implies that

$$\sum_{k=l+1}^{m}\frac{\alpha_k}{\widetilde{\alpha}}S_k(t_r|\boldsymbol{x}) = S_o(t_r|\boldsymbol{x}),$$

which contradicts assumption A3. Therefore, by the non-negativity constraint, $\widetilde{\alpha} = 0$ and hence $\sum_{k=1}^{l}\hat{\alpha}_k \to 1$ as $n \to \infty$.

### 3.2 *Proof of Theorem 4.3*

Define the random variables: $A_n^k = \sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})|$. By assumption A1, as $n \to \infty$, $A_n^k \to 0$ for $k = 1, ...l$, while $A_n^k \to c_k$ for some $c_k > 0$ $(k = l+1, ..., m)$. Now

$$\begin{aligned}
\sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\sum_{k=1}^{m} \hat{\alpha}_k \hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})| &= \sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\sum_{k=1}^{m} \hat{\alpha}_k \hat{S}_k(t|\boldsymbol{x}) - \sum_{k=1}^{m} \hat{\alpha}_k S_o(t|\boldsymbol{x})|, \text{ since } \sum_{k=1}^{m} \hat{\alpha}_k = 1 \\
&= \sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\sum_{k=1}^{m} \hat{\alpha}_k \{\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})\}| \\
&\leqslant \sum_{k=1}^{m} \hat{\alpha}_k \sup_{t\epsilon\Omega} \sup_{\boldsymbol{x}} |\hat{S}_k(t|\boldsymbol{x}) - S_o(t|\boldsymbol{x})|, \text{ triangle inequality} \\
&= \sum_{k=1}^{l} \hat{\alpha}_k A_n^k + \sum_{k=l+1}^{m} \hat{\alpha}_k A_n^k \\
&\to 1 \times 0 + \sum_{k=l+1}^{m} \{0 \times c_k\} = 0,
\end{aligned}$$

by Slutsky's lemma and Theorem 1. This implies that the stacked estimator is uniformly consistent as long as the correctly specified models are uniformly consistent.

### 3.3 *Distributional Results*

The main paper briefly mentions that the stacked estimate of the conditional survival function is likely an intractable distribution. This is specifically due to the constrained minimization of $\boldsymbol{\alpha}$. For example, the estimation procedure for $\boldsymbol{\alpha}$ has a strong connection with the LASSO; in particular, replace the sum-to-one constraint with a sum-less-than-one constraint. Potscher and Leeb (2009) show that the asymptotic distribution of LASSO parameter estimates depends on the asymptotic behavior of the penalty term. In some cases, the asymptotic distribution is Normal, but Potscher and Leeb (2009) argue that the asymptotic distribution is, in general, a poor representation of the finite sample distribution which is always non-Normal. In addition, a major practical difficulty is that there exists no uniformly consistent estimator of the distribution of LASSO parameter estimates. These issues present major issues for a distributional result for

stacked survival models.

## 4. Time-Dependent Stacking

Potential added flexibility for stacked survival models allows the weights to depend on time, i.e., $\hat{\alpha}_k(t)$. Similar to the approach proposed by Fan and Zhang (2000) for functional data analysis, a two-step estimation procedure is investigated that first obtains "raw estimates" at event times, then smoothes the raw estimates to obtain the final refined time-dependent weights.

The first step estimates the stacking weights for each $N$ event times (i.e., $t_{(1)}, ..., t_{(N)}$). That is, for a given $t_{(r)}$,

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

with the additional constraint that $\sum_{k=1}^{m} \hat{\alpha}_k(t_{(r)}) = 1$. The $\hat{\alpha}_k(t_{(r)})$ are called the "raw estimates". Since the raw estimates can vary substantially across time, the second step smoothes the raw estimates to decrease variability.

While there are several potential avenues for smoothing the raw estimates, local constant regression, e.g., see Ruppert *et al.* (2003), stabilizes the estimates while maintaining both the sum-to-one and non-negativity constraints. In particular,

$$\hat{\alpha}_k^{TD}(t) = \sum_{r=1}^{N} K\left(\frac{t_{(r)} - t}{h}\right) \hat{\alpha}_k(t_{(r)}) / \sum_{r=1}^{N} K\left(\frac{t_{(r)} - t}{h}\right),$$

where $t_{(r)}$ the $r^{th}$ ordered event time, and $K(\cdot)$ is a symmetric probability density. The final estimate for the time-dependent stacking procedure is

$$\hat{S}^{TD}(t|\boldsymbol{x}) = \sum_{k=1}^{m} \hat{\alpha}_k^{TD}(t) \hat{S}_k(t|\boldsymbol{x}), \tag{4.2}$$

where $\hat{S}_k(t|\boldsymbol{x})$ is the $k^{th}$ conditional survival estimate using all the data. This two-step approach to estimating time-dependent weights is appealing for its simplicity and straightforward computational implementation.

Conceptually, time-dependent weights may perform better by shifting weight between survival models as the appropriateness of the distributional assumptions vary across time. However, time-dependent weights increase the flexibility of stacked survival models and, therefore, generally possess a larger variance than time-independent weights. As such, when the stack includes a correctly specified model, time-independent weights will likely perform better than time-dependent weights. In addition, the conditional survival function with time-dependent weights, i.e., $\hat{S}^{TD}(t|\boldsymbol{x})$, is not guaranteed to be a non-increasing function with respect to survival time (which is an essential characteristic of survival functions). In fact, an increasing survival function occurred for a handful of points in the GBCS analysis (see Section 6 in the main paper). As such, time-dependent weights may improve predictive performance, but the conceptual cohesion of the conditional survival function is potentially compromised.

We note that adding a non-decreasing constraint on all of the time-dependent weights would ensure a non-increasing survival function. However, it is easy to show that a non-decreasing constraint on all of the time-dependent weights would result in constant (i.e., time-independent) weights due to the sum-to-one constraint. In addition, we note that a non-decreasing constraint on one survival model and a non-increasing constraint on a separate survival model will not ensure a non-increasing survival function. Thus, it is difficult to fix the conceptual cohesion of the time-dependent stacking.

**Remark 1.** Estimating time-dependent weights requires the selection of a bandwidth $h$. A reasonable approach sets the bandwidth to the standard error of the observed event times. This ensures $h$ approaches 0 at a correct speed for the asymptotic results. However, there may exist a more optimal approach.

**Remark 2.** Tables 2 and 3 compare time-dependent stacking to time-independent stacking (i.e., the approach in the main paper). Time-independent stacking performs as good, or slightly better,

than time-dependent stacking. Although, time-dependent stacking was slightly better for non-linear effects with a large covariate space.

## 5. Furthering the Simulation Study

In this section, we consider several extensions to the simulation study introduced in the main paper. Section 5.1 investigates the simulation scenarios in the main paper when the sample size is doubled. Section 5.2 investigates the robustness of the stacking procedure to the misspecification of a censoring model. Section 5.3 investigates the computational time required for different aspects of stacking (e.g., minimizing the weighted least squares problem). Section 5.4 investigates the effect of the out-of-bag estimator for RSF on the performance of the stacked estimator. Section 5.5 investigates the impact of including additional points when estimating the stacking weights. Section 5.6 considers simulation scenarios with a high censoring rate (e.g., administrative censoring). Section 5.7 considers simulation scenarios with quadratic covariate effects rather than a smooth step function.

### 5.1 *Larger Sample Size*

The simulation scenarios presented in the main paper are extended with a larger sample size ($n = 400$). Tables 4 and 5 illustrate that the qualitative observations in the main paper (e.g., the stacking procedure performing well in a wide variety of situations) remain the same. In addition, the ISSE decreases as expected in each scenario with a larger sample size.

### 5.2 *Misspecified Censoring Distribution*

The simulation scenarios presented in the main paper are modified to have a conditional censoring distribution. In particular, the censoring distribution is the same as the event distribution except the scale/mean parameter is scaled to ensure approximately 25% censoring. Similar to the main

paper, the stacking procedure and the cross-validated estimator use a Kaplan-Meier, which is misspecified in these scenarios, to estimate the censoring distribution in the Brier Score. Tables 6 and 7 illustrate that the stacking procedure and cross-validation are robust against a misspecified censoring distribution for the scenarios investigated here. In particular, the stacking procedure remains a top two estimator in every simulation scenario.

### 5.3 *Computational Time*

For the simulation scenarios presented in the main paper, we investigate the average elapsed time (in seconds) for the stacking procedure with time-independent and time-dependent weights. The average elapsed time is also presented for the minimization problem. Tables 8 and 9 show that the estimation of individual models and the out-of-bag estimators are the main computational cost for stacked survival models. In particular, the 'minimization' of the time-independent stacking is always a fraction of the elapsed time for time-independent stacking. In addition, time-dependent stacking almost always takes longer than time-independent stacking. This is not surprising considering that time-dependent stacking requires $N$ minimizations, where $N$ is the number of unique events, of the weighted least squares problem.

It is important to note that this is the elapsed time rather than computational time. As such, the presence of other processes on shared Unix serves affects the average time presented here. However, these results provide a rough guide on the ordering of the computational requirements for each method. In particular, the minimal effect of the minimization procedure for time-independent stacking is not surprising, while the time-dependent stacking procedure is expected to take longer than the time-independent stacking procedure.

### 5.4 *Out-of-bag estimator for RSF*

In the main paper, the `rsf` function implicitly estimates the out-of-bag estimate for random survival forests (RSF) [see Ishwaran *et al.* (2008) for details]. This out-of-bag estimate is not the same as the five-fold cross-validation estimator used for the parametric and semi-parametric models. The out-of-bag estimate was used from the `rsf` function for computational convenience. However, the differential estimation of the out-of-bag estimate may have influenced the performance of the stacked estimator. To investigate this point, Table 10 presents the ratio of ISSE for the stacked estimator when the out-of-bag estimator for RSF is the `rsf` function versus five-fold cross-validation. In every scenario, the performance of the stacked estimator was insensitive to the out-of-bag estimator of RSF. In addition, Table 11 illustrates that, in every scenario, the `rsf` function is approximately 65% faster (computationally) than the five-fold cross-validation.

### 5.5 *Influence of the Number of Time Points*

As mentioned in the main paper, the performance of the stacked estimator is influenced by the number time points in the minimization procedure. However, there is a point of diminishing returns where the addition of more time points does not appreciably improve the stacked estimator. We have generally found that nine equally spaced quantiles of the observed event distribution is past the point of diminishing returns. Table 12 shows that the ratio of ISSE is essentially one for 9 and 19 equally spaced quantiles of the observed event distribution. Despite the extremely marginal improvement in ISSE, Table 13 shows that 9 equally spaced quantiles takes approximately 40% less computational time than 19 equally spaced quantiles of the observed event distribution.

### 5.6 *High Censoring*

This setting is similar to Section 5.1 in the main paper except that the censoring rate is approximately 75% and the sample size is $n = 1000$. In addition, the censoring distribution is 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 14 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 20% 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. For the non-linear scenario, the stacking possesses 10% to 20% lower ISSE than second best survival model (i.e., stacking is outperforming each model in the set of candidate survival models).

### 5.7 *Quadratic Non-Linearity*

These settings are similar to both simulation scenarios in the main paper, but the non-linearity is induced by a quadratic function rather than a 'smooth' monotonic step function. In particular, the only difference is that $\boldsymbol{\gamma} = \boldsymbol{x}_p^2$ rather than $\boldsymbol{\gamma}^1 = \boldsymbol{x}_p$ or $\boldsymbol{\gamma}^2 = \Phi(4 \times \boldsymbol{x}_p)$, which are studied in the main paper. Similar to Section 5 in the main paper, the censoring rate is a uniform distribution designed to enforce approximately 25% censoring.

Tables 15 presents the results for the modest and large covariate spaces with quadratic covari-

ate effects. In contrast to the non-linear scenario in the main paper, which has a 'smooth' and monotone step function, the random survival forests (RSF) perform substantially better relative to the parametric and semi-parametric models. In addition, the stacked estimator is, again, a top two estimator in every scenario with larger relative differences compared to the scenarios with a 'smooth step function'. The stacked estimator also performs better in every scenario than the cross-validated estimator.

## References

Fan, Jianqing and Zhang, Jin-Ting. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society. Series B* **62**, 303–322.

Ishwaran, Hemant, Kogalur, Udaya B., Blackstone, Eugene H. and Lauer, Michael S. (2008). Random survival forests. *Annals of Applied Statistics* **2**, 841–860.

Newey, Whitney K. and McFadden, Daniel. (1994). Large sample estimation and hypothesis testing. In: *The Handbook of Econometrics*, Volume 4. Amsterdam: North-Holland.

Potscher, Benedikt M. and Leeb, Hannes. (2009). On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis* **100**, 2065–2082.

Ruppert, David, Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Fig. 1. The difference between the Brier Score calculated on complete data ($BS(t)$) versus the inverse probability-of-censoring weighted Brier Score ($IPCW\text{-}BS(t)$). The sample size for the illustration is 5000.

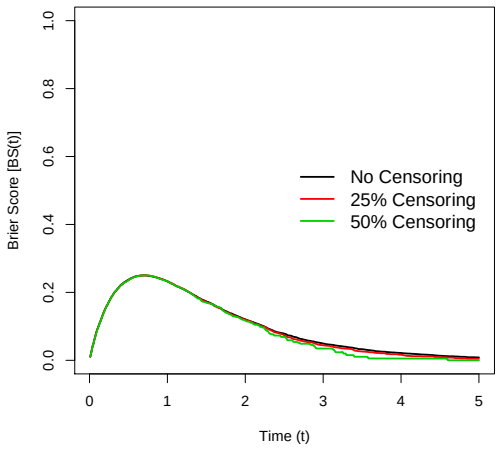


Fig. 2. The example of stacking a misspecified, but efficient, parametric model and a low biased, but highly variable, non-parametric model (see Section 2). Figure (a) shows a non-parametric estimator with no bias, while Figure (b) shows a non-parametric estimator with a small amount of bias. Note that the effectiveness of stacking decreases as the correlation ($\rho$) increases. The mean-squared error of both candidate survival models is 10, which is represented by the dotted line.

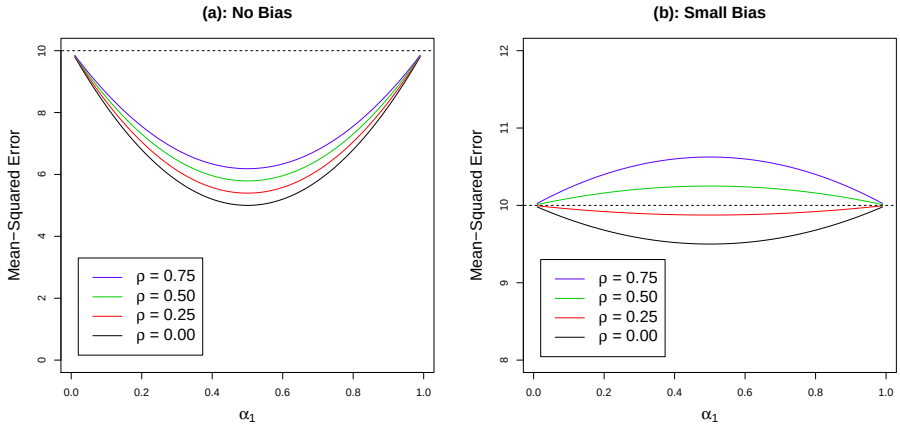

Table 1. Simulation results for Section 2 in the Supplementary Materials. The scenarios investigate the effect of including highly correlated candidate survival models, and correspond to the Gamma and Weibull scenarios with linear covariate effects (i.e., following notation in the main paper: $\{d, q\} = \{2, 1\}$ and $\{d, q\} = \{3, 1\}$ for the Weibull and gamma scenarios, respectively, with $n = 200$). The simulation scenario is replicated 2000 times.

| Number of Candidate Survival Models | Number of Models with Non-Zero Weights | | % of Times Only One Model Recieved Weight | |
|---|---|---|---|---|
| Scenario | Gamma | Weibull | Gamma | Weibull |
| 2 | 1.73 | 1.66 | 26.6% | 33.7% |
| 3 | 2.17 | 2.06 | 13.8% | 20.0% |
| 4 | 2.49 | 2.34 | 9.0% | 13.0% |
| 5 | 2.73 | 2.56 | 6.8% | 9.9% |
| 6 | 2.93 | 2.74 | 5.2% | 8.3% |

Table 2. Simulation results for Section 5.1 in the main paper ($n = 200$, $p = 8$ 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 10. 'Stacking (TI)' is stacking with time-independent weights, and 'Stacking (TD)' is stacking with time-dependent weights. The standard error for the estimate ISSE for each method in each scenario is less than 0.01.

| | Models | log-Normal | Weibull | Gamma |
|---|---|---|---|---|
| Linear | Stacking (TI) | 0.42 | 0.58 | 0.37 |
| Effects | Stacking (TD) | 0.45 | 0.61 | 0.39 |
| Non-Linear | Stacking (TI) | 3.49 | 2.08 | 3.69 |
| Effects | Stacking (TD) | 3.54 | 2.13 | 3.73 |

Table 3. Simulation results for Section 5.2 in the main paper ($n = 200$, $p = 80$ 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 1. 'Stacking (TI)' is stacking with time-independent weights, and 'Stacking (TD)' is stacking with time-dependent weights. The standard error for the estimate ISSE for each method in each scenario is less than 0.005.

| | Models | log-Normal | Weibull | Gamma |
|---|---|---|---|---|
| Linear | Stacking (TI) | 2.43 | 1.68 | 2.50 |
| Effects | Stacking (TD) | 2.43 | 1.68 | 2.50 |
| Non-Linear | Stacking (TI) | 1.97 | 1.00 | 2.04 |
| Effects | Stacking (TD) | 1.96 | 0.99 | 2.03 |

Fig. 3. The progession of integrated squared survival error (ISSE) as the number of irrelevant candidate survival models increases in the set of candidate survival models.

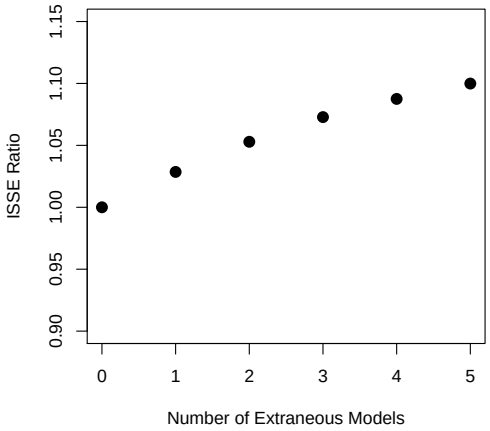


Table 4. Larger sample size results for Section 5.1 in the main paper ($n = 400$, $p = 8$ covariates, and 25% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 2000 times, and the error is multiplied by 10. The two top estimators 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. The standard error for the estimate ISSE for each method in each scenario is less than 0.01.

| | | Models | log-Normal | Weibull | Gamma |
|---|---|---|---|---|---|
| Linear Effects | Single Models | log-Normal | **0.17** | 0.53 | **0.17** |
| | | Weibull | 0.38 | **0.25** | 0.21 |
| | | Cox | 0.51 | 0.34 | 0.36 |
| | | RSF | 6.16 | 4.24 | 6.26 |
| | Flexible Models | Stacking | **0.20** | **0.29** | **0.18** |
| | | CV | 0.45 | 0.41 | 0.27 |
| Non-Linear Effects | Single Models | log-Normal | 4.54 | 2.17 | 4.87 |
| | | Weibull | 4.92 | **1.91** | 5.10 |
| | | Cox | 4.89 | 1.94 | 5.10 |
| | | RSF | **3.56** | 3.20 | **3.67** |
| | Flexible Models | Stacking | **2.91** | **1.70** | **3.08** |
| | | CV | 4.76 | 2.07 | 4.94 |

Table 5. Larger sample size results for Section 5.2 in the main paper ($n = 400$, $p = 80$ covariates, and 25% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 1000 times, and the error is multiplied by 1. 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. The standard error for the estimate ISSE for each method in each scenario is less than 0.005.

| | | Models | log-Normal | Weibull | Gamma |
|---|---|---|---|---|---|
| Linear Effects | Single Models | Cox - Lasso | **2.40** | **1.57** | **2.48** |
| | | Cox - Boosting | 2.60 | 1.72 | 2.66 |
| | | RSF | 2.44 | 1.84 | **2.48** |
| | Flexible Models | Stacking | **2.40** | **1.57** | **2.48** |
| | | CV | **2.40** | **1.57** | **2.48** |
| Non-Linear Effects | Single Models | Cox - Lasso | 1.94 | 0.96 | 2.01 |
| | | Cox - Boosting | 1.98 | **0.95** | 2.07 |
| | | RSF | **1.85** | 1.12 | **1.91** |
| | Flexible Models | Stacking | **1.91** | **0.94** | **1.98** |
| | | CV | 1.95 | 0.96 | 2.02 |

Table 6. Misspecified estimator of the censoring distribution for Section 5.1 in the main paper ($n = 200$, $p = 8$ 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 10. 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. The standard error for the ISSE estimate of the Weibull model is 0.04 in the gamma distributed scenario with linear covariate effects, otherwise the standard error for the estimate ISSE for each method in each scenario is less than 0.01.

| | | Models | log-Normal | Weibull | Gamma |
|---|---|---|---|---|---|
| Linear Effects | Single Models | log-Normal | **0.32** | 0.82 | **0.33** |
| | | Weibull | 0.48 | **0.51** | 0.41 |
| | | Cox | 0.74 | 0.68 | 0.67 |
| | | RSF | 8.13 | 5.43 | 8.46 |
| | Flexible Models | Stacking | **0.38** | **0.56** | **0.35** |
| | | CV | 0.58 | 0.73 | 0.50 |
| Non-Linear Effects | Single Models | log-Normal | 4.67 | 2.41 | 5.00 |
| | | Weibull | 5.12 | **2.10** | 5.32 |
| | | Cox | 5.03 | 2.17 | 5.27 |
| | | RSF | **4.61** | 3.60 | **4.88** |
| | Flexible Models | Stacking | **3.65** | **1.95** | **3.92** |
| | | CV | 4.96 | 2.27 | 5.20 |

Table 7. Misspecified estimator of the censoring distribution for Section 5.2 in the main paper ($n = 200$, $p = 80$ 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 1. 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. The standard error for the estimate ISSE for each method in each scenario is less than 0.004.

| | | Models | log-Normal | Weibull | Gamma |
|---|---|---|---|---|---|
| Linear Effects | Single Models | Cox - Lasso | **2.41** | **1.65** | **2.47** |
| | | Cox - Boosting | 2.55 | 1.72 | 2.61 |
| | | RSF | 2.49 | 1.86 | 2.54 |
| | Flexible Models | Stacking | **2.41** | **1.65** | **2.47** |
| | | CV | 2.43 | 1.67 | 2.49 |
| Non-Linear Effects | Single Models | Cox - Lasso | 1.99 | 1.02 | 2.05 |
| | | Cox - Boosting | 1.98 | **0.99** | 2.05 |
| | | RSF | **1.88** | 1.13 | **1.94** |
| | Flexible Models | Stacking | **1.95** | **0.99** | **2.02** |
| | | CV | 1.98 | 1.01 | 2.04 |

Table 8. Average elapsed time (in seconds) for the simulation scenarios presented in Section 5.1 in the main paper ($n = 200$, $p = 8$ covariates, and 25% censoring). Each simulation is replicated 2000 times. 'Time-Independent Stacking' is the average time for the entire stacking procedure for time-independent weights, 'Minimization' is the average time to solve the weighted least squares problem for only time-independent weights, and 'Time-Dependent Stacking' is the average time for the entire stacking procedure for time-dependent weights. The standard error for the average time for each method in each scenario is less than 1 second.

| | Models | log-Normal | Weibull | Gamma |
|---|---|---|---|---|
| Linear Effects | Time-Independent Stacking | 5 | 7 | 7 |
| | Minimization | 0 | 1 | 1 |
| | Time-Dependent Stacking | 40 | 38 | 35 |
| Non-Linear Effects | Time-Independent Stacking | 4 | 5 | 6 |
| | Minimization | 0 | 0 | 1 |
| | Time-Dependent Stacking | 35 | 38 | 37 |

Table 9. Average elapsed time (in seconds) for the simulation scenarios presented in Section 5.2 in the main paper ($n = 200$, $p = 80$ covariates, and 25% censoring). Each simulation is replicated 1000 times. 'Time-Independent Stacking' is the average time for the entire stacking procedure for time-independent weights, 'Minimization' is the average time to solve the weighted least squares problem for only time-independent weights, and 'Time-Dependent Stacking' is the average time for the entire stacking procedure for time-dependent weights. The standard error for the average time for each method in each scenario is less than 2 seconds.

| | Models | log-Normal | Weibull | Gamma |
|---|---|---|---|---|
| Linear | Time-Independent Stacking | 121 | 172 | 210 |
| Effects | Minimization | 0 | 1 | 1 |
| | Time-Dependent Stacking | 161 | 143 | 173 |
| Non-Linear | Time-Independent Stacking | 120 | 95 | 180 |
| Effects | Minimization | 1 | 0 | 1 |
| | Time-Dependent Stacking | 133 | 122 | 146 |

Table 10. The ratio of ISSE for the stacked survival model with different estimators for the RSF out-of-bag estimate. The numerator is the ISSE of the out-of-bag estimate from the `rsf` function, while the denominator is the ISSE of the five-fold cross-validation based estimate. Each simulation is replicated 2000 times. The standard error for the average ratio for each method in each scenario is less than 0.001.

| Covariate Dimension | Type of Effect | log-Normal | Weibull | Gamma |
|---|---|---|---|---|
| $p = 8$ | Linear | 1.00 | 1.00 | 1.00 |
| | Non-Linear | 0.99 | 1.01 | 0.99 |
| $p = 80$ | Linear | 1.00 | 1.00 | 1.00 |
| | Non-Linear | 0.99 | 1.00 | 0.99 |

Table 11. The ratio of computational time for the stacked survival model with different estimators for the RSF out-of-bag estimate. The numerator is the time of the out-of-bag estimate from the `rsf` function, while the denominator is the time of the five-fold cross-validation based estimate. Each simulation is replicated 2000 times. The standard error for the average ratio for each method in each scenario is less than 0.0007.

| Covariate Dimension | Type of Effect | log-Normal | Weibull | Gamma |
|---|---|---|---|---|
| $p = 8$ | Linear | 0.35 | 0.34 | 0.34 |
| | Non-Linear | 0.33 | 0.34 | 0.33 |
| $p = 80$ | Linear | 0.36 | 0.36 | 0.36 |
| | Non-Linear | 0.35 | 0.36 | 0.35 |

Table 12. The ratio of ISSE for the stacked survival model with different numbers of time points in the minimization procedure. The numerator is the ISSE for the estimator with 9 time points, while the denominator is the ISSE for the estimator with 19 time points. Each simulation has a sample size of 400. The $p = 8$ simulations are replicated 2000 times, while the $p = 80$ simulations are replicated 1000 times due to computational requirements. The standard error for the average ratio for each method in each scenario is less than 0.002.

| Covariate Dimension | Type of Effect | log-Normal | Weibull | Gamma |
|---|---|---|---|---|
| $p = 8$ | Linear | 1.01 | 1.01 | 1.01 |
| | Non-Linear | 1.00 | 1.01 | 1.00 |
| $p = 80$ | Linear | 1.00 | 1.00 | 1.00 |
| | Non-Linear | 1.00 | 1.00 | 1.00 |

Table 13. The ratio of elapsed time for the stacked survival model with different numbers of time points in the minimization procedure. The numerator is the ISSE for the estimator with 9 time points, while the denominator is the ISSE for the estimator with 19 time points. Each simulation has a sample size of 400. The $p = 8$ simulations are replicated 2000 times, while the $p = 80$ simulations are replicated 1000 times due to computational requirements. The standard error for the average ratio for each method in each scenario is less than 0.006.

| Covariate Dimension | Type of Effect | log-Normal | Weibull | Gamma |
|---|---|---|---|---|
| $p = 8$ | Linear | 0.53 | 0.52 | 0.54 |
| | Non-Linear | 0.55 | 0.54 | 0.54 |
| $p = 80$ | Linear | 0.60 | 0.54 | 0.61 |
| | Non-Linear | 0.60 | 0.59 | 0.60 |

Table 14. Simulation results for a high censoring scenario ($n = 1000$, $p = 8$ covariates, and 75% censoring) presented in integrated squared survival error (ISSE) over the observed support. Each simulation is replicated 1000 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. The standard error for the estimate ISSE for each method in each scenario is less than 0.004.

| | | Models | log-Normal | Weibull | Gamma |
|---|---|---|---|---|---|
| Linear Effects | Single Models | log-Normal | **0.09** | 0.28 | **0.10** |
| | | Weibull | 0.22 | **0.14** | 0.12 |
| | | Cox | 0.25 | 0.16 | 0.16 |
| | | RSF | 3.00 | 1.73 | 3.01 |
| | Flexible Models | Stacking | **0.11** | **0.15** | **0.10** |
| | | CV | 0.23 | 0.21 | 0.13 |
| Non-Linear Effects | Single Models | log-Normal | 2.75 | 0.69 | 2.76 |
| | | Weibull | 2.94 | **0.66** | 2.90 |
| | | Cox | 2.91 | 0.67 | 2.88 |
| | | RSF | **1.33** | 1.05 | **1.30** |
| | Flexible Models | Stacking | **1.26** | **0.57** | **1.23** |
| | | CV | 1.33 | 0.85 | **1.30** |

Table 15. Simulation results for non-monotonic covariate effects presented as integrated squared survival error (ISSE) over the observed support (see Section 5.7). Each simulation is replicated 2000 times. The error is multiplied by 10 for the scenarios with a 'modest covariate dimension' ($p = 8$), while the error is multiplied by 1 for the scenarios with a 'large covariate dimension' ($p = 80$). 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. The standard error for the estimate ISSE for each scenario with a 'modest covariate dimension' ($p = 8$) is less than 0.04, while the standard error for the estimate ISSE for each scenario with a 'large covariate dimension' ($p = 80$) is less than 0.005,.

| | | Models | log-Normal | Weibull | Gamma |
|---|---|---|---|---|---|
| Modest Covariate Dimension | Single Models | log-Normal | 10.4 | 4.67 | 11.2 |
| | | Weibull | 10.9 | 4.48 | 11.5 |
| | | Cox | 10.5 | 4.51 | 11.2 |
| | | RSF | **5.08** | **3.87** | **5.27** |
| | Flexible Models | Stacking | **5.09** | **3.52** | **5.27** |
| | | CV | 5.55 | 4.13 | 5.57 |
| Large Covariate Dimension | Single Models | Lasso | 1.89 | **0.87** | 1.98 |
| | | Boosting | 1.85 | 0.90 | 1.93 |
| | | RSF | **1.49** | 0.93 | **1.55** |
| | Flexible Models | Stacking | **1.55** | **0.83** | **1.60** |
| | | CV | 1.84 | 0.88 | 1.92 |