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

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

Section 1 provides further analyses for the lung transplant example. Section 2 briefly discusses the consistency of stacked survival models under a misspecified censoring distribution. Section 3 presents numerous extensions to the simulation study in the main paper, while Section 4 proves the inequality that places an upper bound on the mean-squared error of the restricted mean treatment effect (RMTE).

1 Additional Lung Transplantation Analyses

We present additional analyses regarding the lung transplantation example from the main paper. Table 1 presents descriptive statistics for the covariates stratified by low volume and high volume centers, which demonstrates the need for an adjusted analysis due to the many covariate imbalances.

We define the restricted mean $R^2(\tau)$ value as

$$\hat{R}^2(\tau) \;=\; 1 - \frac{\widehat{SS}_{res}(\tau)}{\widehat{SS}_{tot}(\tau)}, \tag{1}$$

where $\widehat{SS}_{res}(\tau)$ and $\widehat{SS}_{tot}(\tau)$ are censored-data extensions of, respectively, the residual sum

of squares and total sum of squares based on inverse probability-of-censoring weighting:

$$\widehat{SS}_{res}(\tau) = \sum_{i \epsilon \Gamma_a} \frac{\tilde{\Delta}_i(\tau)}{G(\min\{t_i, \tau\})} \times \left\{ \min(t_i, \tau) - \int_0^\tau \hat{S}(t|\boldsymbol{x}_i, a_i) dt \right\}^2 \quad (2)$$

$$\widehat{SS}_{tot}(\tau) = \sum_{i \epsilon \Gamma_a} \frac{\tilde{\Delta}_i(\tau)}{G(\min\{t_i, \tau\})} \times \left\{ \min(t_i, \tau) - \int_0^\tau \hat{S}(t|a_i) dt \right\}^2, \quad (3)$$

where $\tilde{\Delta}_i(\tau) = I(\min\{t_i, \tau\} \leq c_i)$, $\int_0^\tau \hat{S}(t|\boldsymbol{x}_i, a_i)dt$ is the estimated restricted mean based on the conditional survival function estimator for treatment a_i, and $\int_0^\tau \hat{S}(t|a_i)dt$ is the estimated treatment-specific restricted mean based on the treatment-specific Kaplan-Meier survival curve. Recall that Γ_a is the set of patients on treatment a. This helps emphasize that, in our situation, the $R^2(\tau)$ depends on the given treatment-specific conditional survival function.

Table 2 presents the $R^2(\tau)$ value for each treatment-specific survival model included in the stack. In general, the restricted mean $R^2(\tau)$ value and, thus, the explained variation is low for each treatment-specific survival model. However, explained variation is neither a sufficient nor necessary condition for correcting imbalances between two non-randomized groups. For example, a correctly specified model can explain a low proportion of variation due to a large error term, which plays a significant role in explained variation measures such as R^2 (Royston, 2006).

2 Misspecification of Censoring Distribution

In the main paper, we note that stacked survival models are, under certain conditions, consistent even when the censoring distribution is misspecified. In this section, we sketch the necessary conditions and arguments to show consistency. In this situation, we estimate the censoring distribution with a Kaplan-Meier, which we assume converges to a constant function even for situations when the censoring distribution depends on the covariates; that is, $\hat{G}(t) \to G(t)$ for all $t \, \epsilon \, (0, \tau)$ and there exists a $\delta > 0$ such that $G(\tau) > \delta$. Wey et al. (2015)

show that the Brier Score approaches its expectation. Note that $E[\frac{\Delta(t_r)}{G(T(t_r))}|T, \boldsymbol{x}] = \frac{1}{G(T(t_r))} \times E[\Delta(t_r)|T, \boldsymbol{x}] = \frac{G(T(t_r)|\boldsymbol{x})}{G(T(t_r))}$, which is finite and bounded away from zero. Let $C(\boldsymbol{x}) \equiv \frac{G(T(t_r)|\boldsymbol{x})}{G(T(t_r))}$, then the asymptotic minimization problem becomes

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} E_{\boldsymbol{x}} \left\{ C(\boldsymbol{x}) \sum_{r=1}^{s} [S_o(t_r|\boldsymbol{x}) - \{S_o(t_r|\boldsymbol{x}) \sum_{k=1}^{l} \alpha_k + \sum_{k=l+1}^{m} \alpha_k S_k(t_r|\boldsymbol{x})\}]^2 \right\}, \quad (4)$$

where the $S_k(t|\boldsymbol{x})$ for $k = 1, ..., l$ are the correctly specified survival models. Equation 4 minimizes the expected value of a non-negative function and, therefore, is minimized when $\sum_{k=1}^{l} \alpha_k = 1$, which holds under the identifiability constraint outlined in Wey et al. (2015); see assumption A3.

3 Furthering the Simulation Study

We present several extensions to the simulation study. Section 3.1 investigates the bias and MSE of the restricted mean estimators in simulations at a larger sample size. Section 3.2 investigates the influence of time point selection in minimizing the Brier Score for stacked survival models. Lastly, Section 3.3 presents simulation scenarios with covariate independent censoring rather than covariate dependent censoring.

3.1 Larger Sample Sizes

We evaluate the performance of the simulation scenarios at a larger sample size ($n = 900$). The idea is that a larger sample size may improve the performance of the semi-parametric and non-parametric models included in the set of candidate survival models.

Tables 3 and 4 present the bias and MSE for the exponential and gamma distributed scenarios presented in the main paper at a sample size of $n = 900$. The Stacked estimator with and without RSF generally performs relatively well although there is usually an

individual estimator that performs better for a given scenario (e.g., the Weibull estimator in the exponential scenario with linear covariate effects). Finally, while the RSF estimator performed surprisingly well for estimating the restricted mean treatment effect in the main paper, the relative performance of the RSF actually worsened particularly for the exponential scenarios. Despite being a non-parametric estimator of a conditional survival function, the ISSE for RSF surprisingly worsened relative to the Cox model with larger sample sizes compared to the scenarios presented in the main paper. It is possible that RSF requires that the minimum number of observations in each node increases to ensure better performance with larger sample sizes.

3.2 Stacking Question: Selection of t_s

As noted by Wey et al. (2015), 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. (2015) found minimizing over nine equally spaced quantiles of the observed event distribution performed well for estimating the conditional survival function. However, when estimating restricted mean treatment effects, stacked survival models may perform better by restricting to time points within the support of interest, i.e., between 0 and τ. For example, in the simulation scenarios, the largest time point of interest is $\tau = 50$. Yet Table 5 shows that, for the gamma distributed scenarios, nine equally spaced quantiles of the observed event distribution results in approximately 40% of these points beyond $\tau = 50$. We evaluate the performance of stacking when the time points are restricted to less than $\tau = 50$ for the exponential and gamma distributed scenarios in the main paper.

Tables 6 and 7 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 had a negligible effect on bias and MSE, but it is

associated with up to 5% larger ISSE in the gamma scenarios.

3.3 Covariate Independent Censoring

The censoring distribution in the simulation study of the main paper is covariate-dependent, and the Stacked estimator performs well despite a misspecified Kaplan-Meier estimator of the censoring distribution. This section evaluates the performance of the Stacked estimator in the presence of covariate-independent censoring for the exponential and gamma distributed scenarios, which implies a correctly specified Kaplan-Meier estimator of the censoring distribution. For the linear scenarios, the intercept of the censoring distribution ensures a similar censoring rate as the covariate dependent censoring scenarios. In contrast, for the non-linear scenarios, the intercept of the censoring distributions have to be modified to ensure similar censoring rates. In particular, the parameter values for the censoring distributions are defined as: $\gamma_2^0 = -5$ and $\gamma_2^1 = -6$ for the exponential non-linear scenarios, and $\gamma_4^0 = 3.9$ and $\gamma_4^1 = 3.4$ for the gamma non-linear scenarios.

Tables 8 and 9 present the covariate independent censoring results for, respectively, the exponential and gamma distributed scenarios. The Stacked estimators are competitive in each scenario regardless of the distribution or functional form of the covariates, which is qualitatively similar to the simulation scenarios in the main paper. The bias is surprisingly large in the non-linear scenarios for every estimator although the Stacked estimators remain competitive with the Splines estimator in both scenarios.

4 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. (2015), 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 $(0, \tau)$:

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

where 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. (2015) is the addition of the treatment indicator a. We can then show that the mean squared error of restricted mean treatment effect is bounded by the mean-squared error of the treatment-specific conditional survival functions:

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*

$$MSE[\hat{\gamma}(\tau)] \;\leq\; \tau \times \Big[MSE_\tau\{\hat{S}^{(0)}\} + MSE_\tau\{\hat{S}^{(1)}\} - 2 \times [Cov_\tau\{\hat{S}^{(0)}, \hat{S}^{(1)}\} + Bias_\tau\{\hat{S}^{(0)}, \hat{S}^{(1)}\}]\Big].$$

The result - which is a consequence of a sequential application of Jensen and Schwarz inequalities - helps justify the strong association of the restricted mean treatment effect 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 distribu-

tion is independent of the covariate distribution (and the survival time distribution).

$$
\begin{aligned}
E\{\hat{\gamma}(\tau)-\gamma(\tau)\}^2 &= E_{LS}\left\{E_{\boldsymbol{X}|LS}\int_0^\tau[\hat{S}^{(1)}(t|\boldsymbol{x})-\hat{S}^{(0)}(t|\boldsymbol{x})]dt- \right. \\
&\qquad \left. E_{\boldsymbol{X}|LS}\int_0^\tau[S^{(1)}(t|\boldsymbol{x})-S^{(0)}(t|\boldsymbol{x})]dt\right\}^2 \\
&\leq E_{LS}E_{\boldsymbol{X}|LS}\left\{\int_0^\tau[\hat{S}^{(1)}(t|\boldsymbol{x})-S^{(1)}(t|\boldsymbol{x})+ \right. \qquad (5)\\
&\qquad \left. S^{(0)}(t|\boldsymbol{x})-\hat{S}^{(0)}(t|\boldsymbol{x})]dt\right\}^2 \\
&\leq \tau\times E_{LS,\boldsymbol{X}}\int_0^\tau\left[\hat{S}^{(1)}(t|\boldsymbol{x})-S^{(1)}(t|\boldsymbol{x})+ \right. \qquad (6)\\
&\qquad \left. S^{(0)}(t|\boldsymbol{x})-\hat{S}^{(0)}(t|\boldsymbol{x})\right]^2 dt \\
&= \tau\times E_{LS,\boldsymbol{X}}\int_0^\tau\left[[\hat{S}^{(1)}(t|\boldsymbol{x})-S^{(1)}(t|\boldsymbol{x})]^2+ \right. \\
&\qquad [S^{(0)}(t|\boldsymbol{x})-\hat{S}^{(0)}(t|\boldsymbol{x})]^2- \\
&\qquad \left. 2\times[\hat{S}^{(1)}(t|\boldsymbol{x})-S^{(1)}(t|\boldsymbol{x})][\hat{S}^{(0)}(t|\boldsymbol{x})-S^{(0)}(t|\boldsymbol{x})]\right]dt \\
&= \tau\times\left\{\mathrm{MSE}_\tau\{\hat{S}^{(0)}\}+\mathrm{MSE}_\tau\{\hat{S}^{(1)}\}-2\times[\mathrm{Cov}_\tau\{\hat{S}^{(0)},\hat{S}^{(1)}\}+ \right. \\
&\qquad \left.\mathrm{Bias}_\tau\{\hat{S}^{(0)},\hat{S}^{(1)}\}]\right\},
\end{aligned}
$$

where the last line holds by adding and subtracting the expectation of the appropriate conditional survival function, line (5) holds by Jensen's inequality, and line (6) holds by Schwarz's inequality. Finally, $\mathrm{Cov}_\tau\{\hat{S}^{(0)},\hat{S}^{(1)}\}=E_{LS,\boldsymbol{X}}\int_0^\tau[\hat{S}^{(1)}(t|\boldsymbol{x})-E_{LS|\boldsymbol{X}}\hat{S}^{(1)}(t|\boldsymbol{x})][\hat{S}^{(0)}(t|\boldsymbol{x})-E_{LS|\boldsymbol{X}}\hat{S}^{(0)}(t|\boldsymbol{x})]$ and $\mathrm{Bias}_\tau\{\hat{S}^{(0)},\hat{S}^{(1)}\}=E_{LS,\boldsymbol{X}}\int_0^\tau[E_{LS|\boldsymbol{X}}\hat{S}^{(1)}(t|\boldsymbol{x})-S^{(1)}(t|\boldsymbol{x})][E_{LS|\boldsymbol{X}}\hat{S}^{(0)}(t|\boldsymbol{x})-S^{(0)}(t|\boldsymbol{x})]$.

References

Royston, P. (2006). Explained variation for survival models. *The Stata Journal* **6,** 83–96.

Wey, A., Connett, J., and Rudser, K. (2015). Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics* **16,** 537–549.

Table 1: Descriptives for the covariate included in the lung transplantation analysis by low volume and high volume centers. Continuous variables are summarized by means and standard deviations, while categorical variables are presented as frequencies and percents.

Covariate	Low Volume Centers ($n = 3646$)	High Volume Centers ($n = 1853$)
Single Transplant	1248 (34.2%)	536 (28.9%)
Male	2151 (59.0%)	1131 (61.0%)
Donor Age > 55 years	253 (6.9%)	200 (10.8%)
Disease Grouping		
Obstructive	1254 (34.4%)	567 (30.6%)
Vascular	100 (2.7%)	79 (4.2%)
Cystic	468 (12.8%)	196 (10.6%)
Restrictive	1824 (50.0%)	1011 (54.6%)
African American Donor*	689 (18.9%)	384 (20.7%)
Smoking Donor	330 (9.1%)	251 (13.5%)
Ventilation Status	569 (15.6%)	343 (18.5%)
Recipient Age (Years)	54.4 (12.9)	56.6 (13.1)
Lung Allocation Score	45.6 (15.9)	47.8 (17.2)
Height Difference†	-1.9 (9.3)	-1.1 (9.6)
Six Minute Walk	723 (420)	477 (477)
O_2 Use	6.0 (5.6)	6.5 (5.8)

*Versus non-African American donors

†Between donor and recipient

Table 2: The estimated $R^2(\tau = 3)$ value for the models in each treatment-specific (i.e., center volume-specific) stacked survival model for the lung transplant example. The $R^2(\tau)$ value is generalization of the R^2 value from linear regression [see Equation (1)].

Model	Low Volume Centers	High Volume Centers
Cox	0.036	0.034
Cox with Splines	0.049	0.066
log-Normal	0.040	0.023
Weibull	0.045	0.036

Table 3: Simulation results for the exponential distributed scenarios: $N = 900$, $N_{SIM} = 1000$, and a marginal censoring rate of 30% for the linear scenario and 20% for the non-linear scenario. '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(0)' and 'ISSE(1)' are the treatment-specific ratios of integrated squared survival error, which corresponds to the mean-squared of the treatment-specific conditional survival function, relative to the Cox estimator. $\gamma(50)$ is the true restricted mean treatment effects for $\tau = 50$.

Scenario	Estimator	Percent Relative Bias	MSE Ratio	ISSE(0)	ISSE(1)
Linear $\gamma(50) = -12.318$	Cox	0%	1.00	1.00	1.00
	Splines	0%	1.09	4.83	4.56
	log-Normal	3%	1.05	2.34	1.83
	Weibull	0%	0.92	0.88	0.89
	RSF	18%	4.31	14.38	11.73
	Stacked with RSF	1%	0.97	1.26	1.13
	Stacked without RSF	0%	0.94	1.13	1.04
Non-Linear $\gamma(50) = -10.334$	Cox	-9%	1.00	1.00	1.00
	Splines	2%	0.58	0.55	0.57
	log-Normal	4%	0.63	1.13	1.15
	Weibull	8%	0.86	0.96	0.98
	RSF	-12%	1.06	2.37	2.15
	Stacked with RSF	2%	0.58	0.45	0.49
	Stacked without RSF	3%	0.57	0.45	0.48

Table 4: Simulation results for the gamma distributed scenarios: $N = 900$, $N_{SIM} = 1000$, and a marginal censoring rate of 30% for the linear scenario and 30% for the non-linear scenario. '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(0)' and 'ISSE(1)' are the treatment-specific ratios of integrated squared survival error, which corresponds to the mean-squared of the treatment-specific conditional survival function, relative to the Cox estimator. $\gamma(50)$ is the true restricted mean treatment effects for $\tau = 50$.

Scenario	Estimator	Percent Relative Bias	MSE Ratio	ISSE(0)	ISSE(1)
Linear $\gamma(50) = -6.599$	Cox	-7%	1.00	1.00	1.00
	Splines	-6%	1.04	4.38	4.40
	log-Normal	-4%	0.92	1.22	1.43
	Weibull	2%	0.66	0.80	0.92
	RSF	-1%	0.85	8.61	14.13
	Stacked with RSF	-1%	0.73	1.01	1.14
	Stacked without RSF	-1%	0.74	0.96	1.03
Non-Linear $\gamma(50) = -6.407$	Cox	-15%	1.00	1.00	1.00
	Splines	-8%	0.67	0.91	0.61
	log-Normal	-7%	0.57	0.86	0.96
	Weibull	0%	0.40	0.95	0.97
	RSF	3%	0.47	1.66	2.10
	Stacked with RSF	-6%	0.55	0.68	0.51
	Stacked without RSF	-6%	0.57	0.69	0.50

Table 5: The number of t_s for the stacked survival model that occur beyond $\tau = 50$ 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.

	Number	Exponential		Gamma	
	Past τ	Linear	Non-Linear	Linear	Non-Linear
$\tau = 50$	0	0	4	0	0
	1	19	957	0	0
	2	880	39	2	0
	3	101	0	405	316
	4	0	0	589	679
	5	0	0	4	5

Table 6: Simulation results for restricting the minimization procedure for the exponential distributed scenarios: $N = 300$, $N_{SIM} = 1000$, and a marginal censoring rate of 30% for the linear scenario and 20% for the non-linear scenario. '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 average mean-squared across the treatment-specific conditional survival functions, relative to the 'Stacked' estimator. The 'Equally Spaced' estimators use nine equally spaced quantiles of the unrestricted observed event distribution, while the 'Restricted < 50' estimators use nine equally spaced quantiles of the observed event distribution *restricted* to $\tau = 50$.

	Estimator	Relative Bias	MSE Ratio	ISSE Ratio
Linear $\gamma(50) = -12.318$	Equally Spaced	0.02	1.00	1.00
	Restricted < 50	0.01	1.00	1.01
	Equally Spaced	0.03	1.00	1.00
	Restricted < 50	0.02	0.98	1.00
Non-Linear $\gamma(50) = 7.929$	Equally Spaced	0.06	1.00	1.00
	Restricted < 50	0.07	1.01	1.01
	Equally Spaced	0.06	1.00	1.00
	Restricted < 50	0.06	0.99	1.02

Table 7: Simulation results for the gamma distributed scenarios: $N = 300$, $N_{SIM} = 1000$, and a marginal censoring rate of 30% for the linear scenario and 30% for the non-linear scenario. '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 average mean-squared across the treatment-specific conditional survival functions, relative to the 'Stacked' estimator. The 'Equally Spaced' estimators use nine equally spaced quantiles of the unrestricted observed event distribution, while the 'Restricted < 50' estimators use nine equally spaced quantiles of the observed event distribution *restricted* to $\tau = 50$.

	Estimator	Relative Bias	MSE Ratio	ISSE Ratio
Linear $\gamma(50) = -6.599$	Equally Spaced	-0.01	1.00	1.00
	Restricted < 50	-0.02	1.01	1.04
	Equally Spaced	-0.02	1.00	1.00
	Restricted < 50	-0.03	1.01	1.05
Non-Linear $\gamma(50) = -6.407$	Equally Spaced	-0.06	1.00	1.00
	Restricted < 50	-0.06	1.03	1.03
	Equally Spaced	-0.07	1.00	1.00
	Restricted < 50	-0.08	1.02	1.05

Table 8: Simulation results for covariate independent censoring with the exponential distributed scenarios: $N = 300$, $N_{SIM} = 1000$, and a marginal censoring rate of 30% for the linear scenario and 20% for the non-linear scenario. '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. The $\gamma(50)$ is the true restricted mean treatment effect $\tau = 50$.

	Estimator	Percent Relative Bias	MSE Ratio
Linear $\gamma(50) = -12.318$	Cox	1%	1.00
	Splines	1%	1.18
	Stacked	2%	0.99
	Stacked (with RSF)	3%	1.01
Non-Linear $\gamma(50) = 7.929$	Cox	42%	1.00
	Splines	34%	0.78
	Stacked	38%	0.85
	Stacked (with RSF)	37%	0.84

Table 9: Simulation results for covariate independent censoring with the gamma distributed scenarios: $N = 300$, $N_{SIM} = 1000$, and a marginal censoring rate of 30% for the linear scenario and 30% for the non-linear scenario. '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. The $\gamma(50)$ is the true restricted mean treatment effect $\tau = 50$.

	Estimator	Percent Relative Bias	MSE Ratio
Linear $\gamma(50) = -6.599$	Cox	-6%	1.00
	Splines	-5%	1.22
	Stacked	-2%	0.95
	Stacked (with RSF)	-3%	0.95
Non-Linear $\gamma(50) = -6.407$	Cox	-28%	1.00
	Splines	-13%	0.70
	Stacked	-20%	0.75
	Stacked (with RSF)	-19%	0.72