Supporting Information for "Estimating restricted mean treatment effects with stacked survival models"

Andrew Wey, David Vock, John Connett, and Kyle Rudser

Section 1 presents several extensions to the simulation study in the main paper, while Section 2 proves that the mean-squared error of the treatment specific conditional survival functions place an upper bound on the mean-squared error of the restricted mean treatment effect.

1 Furthering the Simulation Study

We present several extensions to the simulation study. Section 1.1 in the Supporting Information investigates the bias and MSE of the restricted mean estimators at a larger sample size ($n = 900$). Section 1.2 in the Supporting Information compares the bootstrap variance to the Monte Carlo variance. Section 1.3 in the Supporting Information considers different approaches (including a bias correction) to bootstrap estimation of confidence intervals. Section 1.4 in the Supporting Information investigates the influence of time point selection in minimizing the Brier Score for stacked survival models.

1.1 Larger Sample Size

We present the results of the simulation scenarios from the main paper with a larger sample size ($n = 900$ rather than $n = 300$) as the more robust approaches (i.e., splines and random

survival forests) should perform better with a larger sample size.

Tables 1 and 2 present the relative bias ($[E\hat{\gamma}(\tau) - \gamma(\tau)]/\gamma(\tau)$), the mean-squared error (MSE) ratio relative to the Cox estimator, and the integrated squared survival error (ISSE) ratio relative to the Cox estimator. For the exponential scenarios, both Stacked estimators show improved relative bias and MSE ratio compared to a smaller sample size. In addition, the Splines estimator performs relatively better in terms of MSE and, in particular, ISSE. In contrast, the Cox estimator maintains, as expected, the same relative bias in the exponential scenario with non-linear covariate effects despite the larger sample size. For the gamma scenarios, each estimator experiences a small increase in relative bias. This is likely due to the misspecification of each parametric and semi-parametric model. Both Stacked estimators, which perform very similarly at a larger sample size, still possess good MSE compared to the Cox model, while the Splines estimator improved the most in terms of the MSE and ISSE ratios compared to the scenarios with a smaller sample size.

The correlation between the mean-squared error of the conditional survival function (i.e., 'Integrated Squared Survival Error') with the MSE of the restricted mean treatment effect is stronger at a larger sample size, while the correlation is attenuated for the bias of the restricted mean treatment effect. In particular, Figure 1 illustrates that the Pearson correlation between ISSE and the MSE of the restricted mean treatment effect is close to 1, while the correlation between ISSE and the bias of the restricted mean treatment effect and ISSE is much higher at a larger sample size (~ 0.65 rather than ~ 0.20).

1.2 Standard Errors: Bootstrap versus MC

To help explain the poor performance of the confidence intervals for the Stacked estimator with random survival forests (RSF), we compare the bootstrapped variance to the expected variance from the Monte Carlo simulations (i.e., MC Var = MSE $-$ Bias^2). If the bootstrap variance is less than the Monte Carlo variance, then the bootstrap based confidence intervals

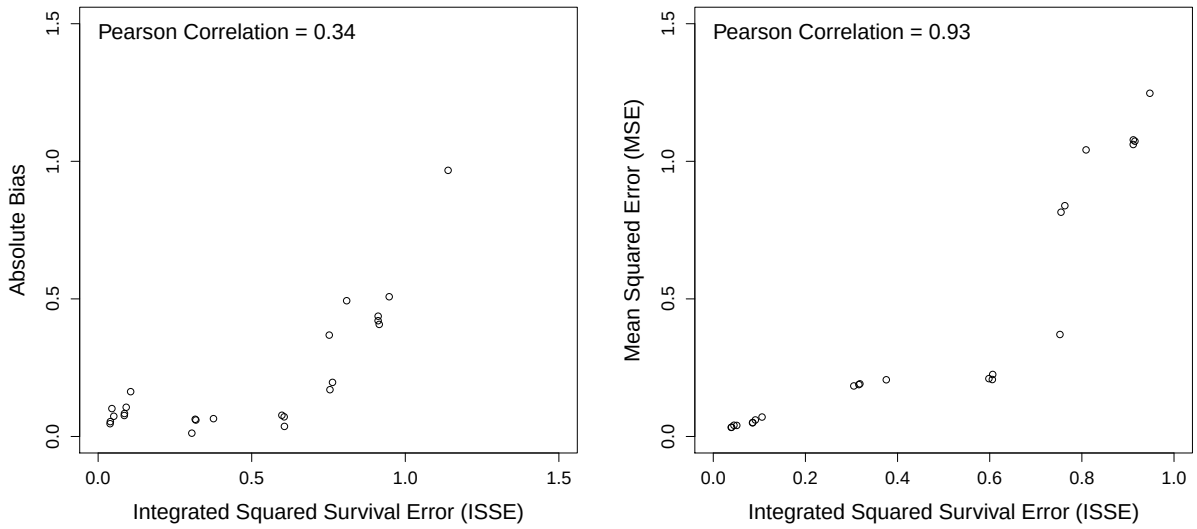
Table 1: Simulation results for the exponential distributed scenarios: $N = 900$, $N_{SIM} = 1000$, and a marginal censoring of 30%. 'Percent Relative Bias' is the ratio of the bias and true restricted mean difference. 'MSE Ratio' is the ratio of MSE relative to the Cox estimator. 'ISSE Ratio' is the ratio of integrated squared survival error, which corresponds to the mean-squared of the conditional survival function, relative to the Cox estimator.

	Estimator	Percent Relative Bias	MSE Ratio	ISSE Ratio
Linear $\gamma(20) = -2.965$	Cox	0	1.00	1.00
	Splines	-2	1.12	1.23
	Stacked	-2	1.03	1.04
	Stacked (with RSF)	-2	1.04	1.04
Non-Linear $\gamma(20) = 2.690$	Cox	9	1.00	1.00
	Splines	1	0.61	0.81
	Stacked	2	0.57	0.80
	Stacked (with RSF)	2	0.56	0.81
Linear $\gamma(50) = -12.318$	Cox	0	1.00	1.00
	Splines	0	1.08	1.12
	Stacked	0	0.99	1.00
	Stacked (with RSF)	1	1.01	0.99
Non-Linear $\gamma(50) = 7.929$	Cox	9	1.00	1.00
	Splines	2	0.58	0.83
	Stacked	3	0.58	0.82
	Stacked (with RSF)	2	0.56	0.82

Table 2: Simulation results for the gamma distributed scenarios: $N = 900$, $N_{SIM} = 1000$, and a marginal censoring of 30%. 'Percent Relative Bias' is the ratio of the bias and true restricted mean difference. 'MSE Ratio' is the ratio of MSE relative to the Cox estimator. 'ISSE Ratio' is the ratio of integrated squared survival error, which corresponds to the mean-squared of the conditional survival function, relative to the Cox estimator.

	Estimator	Percent Relative Bias	MSE Ratio	ISSE Ratio
Linear $\gamma(20) = -0.753$	Cox	-14	1.00	1.00
	Splines	-10	0.98	1.14
	Stacked	-6	0.81	0.87
	Stacked (with RSF)	-7	0.83	0.89
Non-Linear $\gamma(20) = -0.931$	Cox	-18	1.00	1.00
	Splines	-11	0.85	0.86
	Stacked	-8	0.70	0.81
	Stacked (with RSF)	-9	0.72	0.81
Linear $\gamma(50) = -6.599$	Cox	-8	1.00	1.00
	Splines	-6	1.03	1.13
	Stacked	-3	0.78	0.93
	Stacked (with RSF)	-3	0.81	0.94
Non-Linear $\gamma(50) = -6.407$	Cox	-15	1.00	1.00
	Splines	-8	0.68	0.83
	Stacked	-7	0.58	0.80
	Stacked (with RSF)	-7	0.57	0.80

Figure 1: An investigation into the relationship between restricted mean treatment effect performance and the quality of the conditional survival function estimation with a larger sample size ($n = 900$).



are more likely to perform poorly in terms of coverage probability. As illustrated in Table 3, the Stacked estimator with RSF possesses a bootstrapped variance up to 10% lower than the expected Monte Carlo variance for the gamma distributed scenarios (i.e., the situations that the Stacked estimator with RSF achieved less than 90% coverage).

1.3 Confidence Interval Construction

Due to the poor confidence interval performance of the Stacked estimator with random survival forests (RSF), we investigate two additional bootstrap approaches to confidence interval estimation. We first consider estimating the confidence intervals with a Normal approximation based only on $B = 50$ bootstrap replications. The second approach is the 'Bias-Corrected bootstrap' (referred to as 'BC Bootstrap'), which modifies the percentile based confidence intervals with a bias correction (Efron, 1981; Efron and Tibshirani, 1993).

Confidence interval performance is assessed with two measures: the ratio average confi-

Table 3: The ratio of the Monte Carlo variance and the bootstrapped variance for the simulation scenarios in the main paper (i.e., $\frac{\text{Boot Var}}{\text{MC Var}}$). In particular, a ratio greater than one indicates that the bootstrap overestimates the variance, while a ratio less than one indicates that the bootstrap underestimates the variance. $N = 300$ and $N_{SIM} = 1000$ with a marginal censoring of 30%.

		Exponential Scenarios		Gamma Scenarios	
	Estimator	Linear	Non-Linear	Linear	Non-Linear
$\tau = 20$	Cox	1.13	1.04	1.02	1.03
	Splines	1.22	1.11	1.15	1.12
	Stacked	1.18	1.13	1.26	1.24
	Stacked (with RSF)	1.01	0.89	1.13	1.08
$\tau = 50$	Cox	1.02	1.06	0.94	0.94
	Splines	1.07	1.10	0.95	0.95
	Stacked	1.09	1.13	1.05	1.01
	Stacked (with RSF)	0.84	0.77	0.88	0.84

dence interval length (ACL) compared to the Cox estimator and coverage probability. Tables 4 and 5 present the performance of the two confidence interval methods for the simulation scenarios presented in the main paper. As noted in Section 1.2, the Stacked estimator with random survival forests (RSF) occasionally possesses a smaller than expected bootstrap variance. Thus, the Normal based confidence intervals still fail to achieve nominal coverage. Although, the Normal based confidence intervals perform better than the percentile based approach from the main paper. The bias-corrected confidence intervals are associated with substantially worse performance than the percentile based approach for each estimator. For example, the Stacked estimator with RSF possesses coverage levels as low as 59%. Note that Efron and Tibshirani (1993) warn that bias-corrected bootstrap confidence intervals are potentially dangerous in practice.

Table 4: The confidence intervals results for the exponential distributed scenarios in the main paper: $N = 300$, $N_{SIM} = 1000$, and a marginal censoring of 30%. The 'Normal Bootstrap' estimates the variance with $B = 50$ bootstrap replicates, while the 'BC Bootstrap' estimates the confidence interval with $B = 300$ bootstrap replicates. 'ACL Ratio' is the ratio of average confidence interval lengths compared to the Cox estimator. The $\gamma(20)$ and $\gamma(50)$ are the true restricted mean treatment effects for $\tau = 20$ and $\tau = 50$, respectively.

		Normal Bootstrap		BC Bootstrap	
	Estimator	ACL Ratio	Cov.	ACL Ratio	Cov.
Linear $\gamma(20) = -2.965$	Cox	1.00	0.96	1.00	0.95
	Splines	1.12	0.96	1.12	0.95
	Stacked	1.05	0.95	1.03	0.91
	Stacked (with RSF)	0.97	0.95	0.91	0.75
Non-Linear $\gamma(20) = 2.690$	Cox	1.00	0.93	1.00	0.94
	Splines	1.04	0.95	1.05	0.95
	Stacked	0.98	0.95	0.98	0.92
	Stacked (with RSF)	0.86	0.92	0.81	0.63
Linear $\gamma(50) = -12.318$	Cox	1.00	0.95	1.00	0.94
	Splines	1.06	0.95	1.10	0.94
	Stacked	1.00	0.95	1.01	0.90
	Stacked (with RSF)	0.42	0.92	0.77	0.42
Non-Linear $\gamma(50) = 7.929$	Cox	1.00	0.91	1.00	0.93
	Splines	0.99	0.94	1.06	0.94
	Stacked	0.96	0.93	0.98	0.92
	Stacked (with RSF)	0.76	0.88	0.78	0.80

Table 5: The confidence interval results for the gamma distributed scenarios in the main paper: $N = 300$, $N_{SIM} = 1000$, and a marginal censoring of 30%. The 'Normal Bootstrap' estimates the variance with $B = 50$ bootstrap replicates, while the 'BC Bootstrap' estimates the confidence interval with $B = 300$ bootstrap replicates. 'ACL Ratio' is the ratio of average confidence interval lengths compared to the Cox estimator. The $\gamma(20)$ and $\gamma(50)$ are the true restricted mean treatment effects for $\tau = 20$ and $\tau = 50$, respectively.

		Normal Bootstrap		BC Bootstrap	
	Estimator	ACL Ratio	Cov.	ACL Ratio	Cov.
Linear $\gamma(20) = -0.753$	Cox	1.00	0.94	1.00	0.94
	Splines	1.18	0.96	1.20	0.92
	Stacked	1.05	0.97	1.05	0.84
	Stacked (with RSF)	1.02	0.96	0.94	0.63
Non-Linear $\gamma(20) = -0.931$	Cox	1.00	0.92	1.00	0.93
	Splines	1.15	0.95	1.15	0.93
	Stacked	1.04	0.97	1.04	0.87
	Stacked (with RSF)	0.99	0.95	0.90	0.59
Linear $\gamma(50) = -6.599$	Cox	1.00	0.92	1.00	0.93
	Splines	1.17	0.93	1.13	0.93
	Stacked	1.11	0.95	1.03	0.87
	Stacked (with RSF)	1.29	0.92	0.86	0.56
Non-Linear $\gamma(50) = -6.407$	Cox	1.00	0.90	1.00	0.92
	Splines	1.09	0.94	1.07	0.94
	Stacked	1.04	0.94	1.00	0.89
	Stacked (with RSF)	1.21	0.91	0.84	0.59

1.4 Stacking Question: Selection of t_s

As noted by Wey et al. (2013), the selection of time points for minimizing the Brier Score can have a substantial effect on the performance of stacked survival models. Wey et al. (2013) recommend minimizing over nine equally spaced quantiles of the observed event distribution. However, when estimating restricted mean treatment effects, stacked survival models may work better by restricting to time points within the support of interest. For example, in the simulation scenarios, the largest time point of interest is $\tau = 50$. Yet Table 6 shows that for the gamma distributed scenarios the recommended nine equally spaced quantiles of the observed event distribution results in approximately 40% of these points beyond $\tau = 50$.

Tables 7 and 8 compares the performance of stacking when the nine equally spaced quantiles of the observed event distribution are *restricted* to times less than $\tau = 50$ to the performance of stacking over nine equally spaced quantiles of the unrestricted observed event distribution. Restricting the time points has a negligible effect on bias and MSE, but it is associated with up to 9% larger ISSE in the gamma scenarios (i.e., the scenarios expected to benefit the most from restricting the event distribution).

Table 6: The average percent of t_s for the stacked survival model that occur beyond $\tau = 20$ and $\tau = 50$ for the simulation scenarios in the main paper. For each simulation iteration, the t_s were the nine equally spaced quantiles of the observed event distribution.

Distribution	Scenario	$\tau = 20$	$\tau = 50$
Exponential	Linear	58%	23%
	Non-Linear	31%	11%
Gamma	Linear	87%	40%
	Non-Linear	79%	41%

Table 7: Simulation results for restricting the minimization procedure for the exponential distributed scenarios: $N = 300$, $N_{SIM} = 1000$, and a marginal censoring of 30%. 'Relative Bias' is the ratio of the bias and true restricted mean difference. 'MSE Ratio' is the ratio of MSE relative to the 'Stacked' estimator. 'ISSE Ratio' is the ratio of integrated squared survival error, which corresponds to the mean-squared of the conditional survival function, relative to the 'Stacked' estimator. The 'Stacked' estimators use nine equally spaced quantiles of the unrestricted observed event distribution, while the 'Res. Stacked' estimators use nine equally spaced quantiles of the observed event distribution *restricted* to $\tau = 50$.

	Estimator	Relative Bias	MSE Ratio	ISSE Ratio
Linear $\gamma(20) = -2.965$	Stacked	-0.02	1.00	1.00
	Res. Stacked	-0.02	0.99	1.00
	Stacked (with RSF)	-0.01	1.00	1.00
	Res. Stacked (with RSF)	-0.01	0.98	1.00
Non-Linear $\gamma(20) = 2.690$	Stacked	0.05	1.00	1.00
	Res. Stacked	0.04	1.02	1.01
	Stacked (with RSF)	0.06	1.00	1.00
	Res. Stacked (with RSF)	0.05	1.00	1.01
Linear $\gamma(50) = -12.318$	Stacked	0.02	1.00	1.00
	Res. Stacked	0.01	1.00	1.01
	Stacked (with RSF)	0.03	1.00	1.00
	Res. Stacked (with RSF)	0.02	0.98	1.00
Non-Linear $\gamma(50) = 7.929$	Stacked	0.06	1.00	1.00
	Res. Stacked	0.07	1.01	1.01
	Stacked (with RSF)	0.06	1.00	1.00
	Res. Stacked (with RSF)	0.06	0.99	1.02

Table 8: Simulation results for the gamma distributed scenarios: $N = 300$, $N_{SIM} = 1000$, and a marginal censoring of 30%. 'Relative Bias' is the ratio of the bias and true restricted mean difference. 'MSE Ratio' is the ratio of MSE relative to the 'Stacked' estimator. 'ISSE Ratio' is the ratio of integrated squared survival error, which corresponds to the mean-squared of the conditional survival function, relative to the 'Stacked' estimator. The 'Stacked' estimators use nine equally spaced quantiles of the unrestricted observed event distribution, while the 'Res. Stacked' estimators use nine equally spaced quantiles of the observed event distribution *restricted* to $\tau = 50$.

	Estimator	Relative Bias	MSE Ratio	ISSE Ratio
	Stacked	-0.01	1.00	1.00
Linear	Res. Stacked	0.00	1.02	1.09
$\gamma(20) = -0.753$	Stacked (with RSF)	-0.04	1.00	1.00
	Res. Stacked (with RSF)	-0.03	0.99	1.08
	Stacked	-0.03	1.00	1.00
Non-Linear	Res. Stacked	-0.03	1.01	1.02
$\gamma(20) = -0.931$	Stacked (with RSF)	-0.07	1.00	1.00
	Res. Stacked (with RSF)	-0.07	1.02	1.06
	Stacked	-0.01	1.00	1.00
Linear	Res. Stacked	-0.02	1.01	1.04
$\gamma(50) = -6.599$	Stacked (with RSF)	-0.02	1.00	1.00
	Res. Stacked (with RSF)	-0.03	1.01	1.05
	Stacked	-0.06	1.00	1.00
Non-Linear	Res. Stacked	-0.06	1.03	1.03
$\gamma(50) = -6.407$	Stacked (with RSF)	-0.07	1.00	1.00
	Res. Stacked (with RSF)	-0.08	1.02	1.05

2 Influence of the Conditional Survival Function

This section proves that the mean-squared error of the treatment specific conditional survival functions places an upper bound on the MSE of the restricted mean treatment effect.

Similar to Wey et al. (2013), we define the mean-squared error for a conditional survival function estimator of the a^{th} treatment as the integral of the squared error at time t over $(0, \tau)$:

$$\text{MSE}_\tau\{\hat{S}^{(a)}(\cdot|\boldsymbol{x})\} = E\int_0^\tau [\hat{S}^{(a)}(t|\boldsymbol{x}) - S^{(a)}(t|\boldsymbol{x})]^2 dt.$$

Note that the expectation is with respect to the covariate distribution and the sampling distribution of the estimator. A significant difference between this investigation and Wey et al. (2013) is the addition of treatment a. Since restricted mean treatment effect estimation requires two conditional survival functions (one for each treatment), we take the simple average of the mean-squared error for both treatments. That is, the main paper uses

$$\text{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\} \;=\; \frac{1}{2} \times \{\text{MSE}_\tau\{\hat{S}^{(0)}(\cdot|\boldsymbol{x})\} + \text{MSE}_\tau\{\hat{S}^{(1)}(\cdot|\boldsymbol{x})\}\}, \tag{1}$$

as the mean-squared error for an estimator of the conditional survival function. We can then show that the mean squared error of restricted mean treatment effect is bounded by the mean-squared error of the conditional survival function:

Theorem 1. *Let the mean squared error of a restricted mean treatment effect be* $MSE[\hat{\gamma}(\tau)] = E\{\hat{\gamma}(\tau) - \gamma(\tau)\}^2$, *then*

$$MSE[\hat{\gamma}(\tau)] \;\leq\; 2\tau \times MSE_\tau\{\hat{S}(\cdot|\boldsymbol{x})\}.$$

The result - which is a consequence of a sequential application of Jensen's, the triangle, and

Schwarz inequalities - helps justify the strong association of the restricted mean squared error with the performance of the conditional survival function estimator. The bias is also bounded, but the limit is less tight due to a positive variance term. This results in a less strong, but still positive, association of bias with the mean-squared error of the conditional survival function.

Proof: We need to first make a distinction between the sampling distribution for the estimator of the conditional survival function [which we call the 'learning sample' (LS) distribution] and the covariate distribution $\boldsymbol{X}$. It is important to note that the learning sample distribution is independent of the covariate distribution (and the survival time distribution).

$$
\begin{aligned}
E\{\hat{\gamma}(\tau) - \gamma(\tau)\}^2 &= E_{LS}\left\{ E_{\boldsymbol{X}|LS} \int_0^\tau [\hat{S}^{(1)}(t|\boldsymbol{x}) - \hat{S}^{(0)}(t|\boldsymbol{x})]dt - \right. \\
&\qquad \left. E_{\boldsymbol{X}|LS} \int_0^\tau [S^{(1)}(t|\boldsymbol{x}) - S^{(0)}(t|\boldsymbol{x})]dt \right\}^2 \\
&\leq E_{LS} E_{\boldsymbol{X}|LS} \left\{ \int_0^\tau [\hat{S}^{(1)}(t|\boldsymbol{x}) - S^{(1)}(t|\boldsymbol{x})]dt + \right. \qquad (2) \\
&\qquad \left. \int_0^\tau [S^{(0)}(t|\boldsymbol{x}) - \hat{S}^{(0)}(t|\boldsymbol{x})]dt \right\}^2 \\
&\leq E_{LS,\boldsymbol{X}} \left(\{ \int_0^\tau [\hat{S}^{(1)}(t|\boldsymbol{x}) - S^{(1)}(t|\boldsymbol{x})]dt \}^2 + \qquad (3) \right. \\
&\qquad \left. \{ \int_0^\tau [S^{(0)}(t|\boldsymbol{x}) - \hat{S}^{(0)}(t|\boldsymbol{x})]dt \}^2 \right) \\
&\leq \tau \times E_{LS,\boldsymbol{X}} \left(\int_0^\tau [\hat{S}^{(1)}(t|\boldsymbol{x}) - S^{(1)}(t|\boldsymbol{x})]^2 dt + \qquad (4) \right. \\
&\qquad \left. \int_0^\tau [S^{(0)}(t|\boldsymbol{x}) - \hat{S}^{(0)}(t|\boldsymbol{x})]^2 dt \right) \\
&= \tau \times \{ \mathrm{MSE}_\tau\{\hat{S}^{(0)}(\cdot|\boldsymbol{x})\} + \mathrm{MSE}_\tau\{\hat{S}^{(1)}(\cdot|\boldsymbol{x})\} \} \\
&= 2\tau \times \mathrm{MSE}_\tau\{\hat{S}(\cdot|\boldsymbol{x})\},
\end{aligned}
$$

where line (2) holds by Jensen's inequality, line (3) holds by the triangle inequality, and line (4) holds by Schwarz's inequality.

References

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics* **9,** 139–172.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman and Hall.

Wey, A., Connett, J., and Rudser, K. (2013). Stacking survival models. arXiv:1309.7936v3.