Supplementary Web Appendix for:
Censoring Unbiased Regression Trees and Ensembles

References to Sections preceded by “S.” and figures, tables, theorems and equations preceded by “S-” are internal to this supplement; all other references refer to the main paper. Example `R` code implementing the *CURE*–L_2 algorithms in the case of the nonparametric bootstrap for the Copenhagen Stroke study is also included in a separate file as additional supplementary material. Other code is available upon request from Dr. Jon Steingrimsson.

The Supplementary Web Appendix is organized as follows.

1. Section S.1 describes the *CURE*-L_2 algorithms using the Brier loss.

2. Section S.2 presents additional simulation results.

3. Section S.3 presents further details on OOB-Based variable importance measures.

4. Section S.4 gives additional results related to the data analysis section.

5. Sections S.5-S.7 give detailed proofs of results from the main paper.

S.1 The CURE$-L_2$ Algorithm With Brier Loss

Let $L_{t;2}(T,\psi_t(W)) = (I(T \geq t) - \psi_t(W))^2$ denote the (full data) Brier loss function; see Section 3.3. Minimizing this squared error loss function directly induces a simple estimator for $S_0(t|W)$. The IPCW Brier loss function

$$L_{t;2,ipcw}(O,\psi_t;G) = \frac{\Delta(I(\tilde{T} \geq t) - \psi_t(W))^2}{G(\tilde{T}|W)}$$

is a CUT for $L_{t;2}(T,\psi_t(W))$ when $G(\cdot|\cdot) = G_0(\cdot|\cdot)$, and leads to one possible observed data estimator for $S_0(t|W)$ when integrated into a tree or ensemble procedure.

Graf et al. (1999) proposed a "time-dependent" Brier loss function with censored outcome data that also requires the specification (or estimation) of $G_0(\cdot|\cdot)$. Calculations similar to Lostritto et al. (2012) show that this loss function is constructed from terms of the form

$$L_{t;2,ipcw}(O(t),\psi_t;G) = \frac{\Delta(t)(I(\tilde{T}(t) \geq t) - \psi_t(W))^2}{G(\tilde{T}(t)|W)} \equiv \frac{\Delta(t)(I(\tilde{T} \geq t) - \psi_t(W))^2}{G(\tilde{T}(t)|W)},$$

where $O(t) = (\tilde{T}(t), \Delta(t), W')'$, $\tilde{T}(t) = \min(T(t), C)$ and $\Delta(t) = I(T(t) \leq C)$ and the stated equivalence follows from the fact that $L_{t;2}(T(t),\psi_t(W)) = L_{t;2}(T,\psi_t(W))$. Similarly to $L_{t;2,ipcw}(O,\psi_t;G)$, straightforward calculations show that $L_{t;2,ipcw}(O(t),\psi_t;G)$ is a CUT for $L_{t;2}(T,\psi_t(W))$ when $G(\cdot|\cdot) = G_0(\cdot|\cdot)$. The value of $I(T \geq t)$ can be unambiguously determined when $\Delta(t) = 1$, which occurs if either $\Delta = 1$ or if $\Delta = 0$ and $\tilde{T} \geq t$. As a result, $L_{t;2,ipcw}(O(t),\psi_t;G)$ uses more of the available information in the data for estimating $S_0(t|W)$ when compared to $L_{t;2,ipcw}(O,\psi_t;G)$.

Mathematically, the equivalence $L_{t;2}(T,\psi_t(W)) = L_{t;2}(T(t),\psi_t(W))$ combined with the construction of $L_{t;2,ipcw}(O(t),\psi_t;G)$ suggests applying (6) to the observed data structure $O(t)$ with $L(Z,\psi_t(W)) = L_{t;2}(T(t),\psi_t(W))$; simplifying the resulting expression, we obtain

$$\frac{\Delta(t)(I(\tilde{T} \geq t) - \psi_t(W))^2}{G(\tilde{T}(t)|W)} + \frac{(1-\Delta(t))m_{t;2}(\tilde{T}(t),W;S)}{G(\tilde{T}(t)|W)} - \int_0^{\tilde{T}(t)} \frac{m_{t;2}(u,W;S)}{G(u|W)} d\Lambda_G(u|W) \quad \text{(S-1)}$$

where $m_{t;2}(u,w;S) = E_S[(I(T \geq t) - \psi_t(W))^2 | T > u, W = w]$ for any proper survival function $S(\cdot|\cdot)$. The leading term in (S-1) is $L_{t;2,ipcw}(O(t),\psi_t;G)$ and it can be shown that (S-1) and $L_{t;2}(T,\psi_t(W))$ have equal conditional (i.e., given W) expectations if either $G(\cdot|\cdot) = G_0(\cdot|\cdot)$ or

$S(\cdot|\cdot) = S_0(\cdot|\cdot)$; hence, it is a doubly robust CUT for $L_{t;2}(T, \psi_t(W))$. Similarly, the direct application of (7) to $O(t)$ combined with the equivalence $L_{t;2}(T, \psi_t(W)) = L_{t;2}(T(t), \psi_t(W))$ gives

$$\Delta(t)(I(\tilde{T} \geq t) - \psi_t(W))^2 + (1 - \Delta(t))m_{t;2}(\tilde{T}(t), W; S) \tag{S-2}$$

as a CUT for $L_{t;2}(T, \psi_t(W))$ when $S(\cdot|\cdot) = S_0(\cdot|\cdot)$.

We respectively refer to (S-1) and (S-2) as the doubly robust and Buckley-James Brier loss functions, denoted respectively by $L_{t;2,d}(O(t), \psi_t; G, S)$ and $L_{t;2,b}(O(t), \psi_t; S)$. Because (S-1) and (S-2) are each CUTs for the Brier loss function (i.e., squared error loss) and focus directly on estimating $S_0(t|W)$, the results of Section 4.1 and Theorem 4.1 can be used to justify implementing the corresponding *CURT*$-L_2$ algorithm by applying *CART* with squared error loss to the imputed dataset $\{(\hat{Z}(O_i(t); G, S), W_i')'; i = 1, \ldots, n\}$ where $\hat{Z}(O_i(t); G, S) = A_{1i}(G) + B_{1i}(G, S) - C_{1i}(G, S)$ is computed by replacing $(\log \tilde{T}_i, \Delta_i)$ with $(I(\tilde{T}_i \geq t), \Delta_i(t))$. Section 4.2 gives the corresponding recipe for implementing *CURE*$-L_2$ with either (S-1) or (S-2). In both cases, the terminal node estimators used to construct the ensemble predictor for $S_0(t|W)$ in Section 5.4 are naturally induced by the choice of loss function.

As described in Section 5.3, the estimator for $G_0(\cdot|\cdot)$ is modified to avoid violations of the positivity assumption using “Method 2” truncation; see Steingrimsson et al. (2016) for a detailed description of “Method 2” truncation. The doubly robust Brier loss at time t automatically induces this kind of truncation with $\hat{\vartheta} = t$; hence as long as $\hat{\vartheta}$ is larger than the time-point used to calculate the Brier loss, “Method 2” truncation has no discernible effect.

The doubly robust Brier loss function (S-1) and the Buckley-James Brier loss function (S-2) are potentially of interest in settings that extend outside the scope of this paper. For example, similarly to Graf et al. (1999), one may find these methods useful in validating prognostic models.

S.2 Additional Simulation Results

In this section we summarize additional simulation results that supplement those given in Section 5 in the main document. We employ *CURE-*L_2 algorithms with the doubly robust and Buckley-James Brier losses. We use *Brier* to denote the *CURE-*L_2 algorithm with the doubly robust loss function given by (S-1); see Section S.1. Similarly to *L2* and *L2-BJ*, $G(\cdot|\cdot)$ and $S(\cdot|\cdot)$ are replaced by the estimates described in Section 5.3 of main paper. Finally, we use *Brier BJ* to denote the same algorithm as *Brier*, but where $G(t|w) = 1$ everywhere. Due to the superior or similar performance of the *RIST* algorithm with at least 6 cases per terminal node compared to having at least 3 cases per terminal node we only include the former in all results presented in the Appendix.

S.2.1 Results for 25th and 75th quantiles

Figures S-1 through S-3 show results estimating $P(T > t|W)$ with t chosen as the 25th, 50th and 75th quantile of the marginal failure time distribution in simulation settings $1 - 4$; these plots augment the comparisons in Figure 1 of the main paper by including results for the Brier loss function (i.e., *Brier* and *Brier BJ*) and considering two additional time points. See Section 5.1 of the main paper for details regarding the simulation setup.

The trends for the $CURE{-}L_2$ algorithm are similar to the ones seen in Figure 1. The $CURE{-}L_2$ algorithms *L2* and *L2 BJ* outperform *RSF* in all settings. Compared to *CI*, these methods are significantly better in Settings 3 and 4 and perform similarly in Settings 1 and 2. For *RIST*, performance is similar in all settings and at all quantiles. There is somewhat greater variation in the performance comparisons for the single time point methods *Brier* and *Brier BJ* methods, though generally speaking these methods are reasonably competitive to the others (each of which uses information across time). The respective performance of the Buckley-James and doubly robust CUTs is similar in all settings, though there are notable improvements using the Buckley-James CUT (*Brier BJ* vs. *Brier*) at the 75th quantile in Settings 3 and 4. It is well known that estimators based on IPCW weights, such as doubly robust estimators, have the disadvantage of not being guaranteed to respect the natural range of the target parameter. Using the Brier loss function, the target parameter is a probability and is therefore constrained to fall in $[0, 1]$; when the terminal node estimators used in *CURT* trees that comprise the *Brier* predictions are truncated to fall in that interval the performance of the modified *Brier* algorithm (not shown) is again comparable to *Brier BJ*.

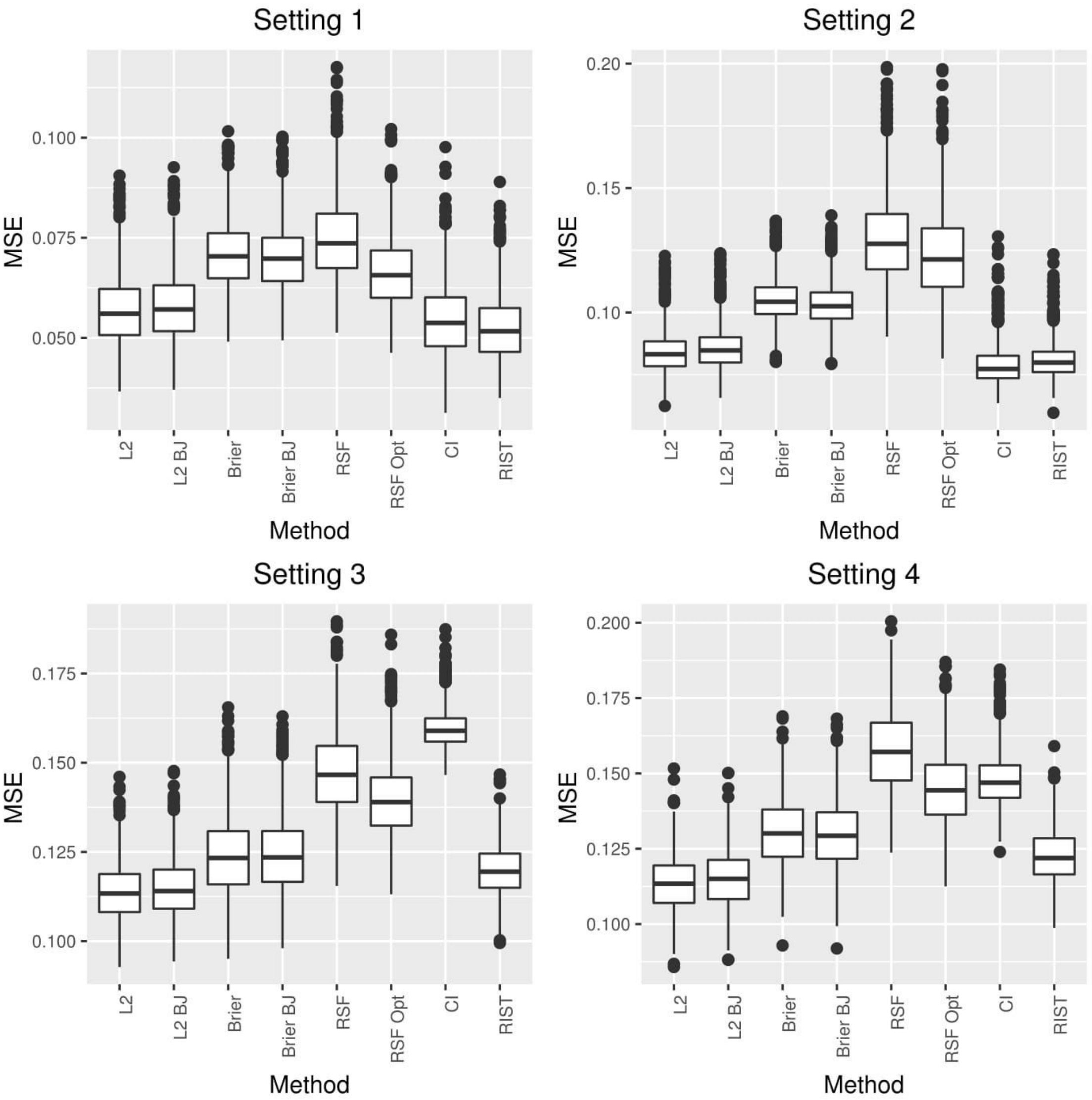


Figure S-1: Boxplots of MSE estimated at the $25th$ quantile of the marginal failure time distribution for the four simulation settings of Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the $CURE{-}L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. *RSF* and *CI* are the default methods for `rfsrc` and `cforest` functions. *RSF Opt* is the default method for `rfsrc` with the `nodesize` parameter tuned. *RIST* is the recursively imputed survival trees.

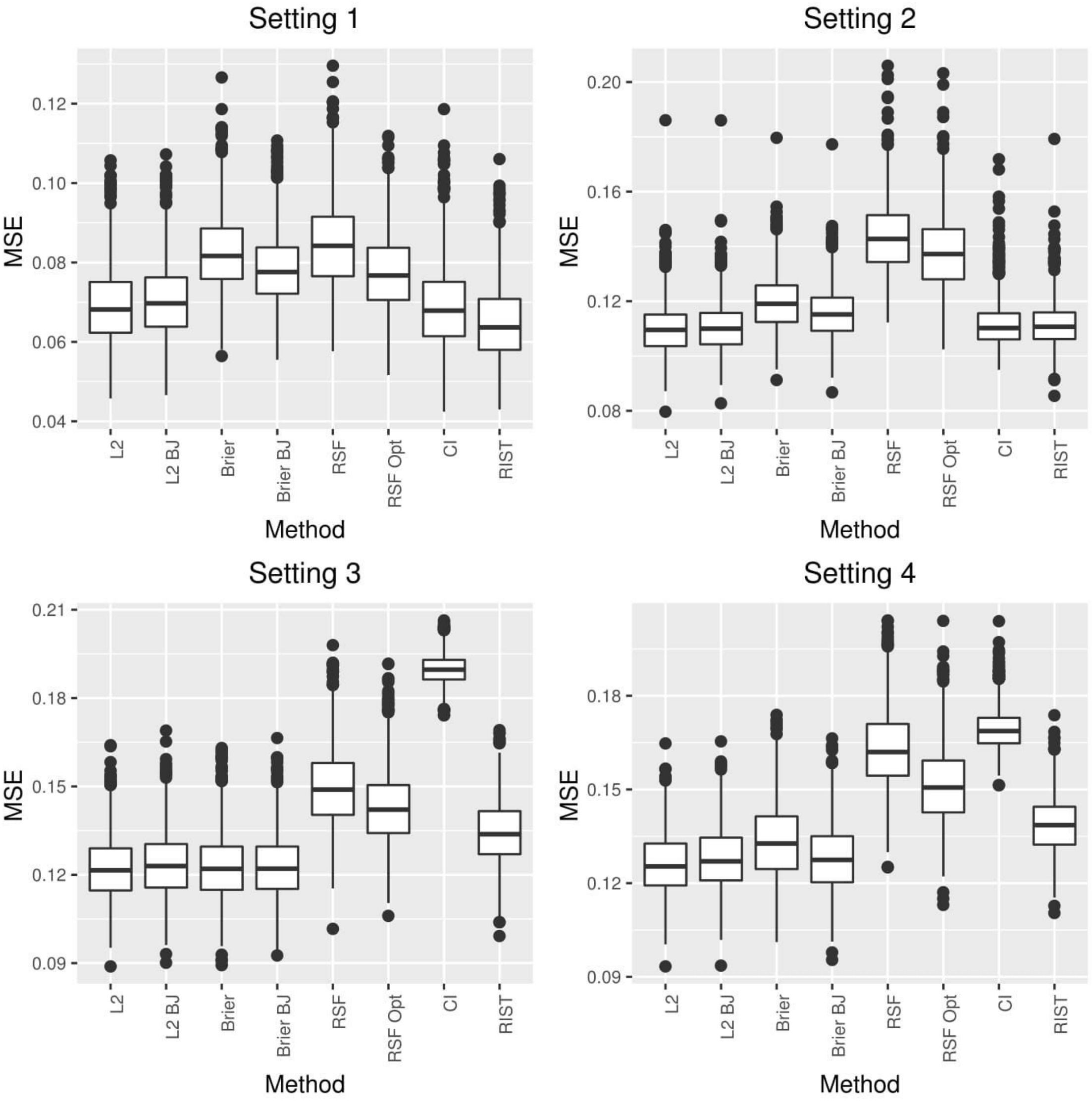


Figure S-2: Boxplots of MSE estimated at the 50th quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the $CURE{-}L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT, and *L2* and *Brier* referring to the choice of loss function. *RSF* and *CI* are the default methods for `rfsrc` and `cforest` functions. *RSF Opt* is the default method for `rfsrc` with the `nodesize` parameter tuned. *RIST* is the recursively imputed survival trees algorithm.

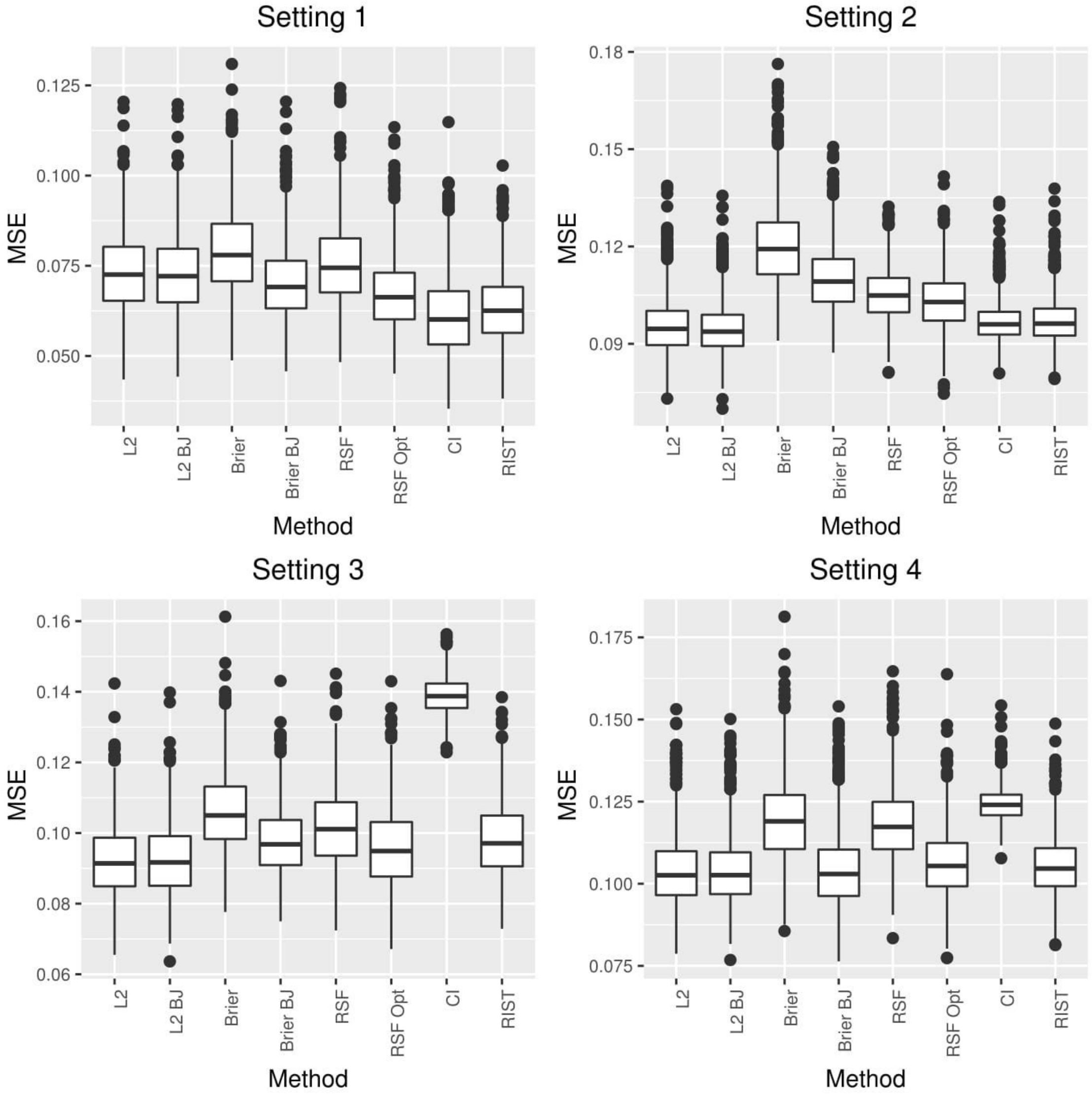


Figure S-3: Boxplots of MSE estimated at the $75th$ quantile of the marginal failure time distribution for the four simulation settings of Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the $CURE{-}L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. *RSF* and *CI* are the default methods for `rfsrc` and `cforest` functions. *RSF Opt* is the default method for `rfsrc` with the `nodesize` parameter tuned. *RIST* is the recursively imputed survival trees.

S.2.2 Nonparametric versus exchangeable weighted bootstrap for ensembles

In this section we first compare the performance of the *CURE*$-L_2$ algorithm implemented using the Bayesian and the nonparametric bootstrap. Both bootstraps are implemented using the `R` function `randomForest` in the `randomForest` package (Liaw and Wiener, 2002) with the Bayesian bootstrap requiring extending the capabilities of the function to allow for arbitrary bootstrap weights. The simulation settings used are the same as used in the main document; see Section 5.1 for further details. The results for each *CURE* algorithm are given in Figures S-4 - S-6.

From Figures S-4 - S-6 we see that the *CURE*$-L_2$ algorithms are not very sensitive to the choice of bootstrap weights. For the Brier loss function, the nonparametric bootstrap does as well or slightly better than the Bayesian bootstrap in all settings and at all quantiles. For the L_2 loss, the relative performance of the two bootstrap procedures depends on the simulation setting and quantile considered.

Weng (1989) shows that if the weights in the i.i.d. weighted bootstrap are simulated from a Gamma(4,1) distribution, the bootstrap weights are second order equivalent to the nonparametric bootstrap weights for bootstrapping the sample mean. The two weights mainly differ in that the former puts positive weights on every observation, while the latter only includes approximately 63% of the observations in each bootstrap sample. Figures S-7 - S-9 show simulation results comparing the non-parametric bootstrap to the $Gamma(4,1)$ bootstrap for the four different settings described in Section 5.1.

The boxplots again show that the results are not sensitive to the choice of bootstrap algorithm. For the L_2 loss, both bootstraps perform similarly and it depends on the setting and quantile considered which performs better. For the Brier loss, the nonparametric bootstrap either performs similarly or better compared to the $Gamma(4,1)$ bootstrap.

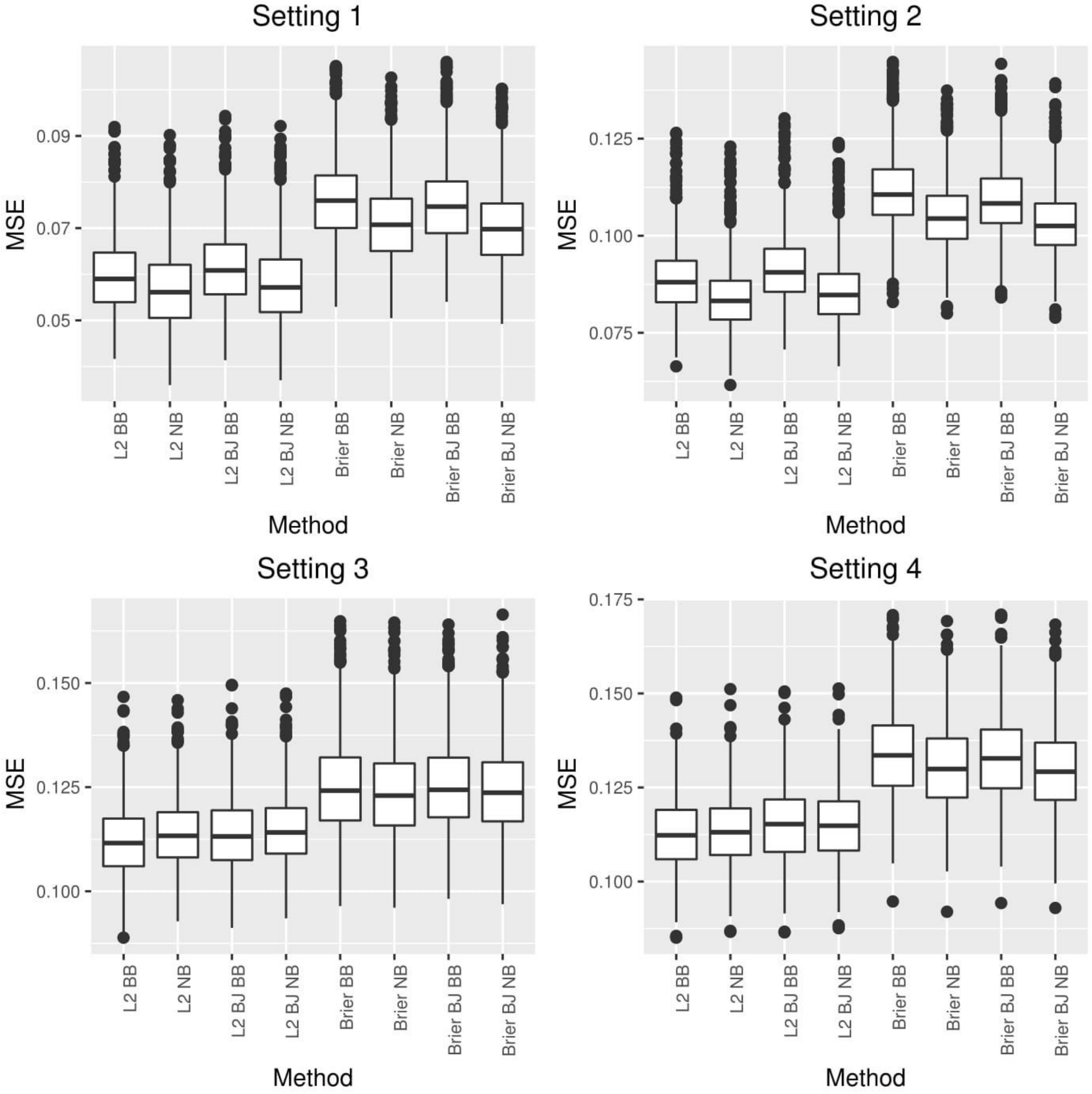


Figure S-4: Boxplots of MSE at the 25*th* quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the *CURE*$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. NB and BB respectively indicate use of the nonparametric and Bayesian bootstrap weights.

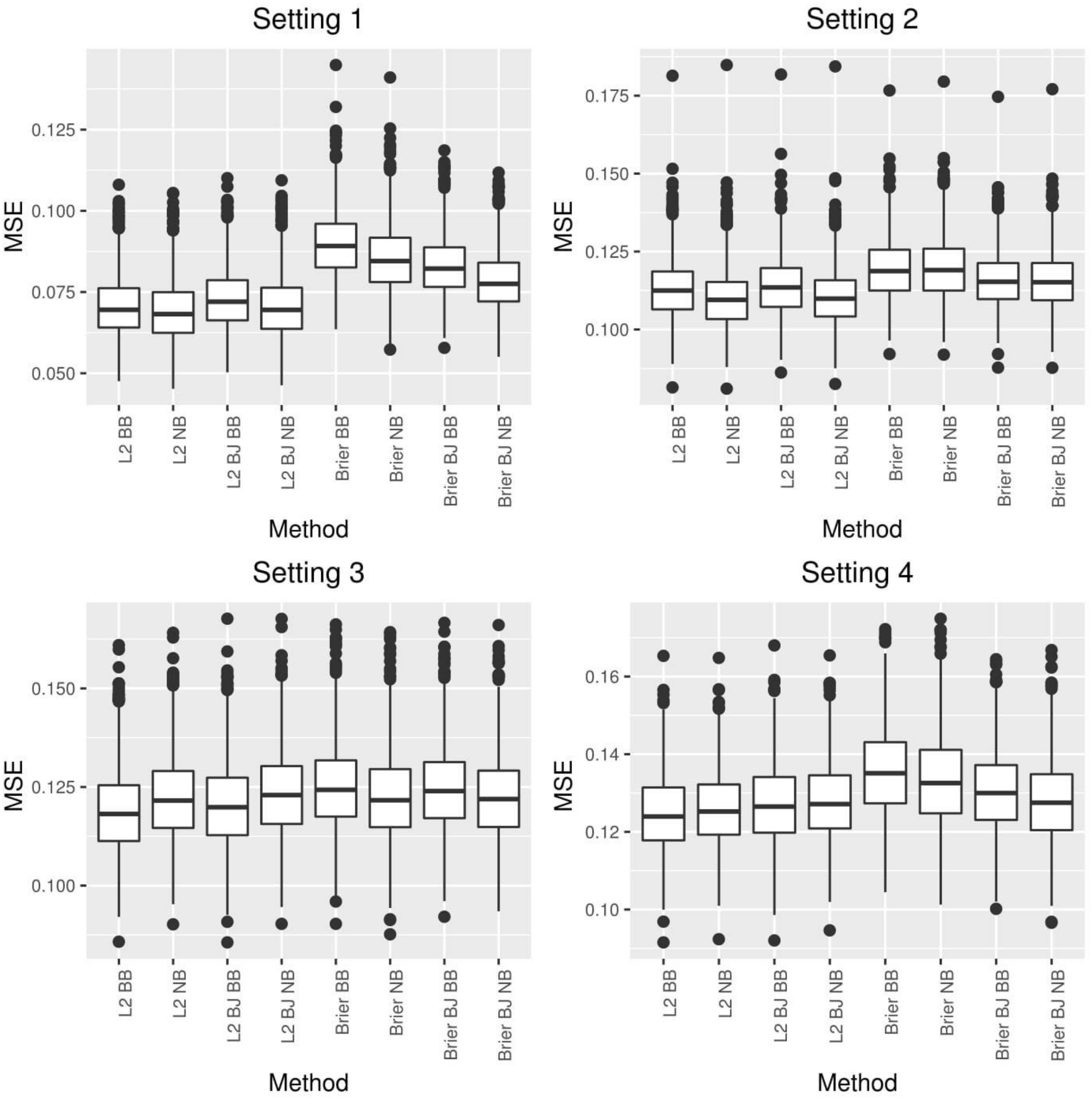


Figure S-5: Boxplots of MSE at the 50*th* quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the *CURE*$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. NB and BB respectively indicate use of the nonparametric and Bayesian bootstrap weights.

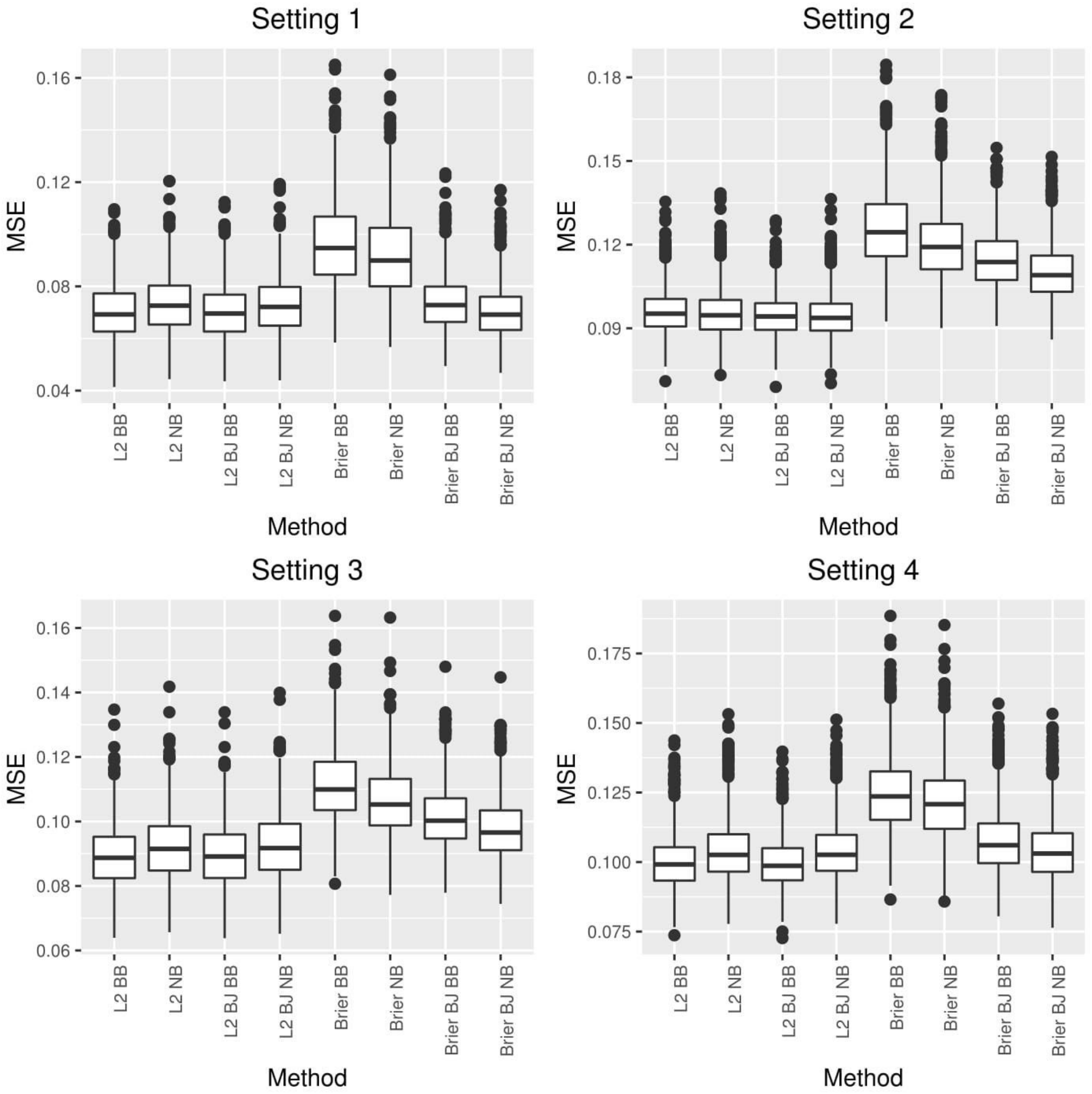


Figure S-6: Boxplots of MSE at the $75th$ quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the *CURE*$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. NB and BB respectively indicate use of the nonparametric and Bayesian bootstrap weights.

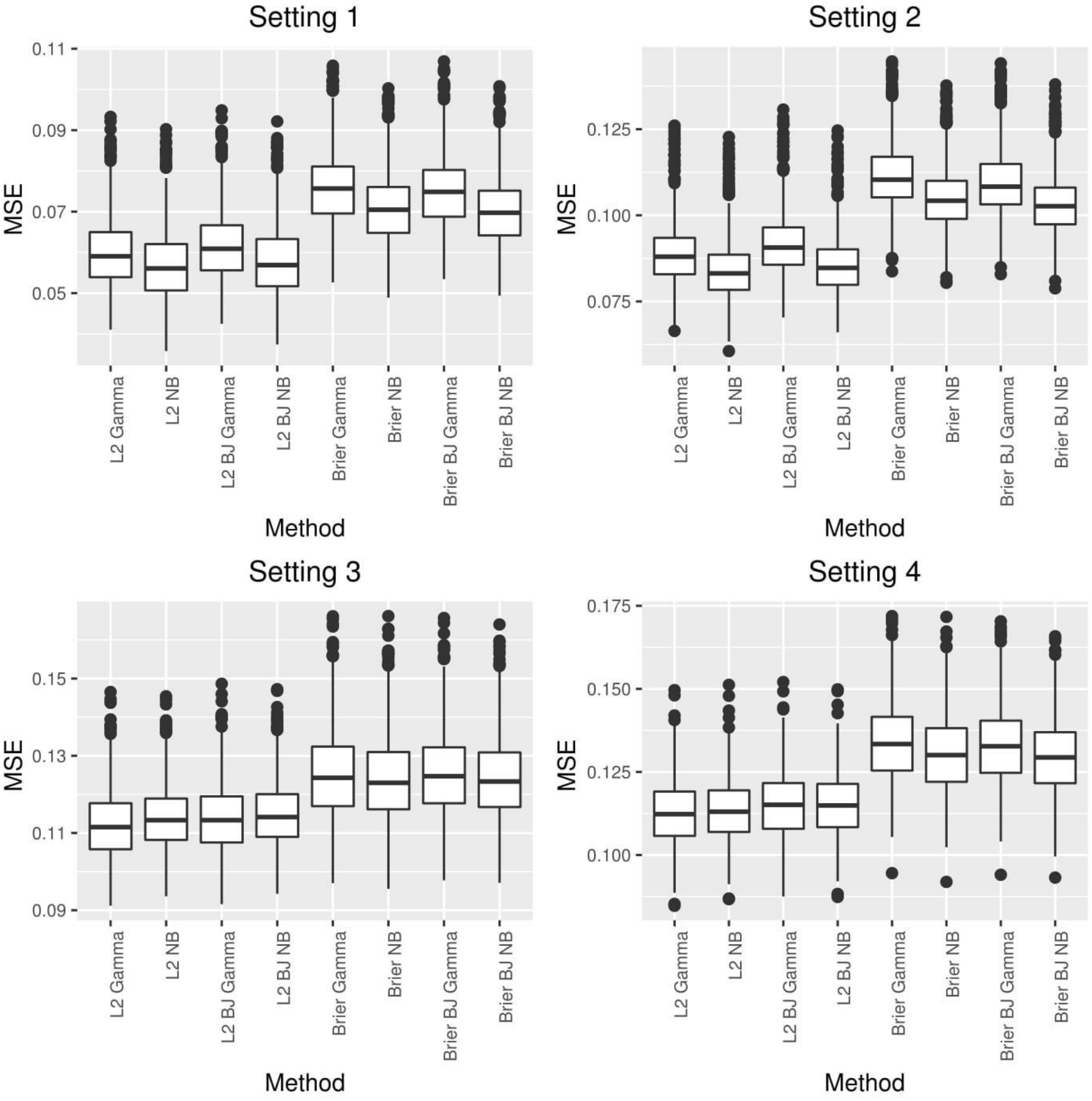


Figure S-7: Boxplots of MSE at the $25th$ quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the $CURE{-}L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT and *L2* and *Brier* as the choice of loss function. NB and Gamma respectively indicate use of the nonparametric bootstrap and the i.i.d. weighted bootstrap with weights simulated using the $Gamma(4, 1)$ distribution.

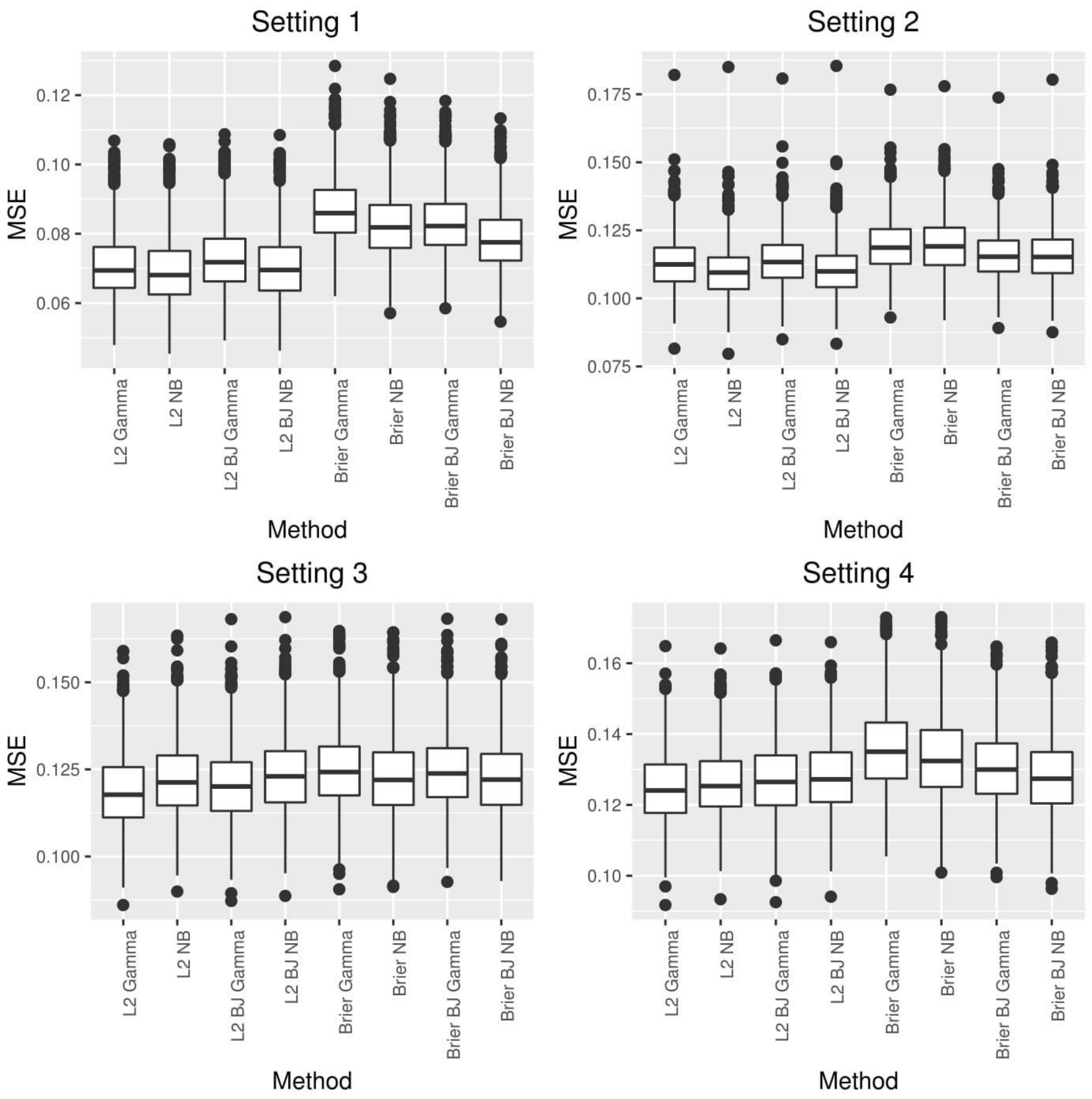


Figure S-8: Boxplots of MSE at the $50th$ quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the $CURE{-}L_2$ algorithms, with BJ referring to the use of the Buckley-James CUT. NB and Gamma respectively indicate use of the nonparametric bootstrap and the i.i.d. weighted bootstrap with weights simulated using the $Gamma(4, 1)$ distribution.

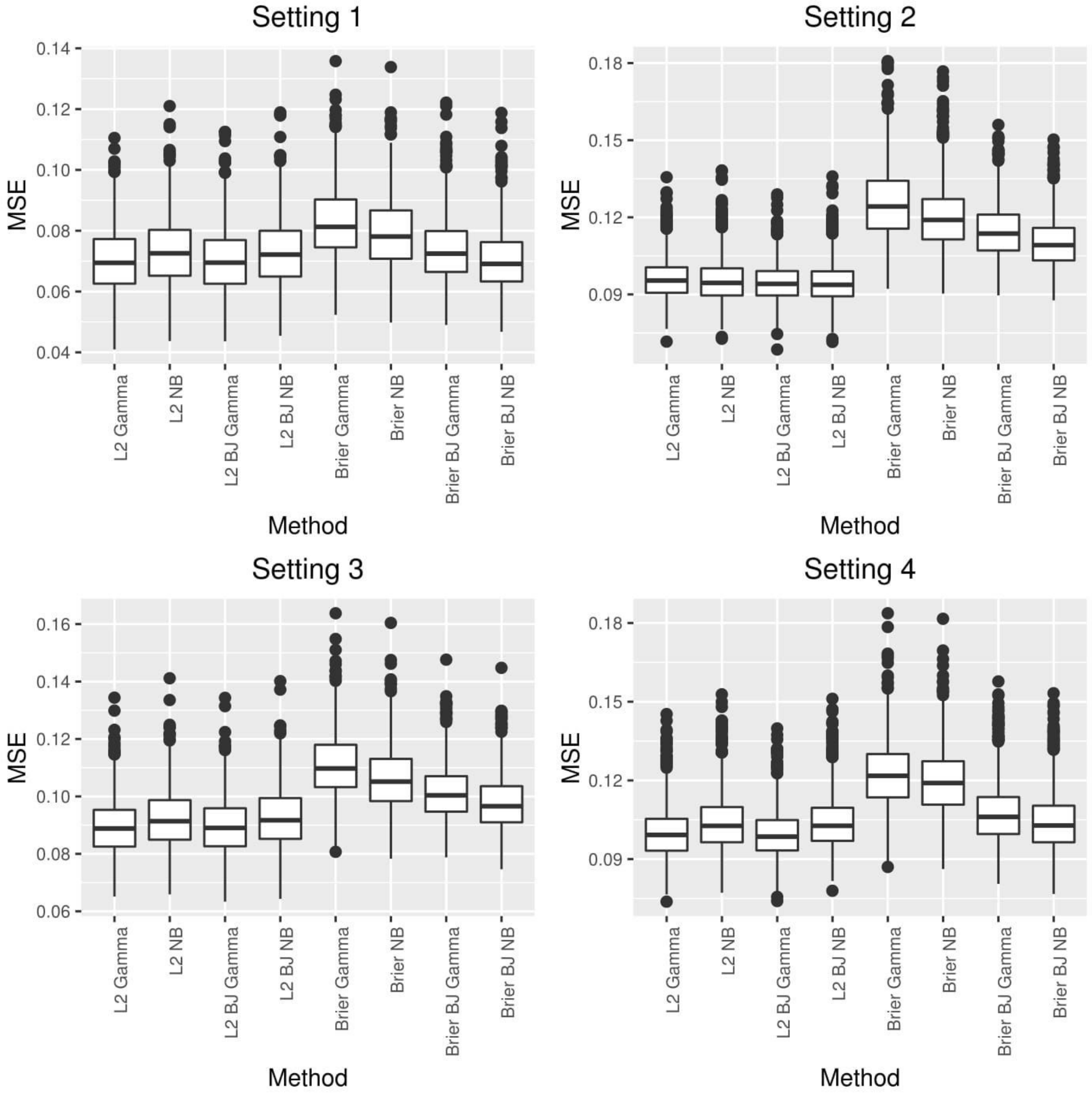


Figure S-9: Boxplots of MSE at the $75th$ quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the $CURE{-}L_2$ algorithms, with BJ referring to the use of the Buckley-James CUT. NB and Gamma respectively indicate use of the nonparametric bootstrap and the i.i.d. weighted bootstrap with weights simulated using the $Gamma(4, 1)$ distribution.

S.2.3 Simulations with G_0 Estimated Using a Kaplan-Meier Estimator and S_0 using an AFT Model

Figure S-10 shows simulation results for the four settings described in Section 5.1 with $G_0(\cdot|\cdot)$ estimated using a Kaplan-Meier estimator. To ensure that $\hat{G}(\cdot|\cdot)$ remains bounded away from zero we use 10% "Method 2" truncation as described in Steingrimsson et al. (2016).

Figure S-11 shows simulation results when $S_0(\cdot|\cdot)$ is estimated using a parametric accelerated failure time model with an error distribution which is assumed to follow a Weibull distribution.

The results show similar trends to those summarized in Section 5.

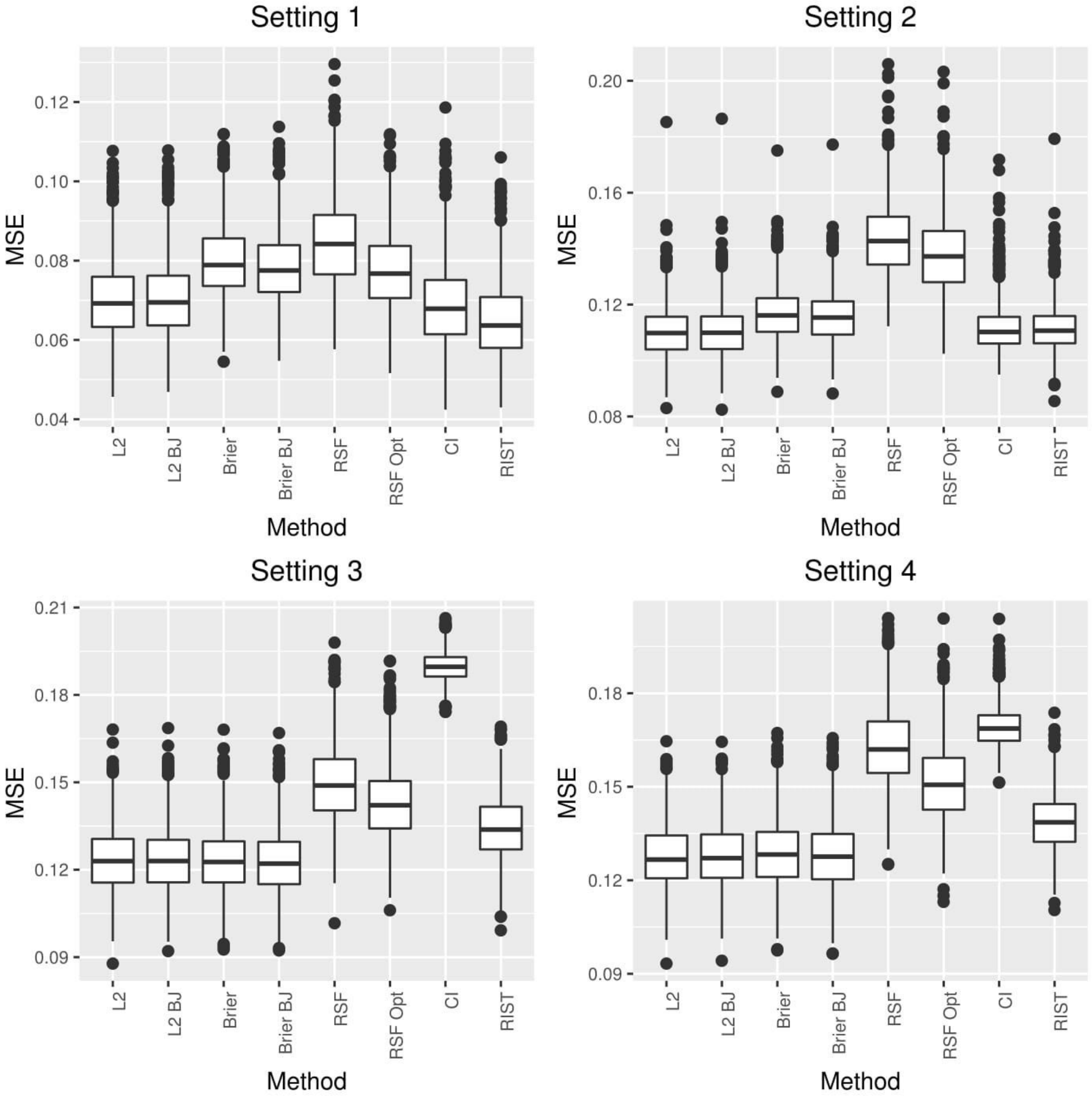


Figure S-10: Boxplots of MSE estimated at the 50*th* quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the *CURE*$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT, and *L2* and *Brier* referring to the choice of loss function. For all the *CURE*$-L_2$ algorithms, $G_0(\cdot|\cdot)$ is estimated using a Kaplan-Meier estimator. *RSF* and *CI* are the default methods for `rfsrc` and `cforest` functions. *RSF Opt* is the default method for `rfsrc` with the `nodesize` parameter tuned. *RIST* is the recursively imputed survival trees algorithm.

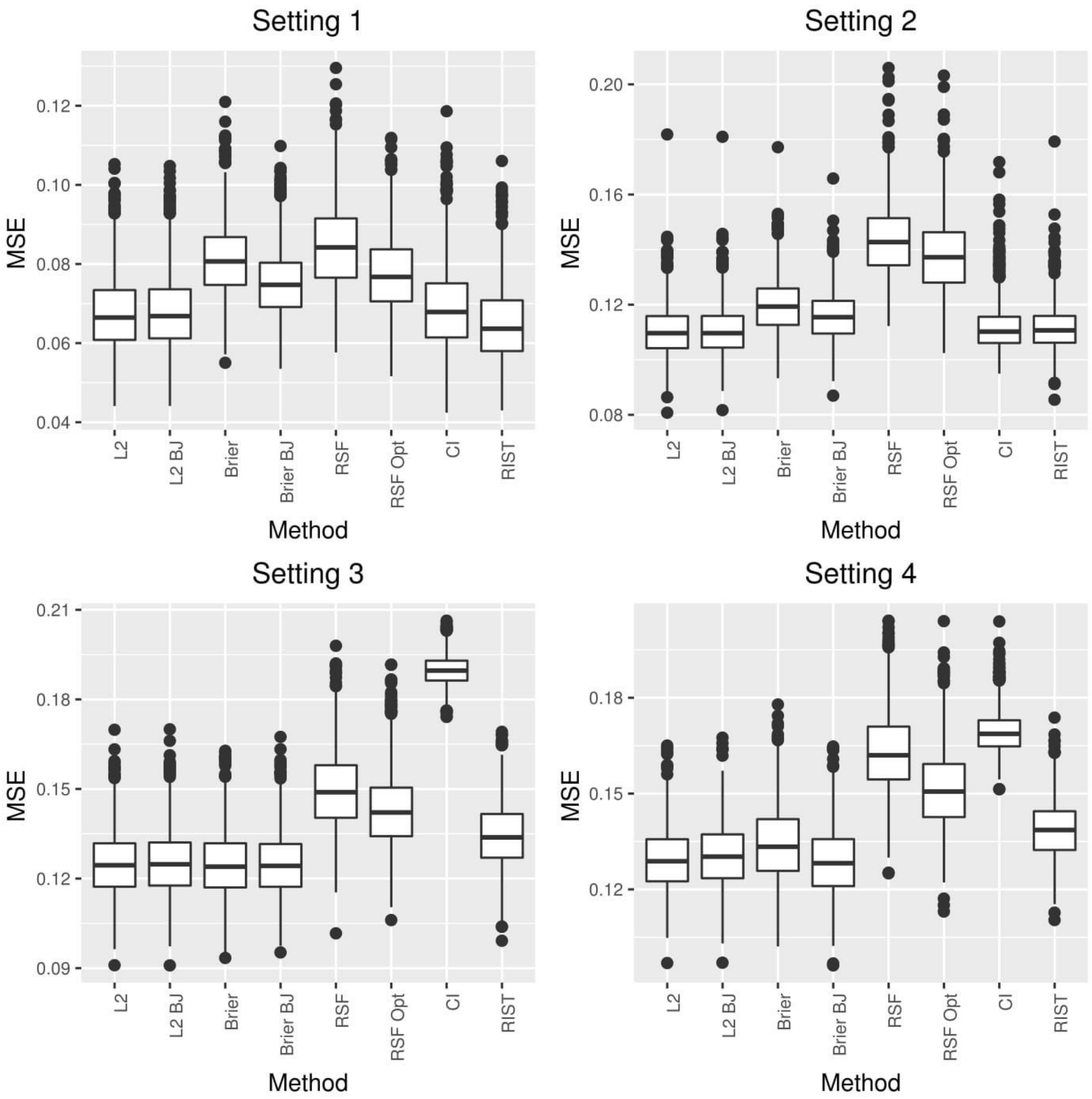


Figure S-11: Boxplots of MSE estimated at the 50th quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the $CURE{-}L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT, and *L2* and *Brier* referring to the choice of loss function. For all the $CURE{-}L_2$ algorithms, $S_0(\cdot|\cdot)$ is estimated using a parametric accelerated failure time model with the error distribution assumed to follow a Weibull distribution. *RSF* and *CI* are the default methods for `rfsrc` and `cforest` functions. *RSF Opt* is the default method for `rfsrc` with the `nodesize` parameter tuned. *RIST* is the recursively imputed survival trees algorithm.

S.2.4 Simulations with increased covariate dimension

Figures S-12 and S-13 show simulation results for the four settings described in Section 5.1 when the covariate dimension is respectively increased to 50 and 100. For all four settings, the dimensionality is increased by adding noise variables. For settings one, three, and four the covariate vector is simulated from a multivariate normal distribution with mean zero and a covariance matrix having element (i, j) equal to $0.9^{|i-j|}$. For setting two, the covariate vector is constructed by simulating i.i.d. uniform random variables on the interval $[0, 1]$.

The results show similar trends to those summarized in Section 5 with the relative performance of the $CURE{-}L_2$ algorithm becoming slightly better as the covariate dimension increases.

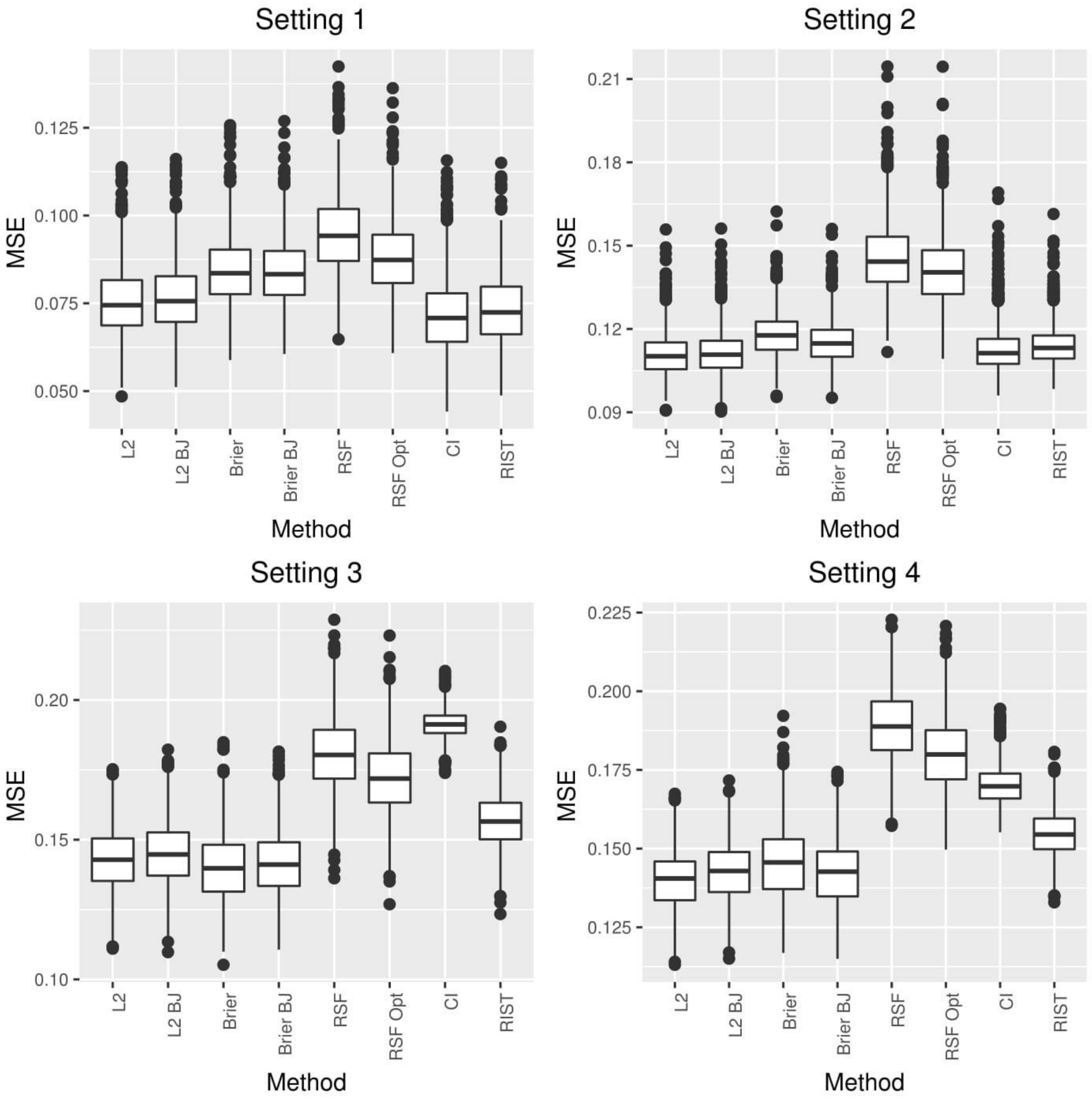


Figure S-12: Boxplots of MSE estimated at the 50*th* quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1 when the covariate dimension is equal to 50. *L2, L2 BJ, Brier* and *Brier BJ* are the $CURE{-}L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT, and *L2* and *Brier* referring to the choice of loss function. *RSF* and *CI* are the default methods for `rfsrc` and `cforest` functions. *RSF Opt* is the default method for `rfsrc` with the `nodesize` parameter tuned. *RIST* is the recursively imputed survival trees algorithm.

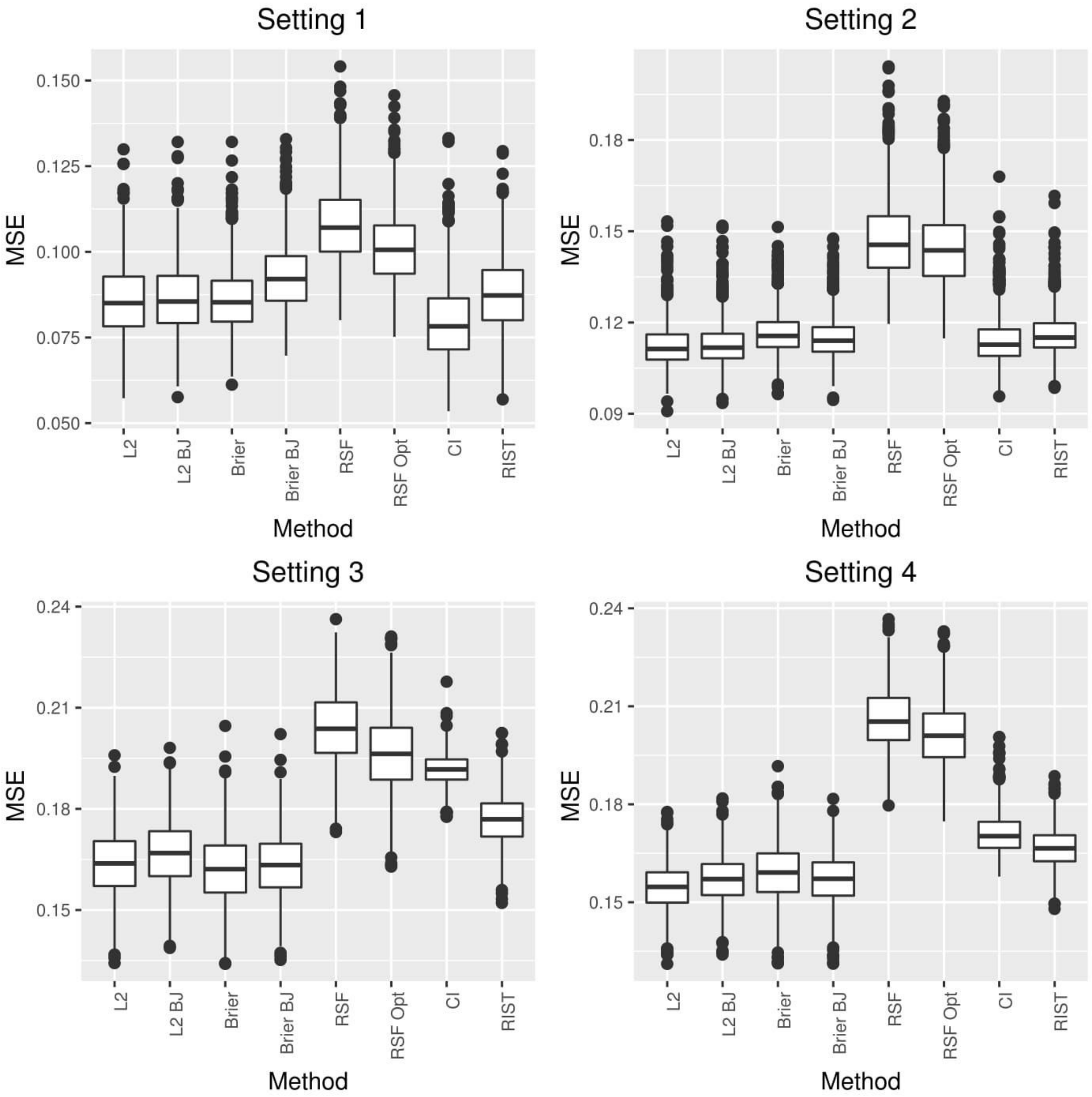


Figure S-13: Boxplots of MSE estimated at the $50th$ quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1 when the covariate dimension is equal to 100. *L2, L2 BJ, Brier* and *Brier BJ* are the *CURE*$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT, and *L2* and *Brier* referring to the choice of loss function. *RSF* and *CI* are the default methods for `rfsrc` and `cforest` functions. *RSF Opt* is the default method for `rfsrc` with the `nodesize` parameter tuned. *RIST* is the recursively imputed survival trees algorithm.

S.2.5 Simulations with 50% censoring rate

Figure S-14 shows simulation results for the four settings described in Section 5 when the overall censoring rate is 50%. For setting one, the rate of the exponential censoring distribution is modified to achieve 50% censoring. For settings two and three, the upper support of the uniform distribution is modified to get 50% censoring. For setting four, the censoring distribution is log-normal with mean $0.1 \left|\sum_{i=1}^{5} W_i\right| + 0.1 \left|\sum_{i=21}^{25} W_i\right|$ and scale parameter 1.

The boxplots in Figure S-14 show similar relative performance to the results shown in Figure 1 in the main paper.

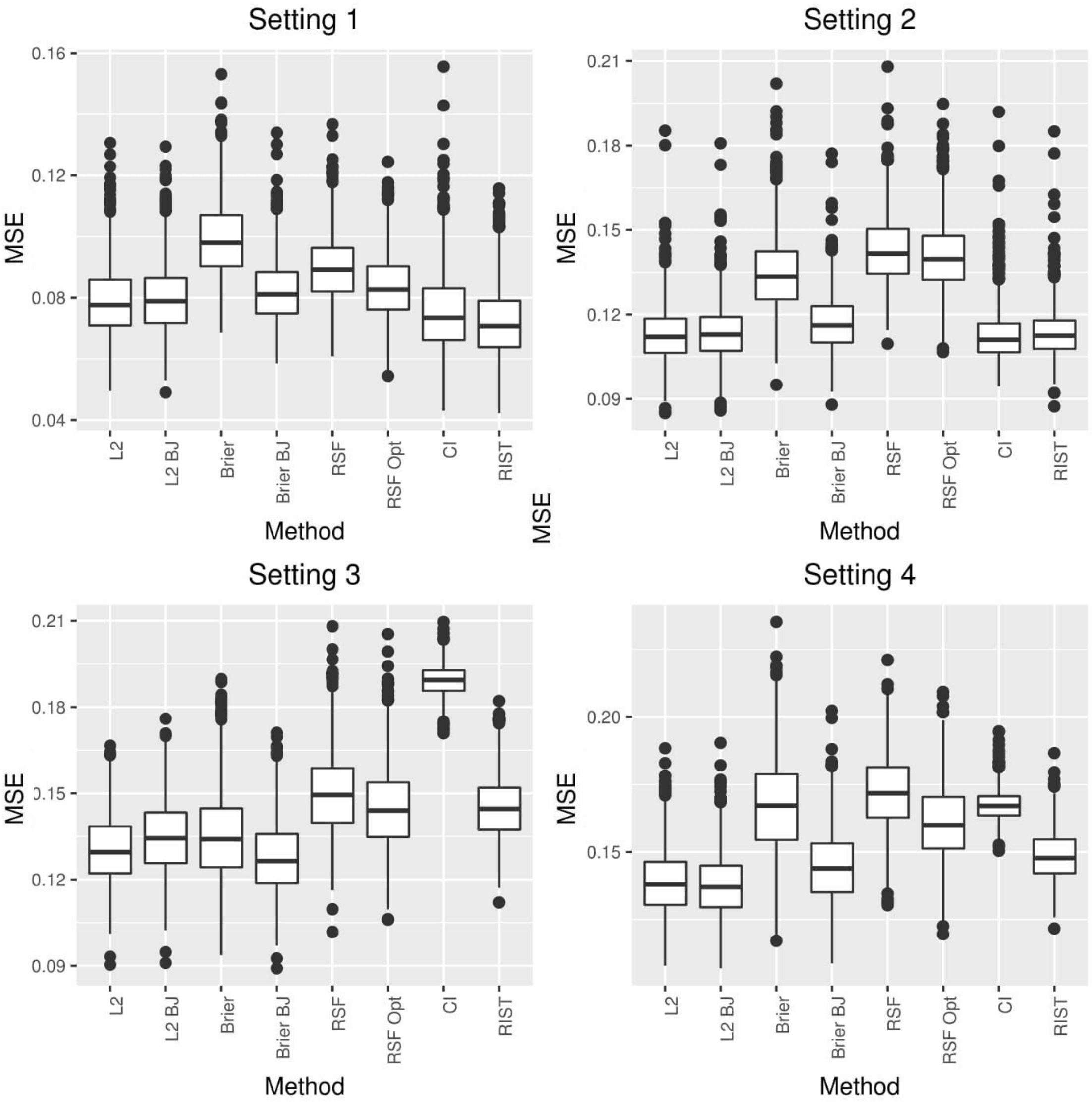


Figure S-14: Boxplots of MSE estimated at the $50th$ quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1 when the censoring rate is 50% for all settings. *L2, L2 BJ, Brier* and *Brier BJ* are the *CURE*$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT, and *L2* and *Brier* referring to the choice of loss function. *RSF* and *CI* are the default methods for `rfsrc` and `cforest` functions. *RSF Opt* is the default method for `rfsrc` with the `nodesize` parameter tuned. *RIST* is the recursively imputed survival trees algorithm.

S.2.6 Revisiting the simulation study in Steingrimsson et al. (2016)

In this section, we revisit the simulation studies for survival trees conducted in Steingrimsson et al. (2016) and compare the performance of the $CURT{-}L_2$ algorithm using the Buckley-James (see (7)) and doubly robust (see (6)) loss functions, with both being implemented using the imputation approach described in Section 4.1. The following two subsections revisit the simulation settings used in Steingrimsson et al. (2016) (Section S.2.6) and summarize the results (Section S.2.6).

S.1.6: Simulation Settings

Steingrimsson et al. (2016) considered two simulation settings. Both settings contain a training set of 250 independent subjects from the observed data distribution (subject to censoring) and a test set of 2000 independent observations from the full data distribution (with no censoring). We simulate 1000 independent training and test set combinations. We briefly review the two settings considered below:

Simulation Setting 1: There are five covariates $W_1, \ldots, W_5$, each of which follows a discrete uniform distribution on the integers 1-100. The response $Z = \log T$ and survival times T are generated from an exponential distribution with a covariate-dependent mean parameter $\mu = aI(W_1 > 50 \mid W_2 > 75) + 0.5I(W_1 \leq 50 \ \& \ W_2 \leq 75)$. We consider "high" ($a = 5$), "medium" ($a = 2$) and "low" ($a = 1$) signal settings representing different degrees of separation in the survival curves. The censoring time C follows an exponential distribution with mean parameter μ_c, where μ_c is chosen to (approximately) achieve a 30% marginal censoring rate, in other words, $P(T \geq C; \mu_c) = 0.3$.

Simulation Setting 2: This simulation setting is similar to setting D in LeBlanc and Crowley (1992). It differs from Setting 1 in that the proportional hazard assumption does not hold. Assume that covariates $W_1, \ldots, W_5$ are independently uniformly distributed on the interval [0,1]. Survival times are generated from a distribution with survivor function $S(t|W) = [1 + t\exp(aI(W_1 \leq 0.5, W_2 > 0.5) + 0.367)]^{-1}$. The choices $a = 2, 1.5$ and 1 respectively correspond to "high", "medium" and "low" signal settings. The censoring times C follow a uniform distribution on $[0, b]$, where b is chosen to (approximately) achieve a 30% marginal censoring rate.

S.1.6: Simulation Results

The censoring distributions in both Settings 1 and 2 are independent of covariates; each is estimated using a Kaplan-Meier estimator. The conditional expectations required for computing the doubly robust and Buckley-James loss functions are respectively estimated using a parametric accelerated failure time (AFT) model with lognormal errors and also using random survival forests; see Section 3.2.2 of Steingrimsson et al. (2016) for details. The performance of the different survival trees for Settings 1 and 2 is respectively summarized in Figures S-15 and S-16 using the mean squared error of survival differences at the 25th, 50th and 75th quantile of the marginal failure time distribution (MSE25, MSE50 and MSE75). Each figure contains 9 plots and summarizes the results for MSE25, MSE50 and MSE75 under high, medium and low signal settings. The 6 boxplots in each plot respectively correspond to the method of LeBlanc and Crowley (1992) as implemented in `rpart` (*EXP*); the inverse probability censoring weighted L_2 loss (*IPCW*); the doubly robust L_2 loss calculated using the parametric AFT model (*DR-AFT*); the doubly robust L_2 loss calculated using random survival forest predictions (*DR-RF*); the Buckley-James L_2 loss calculated using the parametric AFT model (*BJ-AFT*); and, the Buckley-James L_2 loss calculated using random survival forest predictions (*BJ-RF*).

Figure S-15 shows that the performance of *EXP*, *IPCW* and doubly robust survival trees with conditional expectation estimated using either the AFT model or random survival forests are very similar, a result entirely consistent with the results for Simulation 1 in Steingrimsson et al. (2016). The doubly robust trees perform better than the *IPCW* trees in the high and medium signal setting and show similar performance in the low signal setting; performs similarly or slightly better than EXP in high signal setting, however, as well or slightly worse in medium and low signal settings. The performance of Buckley-James trees is essentially the same as the doubly robust trees in nearly all signal settings, with the Buckley-James trees fit using the AFT model having slightly smaller MSE at the 75th quantile.

For Simulation Setting 2, Figure S-16 shows that the doubly robust and Buckley-James trees perform noticeably better than both the *IPCW* trees and *EXP* method in the high and medium setting; performance is comparable for all methods in the low signal setting. Each of *DR-AFT*, *DR-RF*, *BJ-AFT* and *BJ-RF* have comparable performance in high and low signal settings; *BJ-AFT* performs best, with *DR-RF* being second best, in the medium signal setting.

For completeness we also looked at the performance of all survival trees in terms of prediction error as we did in Steingrimsson et al. (2016) (results not shown here). The results for *EXP*, *IPCW*, *DR-AFT* and *DR-RF* are consistent with those results, the pattern of prediction error agreeing with that observed in MSE25, MSE50 and MSE75.

Figure S-15: Boxplots of mean squared error of survival differences at the 25th, 50th and 75th quantile of the marginal failure time distribution (MSE25, MSE50 and MSE75) using the default method in `rpart` (*EXP*), inverse probability censoring weighted loss (*IPCW*), doubly robust L_2 loss with the AFT model (*DR-AFT*), doubly robust L_2 loss with *RSF* (*DR-RF*), Buckley-James L_2 loss with the AFT model (*BJ-AFT*) and Buckley-James L_2 loss with *RSF* (*BJ-RF*), respectively for the high, medium and low signal settings in Setting 1.

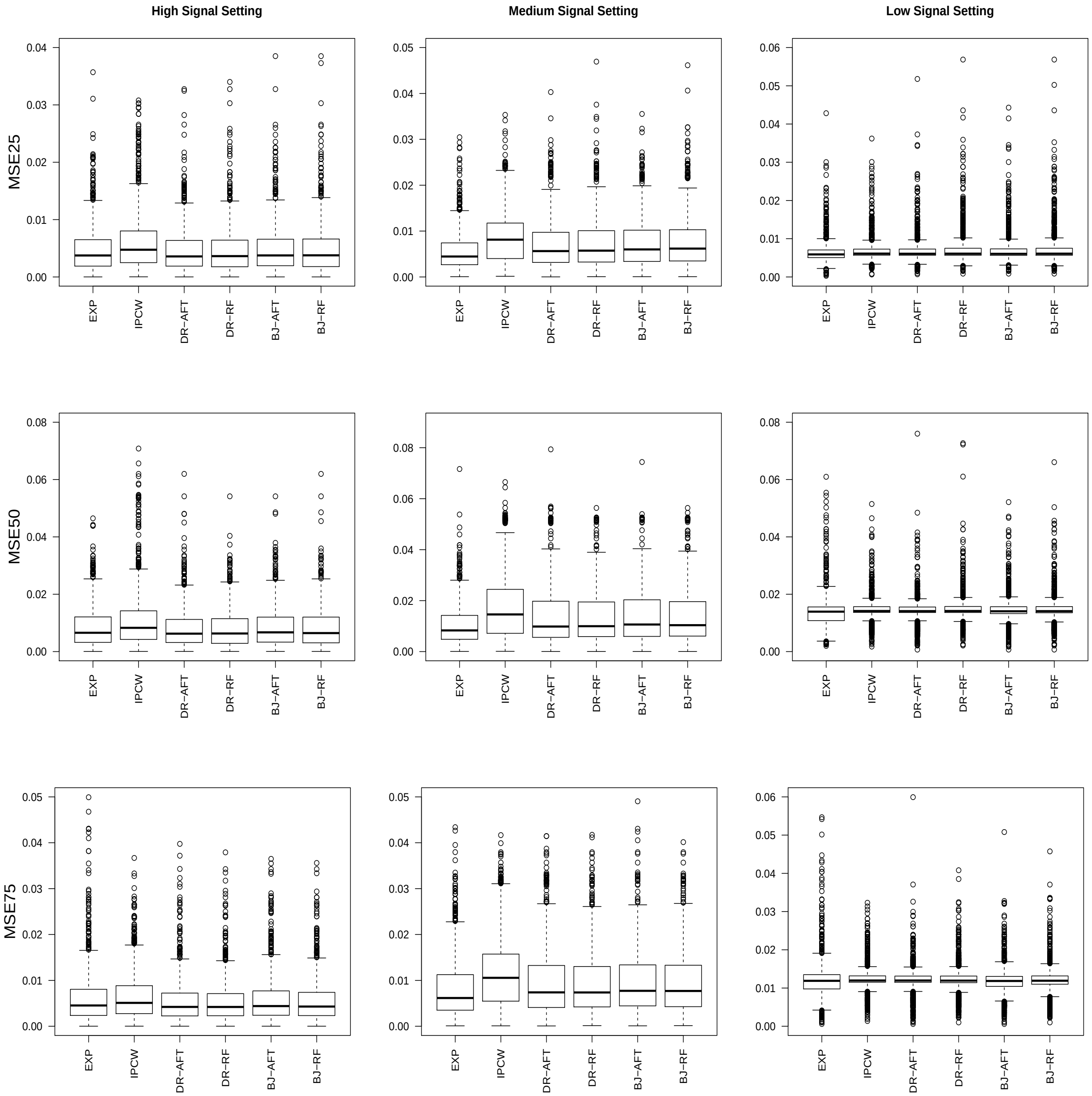
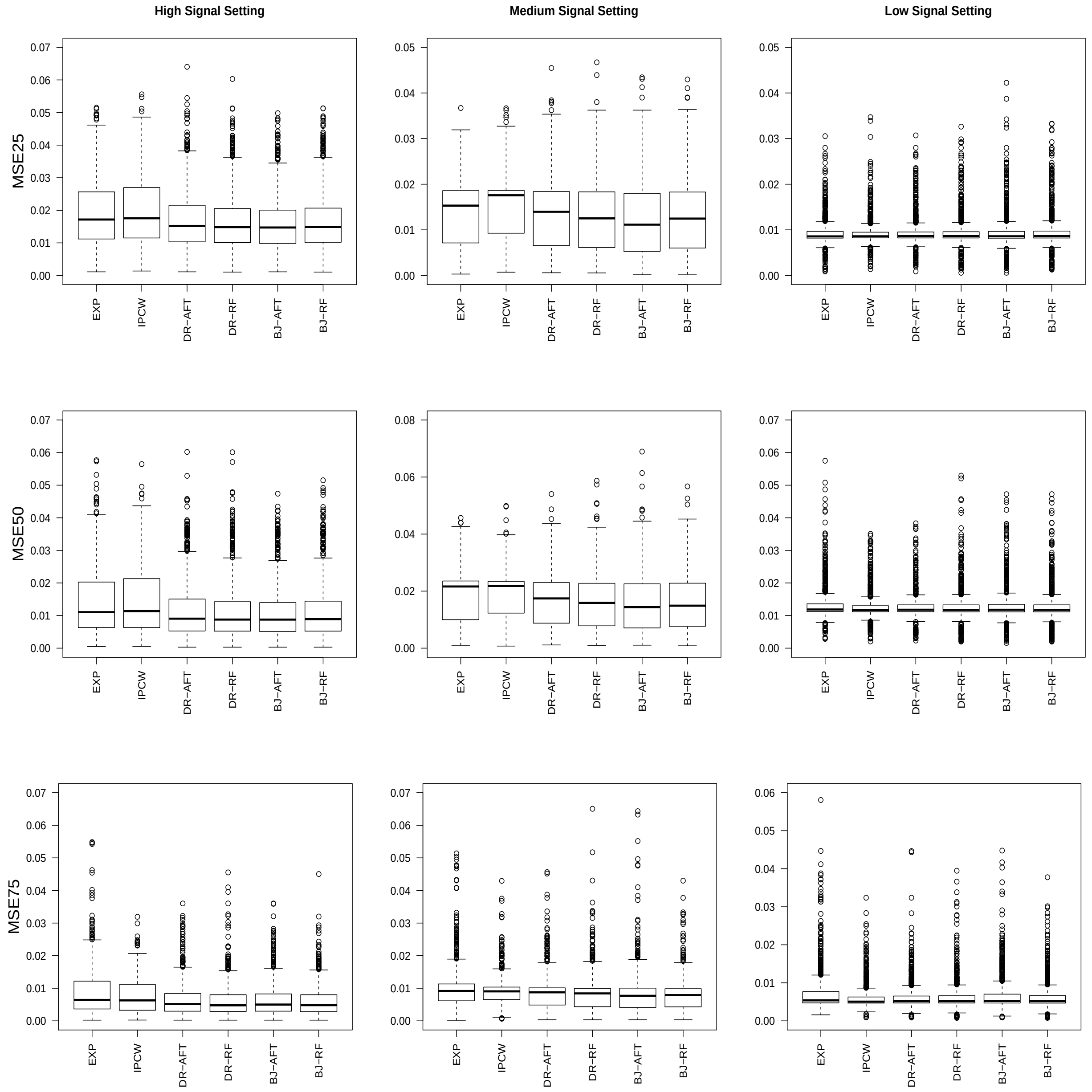


Figure S-16: Boxplots of mean squared error of survival differences at the 25th, 50th and 75th quantile of the marginal failure time distribution (MSE25, MSE50 and MSE75) using the default method in `rpart` (*EXP*), inverse probability censoring weighted loss (*IPCW*), doubly robust L_2 loss with the AFT model (*DR-AFT*), doubly robust L_2 loss with *RSF* (*DR-RF*), Buckley-James L_2 loss with the AFT model (*BJ-AFT*) and Buckley-James L_2 loss with *RSF* (*BJ-RF*), respectively for the high, medium and low signal settings in Setting 2.



S.3 Further Details on OOB-Based Variable Importance Measures

Consider an ensemble generated by the nonparametric bootstrap. Given a tree m from this ensemble, let $\hat{\psi}_m(W)$ be the corresponding prediction for a subject with covariate information W. Let B_m be the set of OOB data associated with the bootstrap sample used to create tree m. The L_2 OOB data prediction error for tree m is defined as

$$\frac{1}{|B_m|}\sum_{i=1}^{n} I(i \in B_m) L_2(Z_i, \hat{\psi}_m(W_i)), \tag{S-3}$$

where $|B_m|$ denotes the size of the OOB sample. For each $i \in B_m$ let $W_i^{(j)}$ be the covariate vector for subject i with the j-th component of the covariate permuted. Define the OOB L_2 loss prediction error using the resulting permuted OOB dataset as

$$\frac{1}{|B_m|}\sum_{i=1}^{n} I(i \in B_m) L_2(Z_i, \hat{\psi}_m(W_i^{(j)})). \tag{S-4}$$

The OOB prediction error VIMP proposed by Breiman (2001) is calculated as the difference between (S-4) and (S-3), averaged over all the trees in the ensemble. That is, for covariate j the OOB prediction error VIMP is defined as

$$\frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{|B_m|}\sum_{i=1}^{n} I(i \in B_m)(L_2(Z_i, \hat{\psi}_m(W_i^{(j)})) - L_2(Z_i, \hat{\psi}_m(W_i)))\right). \tag{S-5}$$

This calculation assumes that $(Z_i, W_i')', i \in B_m$ are fully observed. The corresponding VIMP using a CUT for the L_2 loss function can simply be defined as that which is obtained by replacing the (unobserved) L_2 loss in (S-5) with its corresponding CUT as given in (8).

As the OOB prediction error VIMP is defined as the difference between two loss functions the proof of the following theorem follows from exactly the same arguments as used to prove Theorem 4.1. For the notation used in the theorem we refer to Section 4 in main paper.

Theorem S.3.1. *For each $i = 1, \ldots, n$, define the loss function $L_2(O_i, \psi; G, S, Q) = \psi(W_i)^2 + H(O_i; G, S)\psi(W_i) + Q(O_i; G, S)$ and assume $\max\{|H(O_i; G, S)|, |Q(O_i; G, S)|\} < \infty$. The OOB prediction error VIMPs using the loss function $L_2(O, \psi; G, S, Q)$ do not depend on $Q(O; G, S)$.*

An important implication of Theorem S.3.1 is that the OOB data prediction error VIMPs using the L_2 loss in connection with doubly robust and Buckley-James CUTs can be implemented by running standard software calculating the OOB prediction error VIMPs for fully observed responses on the corresponding "imputed" dataset $\{(\hat{Z}(O_i; G, S), W_i); i = 1, \ldots, n\}$. An easy adaptation of this result allows O_i to be replaced by $O_i(t), i = 1, \ldots, n$.

S.4 Additional Results from Data Analysis Section

In this section, we further analyze the TRACE and Copenhagen datasets analyzed in Section 6 and also give results for two additional publicly available datasets, the Netherlands Breast Cancer Study and the R-Chop Study. For the both studies, the tuned version of the *CURE*$-L_2$ algorithms did not improve upon using the default node size value; hence, those results are not presented. Currently available software for the *RIST* algorithm cannot handle categorical covariates with more than two levels. The Netherlands Breast Cancer Study includes the variable tumor grade which is ordinal with three levels, which for the RIST method we include as a continuous variable.

Section S.4.1 provides prediction errors for the Netherlands data and the R-Chop data. Section S.4.2 gives results on importance of variables for the four datasets. Finally, Section S.4.3 reports prediction error plots for the four datasets when the censored data Brier score is evaluated at six different time units.

S.4.1 Prediction Error for Netherlands and R-Chop Datasets

In this subsection, we compare the performance of the *CURE-*L_2 algorithms *L2* and *L2 BJ* to the *RSF* and the *CI* algorithms using two publicly available datasets. These datasets are:

- *Netherlands Breast Cancer Study Data:* This dataset consists of 144 lymph node positive breast cancer patients and is included in the `R` package `penalized`. The event of interest is time to distant metastasis; subjects who were alive at the end of study, died from causes other than breast cancer, had recurrence of local or regional disease, or developed a second primary cancer were considered censored. The clinical factors measured are: number of affected lymph nodes, age, diameter of the tumor, estrogen receptor status, and grade of the tumor. Additionally, the dataset includes gene expression information for 70 genes that are used to build the prognostic model in both Van't Veer et al. (2002) and Van De Vijver et al. (2002). The censoring rate is 67%.

- *R-Chop Study Data:* This dataset consists of 233 patients with diffuse large B-cell lymphoma undergoing R-Chop treatment and followed until death. The dataset is publicly available in the `R` package `bujar`. The covariate vector consists of microarray data, with 3833 probe sets preselected from a set of 54675 probe sets as described in Wang and Wang (2010). The censoring rate is 74%.

As for the results for the TRACE and Copenhagen study presented in Section 6, we evaluate the Brier score at 3 time units (years for the R-Chop data and months for the Netherlands study). This corresponds to marginal survival probabilities (i.e., estimated using a Kaplan Meier curve) of 0.87 and 0.73 for the Netherlands and R-Chop datasets. Prediction errors for six other choices of t are presented in Section S.4.3. For the $CURE{-}L_2$ algorithms we only report the results for the default choice of `nodesize` (5) as tuning the `nodesize` as described in Section 6 did not result in substantial improvement in prediction accuracy. As in Section 6 we do not include comparisons to the Brier CURE algorithms.

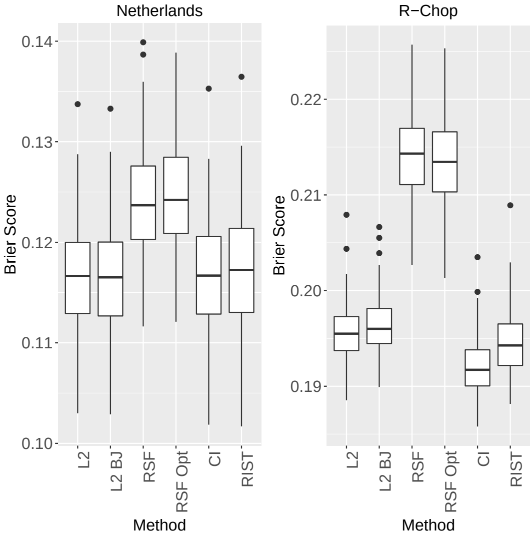


Figure S-17: Censored data Brier Score at $t = 3$ time units for the Netherlands and R-Chop datasets; lower values indicate better prediction accuracy. *L2* and *L2 BJ* are the *CURE*$-L_2$ algorithms with $Z = \log T$. *BJ* refers to the use of the Buckley-James CUT. *RSF* and *CI* are the default methods for the `rfsrc` and `cforest` `R` functions. *RSF Opt* refers to tuning the `nodesize` parameter for the *RSF* method. RIST is the recursively imputed survival trees algorithm.

S.4.2 Importance of Variables for the four Datasets

Table S-1 shows the OOB prediction error variable importance measures for the two *CURE*-L_2 algorithms and the *RSF* method for the TRACE dataset. In agreement with the results obtained from the minimal depth variable importance measures presented in Table 1 in Section 6, Table S-1 shows that ventricular fibrillation is consistently the least important predictor across all methods and age and CHF are consistently the two most influential variables.

The measure of prediction error used to calculate the OOB prediction error VIMPs for the RSF algorithm is the C-index. For the *CURE*-L_2 algorithms, the prediction error used to calculate the OOB prediction error VIMPs uses the $L_{2,d}(O_i, \psi; G, S)$ loss function. Because of the difference in the loss function used to calculate the OOB prediction error VIMPs, the *CURE*-L_2 and the *RSF* OOB prediction error VIMPs cannot be compared in terms of the numerical values but the ranking is comparable between the algorithms.

Tables S-2 and S-3 show the minimal depth and OOB prediction error variable importance measures for the 13 covariates in the Copenhagen Stroke Study. Both the *CURE*$-L_2$ algorithms for both variable importance measures identify the Scandinavian Stroke Score (a measure of stroke

	L2	*L2 BJ*	*RSF*
Age	0.54	0.26	0.08
Clinical Heart Pump Failure	0.25	0.23	0.04
Diabetes	0.17	0.19	0.00
Gender	0.01	0.03	0.01
Ventricular Fibrillation	-0.02	0.01	0.00

Table S-1: Out-of-bag prediction error variable importance measures for the TRACE data; higher values indicate more influential variables. *BJ* refers to the Buckley-James transformation. *RSF* is the default method in the `randomForestSRC` package.

	L2	*L2 BJ*	*RSF*
Scandinavian Stroke Score	0.96	0.94	1.35
Age	1.45	1.39	1.51
Cholesterol level	1.70	1.75	1.71
Atrial fibrillation	4.35	4.53	3.15
Stroke history	3.61	3.49	3.63
History of other disabling diseases	4.42	4.38	3.88
Diabetes	4.22	4.17	3.91
Hypertension	4.16	4.14	4.09
History of ischemic heart disease	4.47	4.27	4.10
Gender	4.50	4.43	4.32
Daily smoking status	4.57	4.55	4.38
Daily alcohol consumption	4.81	4.58	4.53
Hemorrhage	5.81	6.02	5.81

Table S-2: Minimal depth variable importance measures for the Copenhagen Study; lower values indicate more influential variables. *L2* and *L2 BJ* are the CURE L_2 algorithms. *BJ* refers to the Buckley-James transformation. *RSF* is the default method in the `randomForestSRC` package.

severity) as the most important predictor followed by age. Both these variables are known to impact overall survival probabilities for stroke patients.

The Netherlands and R-Chop datasets are comparatively high dimensional and tabular displays of variable importance are not especially informative. In the case of the Netherlands study (Van De Vijver et al., 2002), the genes included for analysis were already selected from a much larger pool, and one of the main conclusions in this study is that models that include these 70 gene expression profiles provide more information than models that do not rely on that information. In Figure S-18, we compare the prediction accuracy of the $CURE{-}L_2$ algorithms built using both clinical information and gene expression measurements to models built using only clinical information. Consistent with the findings in Van De Vijver et al. (2002), we see that adding gene expression information substantially improves the prediction power of the algorithms. In the case of the

	L2	*L2 BJ*	*RSF*
Scandinavian Stroke Score	0.57	0.58	$2.95 * 10^{-2}$
Age	0.18	0.19	$3.95 * 10^{-2}$
Cholesterol level	-0.01	0.01	$5.93 * 10^{-3}$
Daily smoking status	0.01	0.02	$7.83 * 10^{-4}$
Stroke history	0.01	0.00	$3.72 * 10^{-3}$
Diabetes	-0.01	-0.00	$1.12 * 10^{-3}$
Hypertension	0.01	0.01	$1.65 * 10^{-3}$
History of ischemic heart disease	0.02	0.03	$8.61 * 10^{-4}$
Daily alcohol consumption	-0.00	0.00	$1.61 * 10^{-3}$
Gender	0.02	0.01	$2.55 * 10^{-3}$
History of other disabling diseases	0.01	0.02	$2.94 * 10^{-3}$
Atrial fibrillation	-0.00	-0.01	$2.43 * 10^{-3}$
Hemorrhage	-0.00	-0.00	$-3.51 * 10^{-4}$

Table S-3: Out-of-bag prediction error variable importance measures for the Copenhagen Study; higher values indicate more influential variables. *L2* and *L2 BJ* refer to the two CURE L_2 algorithms. *BJ* refers to the Buckley-James transformation. *RSF* is the default method in the `randomForestSRC` package.

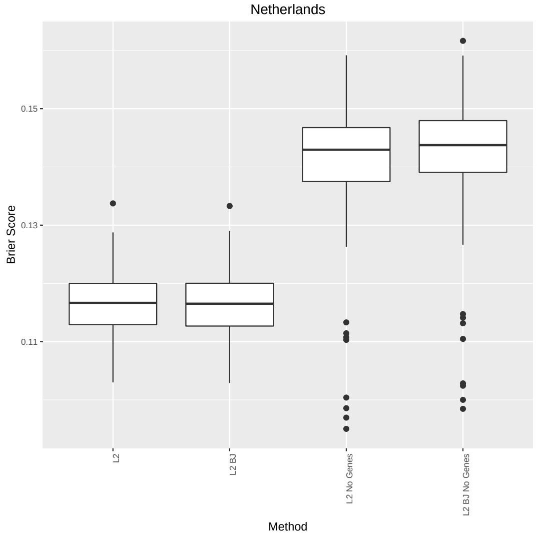


Figure S-18: Prediction error for four *CURE*$-L_2$ algorithms (with and without genetic information) on the Netherlands breast cancer study data. *L2* and *L2 BJ* are the *CURE*$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. *No Genes* refers to the model only being built using clinical factors.

R-Chop data, which involves 3833 probe sets, the minimal depth VIMP measures for both the *CURE*$-L_2$ algorithms and the *RSF* algorithm identify the same probe set as being the most influential. This probe set corresponds to a killer cell lectin-like receptor NKG2A that is a known natural killer (NK) cell receptor; see Brooks et al. (1997). Plonquet et al. (2007) found NK cell counts to be an important predictor for clinical outcomes in diffuse large B-cell lymphoma. We also calculated the OOB prediction error VIMP measures and then evaluated the degree of overlap between the 25 most influential probe sets for both VIMPs. The number of probe sets that were in the top 25 most influential variables for both VIMP measures was 13 and 10 for the doubly robust and Buckley-James *CURE*$-L_2$ algorithms, respectively.

S.4.3 Prediction Error Evaluated at Different Timepoints

Figures S-19-S-22 show prediction errors for the four datasets (TRACE, Copenhagen, Netherlands, and R-Chop) evaluated at the $10, 30, 40, 50, 60$ and 70-th quantiles of the observed times in the datasets. This corresponds to using t equal to $0.8, 3.6, 5.7, 6.3, 6.6,$ and 6.9, for the TRACE data, $0.3, 2.3, 3.6, 4.9, 6.0,$ and 7.7 for the Copenhagen study, $2.4, 5.4, 6.5, 7.2, 8.3,$ and 9.5 for the Netherlands study, and $0.4, 1.2, 1.6, 2.1, 2.7,$ and 3.3 for the R-Chop study. For computational simplicity, the results for the Trace data are only presented for the tuned *CURE*$-L_2$ algorithms and the *RIST* algorithm is not included in the comparisons.

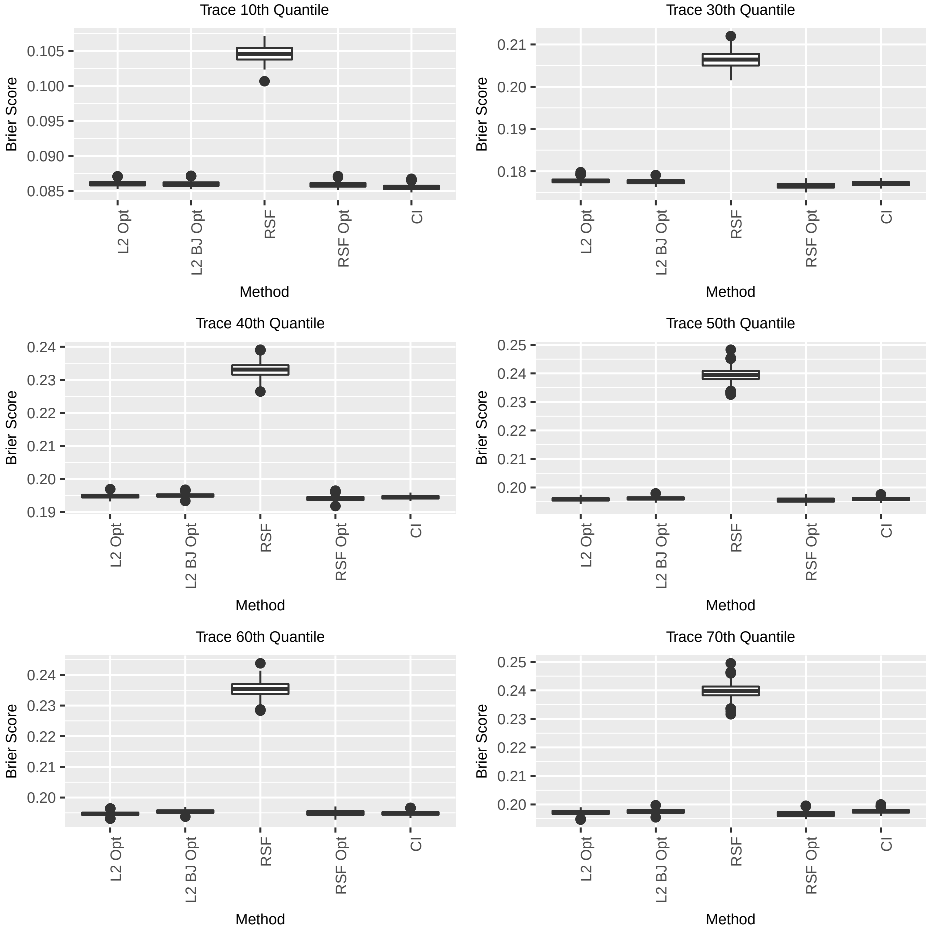


Figure S-19: Censored data Brier Score evaluated at six different time units for the Trace dataset; lower values indicate better prediction accuracy. The time units correspond to the $10, 30, 40, 50, 60$ and 70th quantiles of the observed times. *L2* and *L2 BJ* are the $CURE{-}L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. *RSF* and *CI* are the default methods for the `rfsrc` and `cforest R` functions. *RSF Opt* refers to tuning the `nodesize` parameter for the *RSF* method.

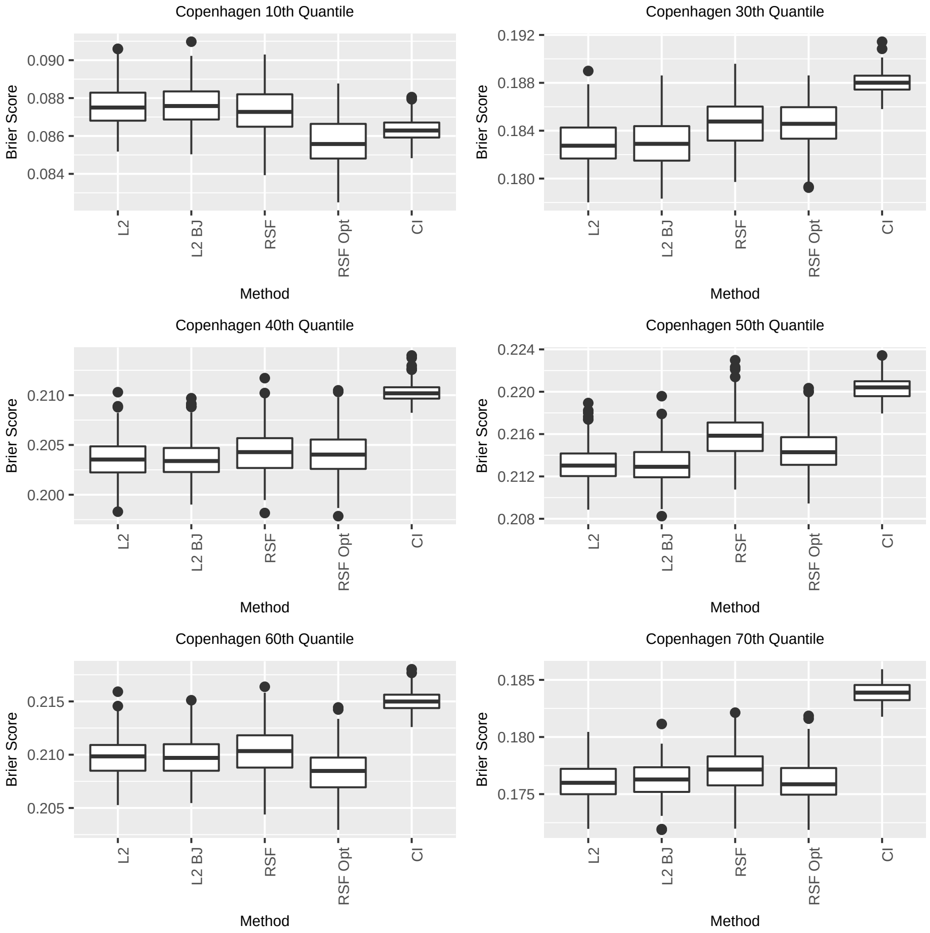


Figure S-20: Censored data Brier Score evaluated at six different time units for the Copenhagen dataset; lower values indicate better prediction accuracy. The time units correspond to the $10, 30, 40, 50, 60$ and 70th quantiles of the observed times. *L2* and *L2 BJ* are the *CURE*$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. *RSF* and *CI* are the default methods for the `rfsrc` and `cforest` `R` functions. *RSF Opt* refers to tuning the `nodesize` parameter for the *RSF* method.

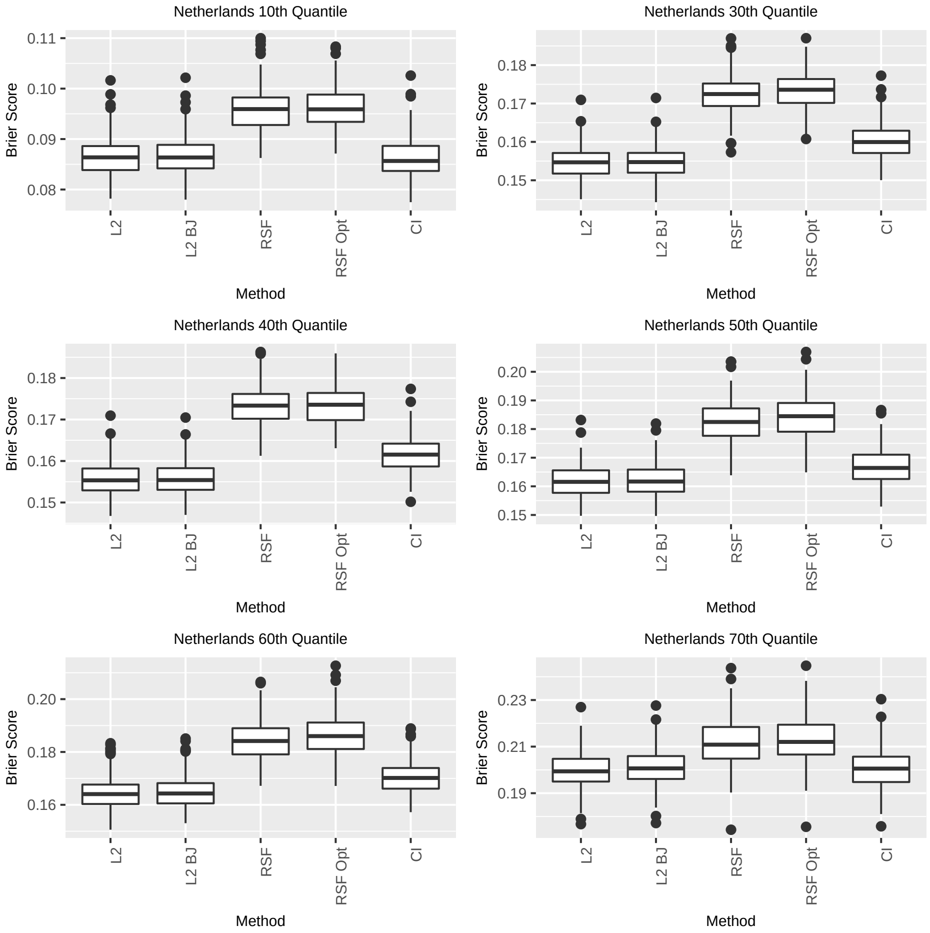


Figure S-21: Censored data Brier Score evaluated at six different time units for the Netherlands dataset; lower values indicate better prediction accuracy. The time units correspond to the $10, 30, 40, 50, 60$ and 70th quantiles of the observed times. *L2* and *L2 BJ* are the *CURE*$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. *RSF* and *CI* are the default methods for the `rfsrc` and `cforest` `R` functions. *RSF Opt* refers to tuning the `nodesize` parameter for the *RSF* method.

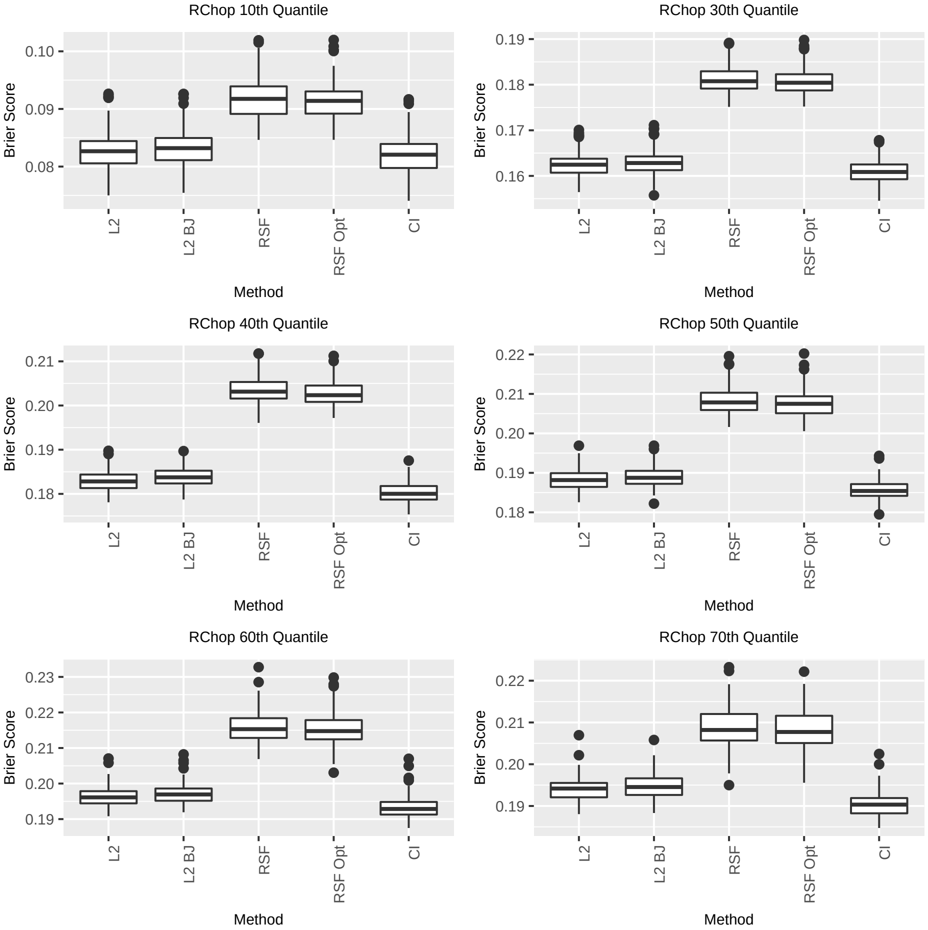


Figure S-22: Censored data Brier Score evaluated at six different time units for the R-Chop dataset; lower values indicate better prediction accuracy. The time units correspond to the 10, 30, 40, 50, 60 and 70th quantiles of the observed times. *L2* and *L2 BJ* are the *CURE*$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. *RSF* and *CI* are the default methods for the `rfsrc` and `cforest` `R` functions. *RSF Opt* refers to tuning the `nodesize` parameter for the *RSF* method.

S.5 Proof of Theorem 3.1

S.5.1 Regularity Conditions

The conditions of Section 3.2 specify that $S_0(t|w)$ and $G_0(t|w)$ are each continuous functions in $t \in \mathbf{R}^+$ for each $w \times \mathcal{S}$ and, in addition, that $\vartheta_{S_0} = \inf\{t : S_0(t|w) = 0\}$ and $\vartheta_{G_0} = \inf\{t : G_0(t|w) = 0\}$ are independent of $w \in \mathcal{S}$. The conditions of Section 3.1 imply that $\phi(h(u), w), (u, w) \in \mathbf{R}^+ \times \mathcal{S}$ is a known scalar function that is continuous in u except possibly at a finite number of points and bounded if $\max\{|r|, \|w\|\} < \infty$. Let $\mu(w) = E[\phi(Z, W)|W = w] = \int_0^\infty \phi(h(u), w) dF_0(u|w) < \infty$ for each $w \in \mathcal{S}$, where $F_0(u|w) = 1 - S_0(u|w)$. We assume that $S(t|w)$ and $G(t|w)$ are each right-continuous, non-increasing functions for $t \geq 0$ that satisfy $S(0|w) = G(0|w) = 1$, $S(t|w) \geq 0$ and $G(t|w) \geq 0$ for each $w \in \mathcal{S}$. Below, let $F(u|w) = 1 - S(u|w)$, $\bar{G}(u|w) = 1 - G(u|w)$, and $\bar{G}_0(u|w) = 1 - G_0(u|w)$.

The conditions imposed are weak enough to accomodate (4) as a special case of (3). For each $w \in \mathcal{S}$, we further assume

(C1) $I_1 = \int_0^\infty \phi(h(u), w) \frac{G_0(u|w)}{G(u-|w)} dF_0(u|w) < \infty$;

(C2) $M_1(r) = \int_0^r \frac{S_0(u|w)}{S(u|w)} \frac{d\bar{G}_0(u|w)}{G(u-|w)} < \infty$ and $M_2(r) = \int_0^r \frac{G_0(u|w) S_0(u|w)}{G(u|w) S(u|w)} \frac{d\bar{G}(u|w)}{G(u-|w)} < \infty$ for each $r > 0$;

(C3) $I_2 = \int_0^\infty \phi(h(u), w) \left[M_1(u-) - M_2(u-)\right] dF(u|w) < \infty$;

(C4) $\int_0^\infty \frac{[\phi(h(u),w)]^2}{G_0(u|w)} dF_0(u|w) < \infty$;

(C5) $\int_0^\infty \frac{m_\phi^2(u,w;S)}{G_0^2(u|w)} S_0(u|w) d\bar{G}_0(u|w) < \infty$;

(C6) $M_3(r) = \int_0^r \frac{|m_\phi(u,w;S)|}{G_0^2(u|w)} d\bar{G}_0(u|w) < \infty$ for each $r > 0$;

S.5.2 Proof that $Y_d^*(O; G, S_0)$, $Y_d^*(O; G_0, S)$ and $Y_d^*(O; G_0, S_0)$ are each CUTs for $Y = \phi(Z, W)$

Assume Conditions (C1)-(C3) hold. Then, calculations similar to those in Rubin and van der Laan (2007) show that $E[Y_d^*(O; G, S)|W = w] = I_1 + I_2$. Consider now the following cases:

- Suppose only that $G(u|w) = G_0(u|w)$ for every (u, w). Then, because $G_0(u|w)$ is continuous,

$G_0(u|w)/G_0(u-\,|w) = 1$ everywhere and it follows that

$$I_1 = \int_0^\infty \phi(h(u), w) dF_0(u|w) = \mu(w).$$

In addition, for every $r \geq 0$, we obtain

$$M_1(r) - M_2(r) = \int_0^r \frac{S_0(u|w)}{S(u|w)} \frac{d\bar{G}_0(u|w)}{G_0(u|w)} - \int_0^r \frac{G_0(u|w)}{G_0(u|w)} \frac{S_0(u|w)}{S(u|w)} \frac{d\bar{G}_0(u|w)}{G_0(u|w)} = 0$$

and hence $I_2 = 0$. Consequently, $E[Y_d^*(O; G_0, S)|W = w] = I_1 + I_2 = \mu(w)$.

- Suppose only that $S(u|w) = S_0(u|w)$ for every (u, w). Then,

$$I_1 + I_2 = \int_0^\infty \phi(h(u), w) \frac{G_0(u|w)}{G(u-\,|w)} dF_0(u|w) + \int_0^\infty \phi(h(u), w) \left[M_1(u-) - M_2(u-)\right] dF_0(u|w)$$

and we see that $E[Y_d^*(O; G, S_0)|W = w] = I_1 + I_2 = \mu(w)$ provided that

$$\frac{G_0(u|w)}{G(u-\,|w)} + [M_1(u-) - M_2(u-)] = 1.$$

Under the assumption $S(u|w) = S_0(u|w)$, the definitions of $M_i(\cdot), i = 1, 2$ and the fact that $G_0(u|w)$ is continuous implies that we need only show

$$\frac{G_0(u|w)}{G(u|w)} + \int_0^u \frac{d\bar{G}_0(r|w)}{G(r-\,|w)} - \int_0^u \frac{G_0(r|w)}{G(r|w)} \frac{d\bar{G}(r|w)}{G(r-\,|w)} = 1 \qquad \text{(S-6)}$$

for every $u \geq 0$. Using integration by parts (e.g., Last and Brandt, 1995, Thm. A.4.6),

$$\frac{G_0(u|w)}{G(u|w)} = 1 + \int_0^u G_0(r|w) \left(\frac{-dG(r|w)}{G(r-\,|w) G(r|w)} \right) + \int_0^u \frac{dG_0(r|w)}{G(r-\,|w)};$$

rearranging this expression, we see

$$\frac{G_0(u|w)}{G(u|w)} + \int_0^u \frac{d\bar{G}_0(r|w)}{G(r-\,|w)} - \int_0^u G_0(r|w) \left(\frac{d\bar{G}(r|w)}{G(r-\,|w) G(r|w)} \right) = 1$$

which is exactly (S-6). This proves $E[Y_d^*(O; G, S_0)|W = w] = \mu(w)$.

The result that $E[Y_d^*(O;G_0,S_0)|W=w]=\mu(w)$ clearly follows from either of the above arguments, completing this part of the proof.

S.5.3 Proof that $Var(Y_d^*(O;G_0,S)|W) \geq Var(Y_d^*(O;G_0,S_0)|W)$.

Let $G(u|w)=G_0(u|w)$ be continuous and consider the class of transformations

$$Y_s^*(O;G_0,\gamma)=\frac{\Delta\phi(\tilde{Z},W)}{G_0(\tilde{T}|W)}+(1-\Delta)\gamma(\tilde{T},W)-\int_0^{\tilde{T}}\gamma(u,W)d\Lambda_{G_0}(u|W), \tag{S-7}$$

where $\gamma(u,W)$ is some specified function. The class of transformations defined by (S-7) is essentially seen to be the same as that considered in Suzukawa (2004, Prop. 3; Eqn. 3.6), but generalized here to allow for covariates and not restricted to depend on $G_0(\cdot|\cdot)$ alone. Importantly, it is also easy to see that selecting $\gamma^*(u,W)=m(u,W;S)/G_0(u|W)$ in (S-7) gives $Y_s^*(O;G_0,\gamma^*)=Y_d^*(O;G_0,S)$. For continuous $G_0(u|w)$, the regularity conditions (C4)-(C6) generalize those in Suzukawa (2004) needed to prove Propositions 3, 5 and 6 in Suzukawa (2004). In particular, we have $E[Y_d^*(O;G_0,S)|W=w]=\mu(w)$ and, mimicking the arguments used to prove Propositions 5 and 6, that $Var[Y_d^*(O;G_0,S)|W=w]=H_1(w;G_0,S_0)+H_2(w;G_0,S_0,S)$, where

$$H_1(w;G_0,S_0)=\int_0^\infty\frac{[\phi(h(x),w)]^2}{G_0(x|w)}dF_0(x|w)-\int_0^\infty\frac{S_0(x|w)[m(x,w;S_0)]^2}{G_0(x|w)^2}d\bar{G}_0(x|w)-\mu^2(w)$$

and

$$H_2(w;G_0,S_0,S)=\int_0^\infty\frac{S_0(x|w)(m(x,w;S)-m(x,w;S_0))^2}{G_0(x|w)^2}d\bar{G}_0(x|w).$$

This proves $Var[Y_d^*(O;G_0,S)|W=w]\geq Var[Y_d^*(O;G_0,S_0)|W=w]=H_1(w,G_0,S_0)$, with strict inequality when $S(t|w)$ and $S_0(t|w)$ differ (hence $m(t,w;S)$ and $m(t,w;S_0)$ differ) for t in some interval with positive length.

S.6 Proof that $K(O_i,G)=1$ for all $i=1,\ldots,n$

In this section we proof that $K(O_i,G)=1$ for all $i=1,\ldots,n$. This follows from the following theorem upon making the identifications $D=\Delta_i$, $\tilde{t}=\tilde{T}_i$, and $w=W_i$.

Theorem S.6.1. *For each $w \in \mathcal{S}$, assume $G(0|w) = 1$, $G(u|w) \geq 0$, and that $G(u|w)$ is a right-continuous, non-increasing function for $u \geq 0$ with at most a finite number of discontinuities on any finite interval. Fix $w \in \mathcal{S}$, let $\tilde{t} > 0$ be finite, suppose $G(\tilde{t}|w) > 0$, and let D be any indicator variable taking on the value 0 or 1. Then,*

$$\frac{D}{G(\tilde{t}|w)} + \frac{(1-D)}{G(\tilde{t}|w)} - \int_0^{\tilde{t}} \frac{d\Lambda_G(u|w)}{G(u|w)} = 1.$$

To proof Theorem S.6.1 we require the following lemma.

Lemma S.6.2. *Let $x \geq 0$ be finite. Let d be any indicator variable taking on the values 0 and 1. Let $B(s)$ be a right-continuous, non-decreasing function for $s \geq 0$ with $B(0) = 0$. Define $\bar{B}(s) = 1 - B(s)$ and $H(s) = \int_0^s \bar{B}^{-1}(u-)dB(u)$ for any s such that $H(s)$ exists.*

Suppose $\bar{B}(x) > 0$. Then, $H(s)$ exists for $s \in [0, x]$ and

$$\frac{d}{\bar{B}(x)} + \frac{1-d}{\bar{B}(x)} - \int_0^x \frac{dH(u)}{\bar{B}(u)} = 1.$$

There is no specific relationship assumed between x, d and $B(\cdot)$, hence $H(\cdot)$. Using notation from both the theorem statement and Lemma S.6.2, we can make the following identifications: $\bar{B}(t) = G(t|w)$, $H(t) = \Lambda_G(t|w)$, $x = \tilde{t}$, and $d = D$. The conditions of Theorem ensure that the conditions of Lemma S.6.2 are satisfied; applying Lemma S.6.2 immediately gives the desired result:

$$\frac{D}{G(\tilde{t}|w)} + \frac{(1-D)}{G(\tilde{t}|w)} - \int_0^{\tilde{t}} \frac{d\Lambda_G(u|t)}{G(u|t)} = 1.$$

Proof of Lemma S.6.2. Because $\bar{B}(x) > 0$ and is non-increasing with $\bar{B}(0) = 1$, right-continuity implies $\inf_{s \leq x} \bar{B}(s) > 0$. Hence, we may write (e.g., Last and Brandt, 1995, Cor A.4.8, p. 426)

$$d\left(\frac{1}{\bar{B}(u)}\right) = -\frac{d\bar{B}(u)}{\bar{B}(u-)\bar{B}(u)} = \frac{dB(u)}{\bar{B}(u-)\bar{B}(u)} \tag{S-8}$$

Let $K(\cdot)$ be any right-continuous function of bounded variation on $[0, x]$. Then, we may write (Last and Brandt, 1995, Thm. A.4.6)

$$\frac{K(x)}{\bar{B}(x)} - \frac{K(0)}{\bar{B}(0)} = \int_0^x K(u-)\; d\left(\frac{1}{\bar{B}(u)}\right) + \int_0^x \left(\frac{1}{\bar{B}(u)}\right) dK(u).$$

Using (S-8) and assuming $K(s) = 1$ for $s \geq 0$, we obtain the identity

$$\frac{1}{\bar{B}(x)} - 1 = \int_0^x d\left(\frac{1}{\bar{B}(u)}\right) + \int_0^x \left(\frac{1}{\bar{B}(u)}\right) d(1) = \int_0^x \frac{dB(u)}{\bar{B}(u-)\bar{B}(u)} = \int_0^x \frac{dH(u)}{\bar{B}(u)}. \qquad \text{(S-9)}$$

Observe that we may also write

$$\frac{d}{\bar{B}(x)} = \frac{1}{\bar{B}(x)} - \frac{1-d}{\bar{B}(x)};$$

using (S-9), it then follows that

$$\frac{d}{\bar{B}(x)} + \frac{1-d}{\bar{B}(x)} - 1 = \int_0^x \frac{dH(u)}{\bar{B}(u)},$$

from which the required identity follows immediately. □

S.7 Proof of Theorem 4.1

Considering $G(\cdot|\cdot)$ and $S(\cdot|\cdot)$ as fixed functions, we will for simplicity rewrite $L_2(O, \psi(W); G, S, Q)$ as $L_2(O, \psi(W); Q) = \psi(W)^2 + H(O)\psi(W) + Q(O)$. Under the stated conditions, we can also assume without loss of generality that $L_2(O, \psi(W); Q) \geq 0$. The proof of this theorem will follow if one can show that all key decisions made by *CART* are invariant to the form of $Q(O)$. The availability of a sample $O_1, \ldots, O_n$ such that $H(O_i)$ and $Q(O_i)$ satisfy the conditions of the theorem for each $i = 1, \ldots, n$ is assumed. Throughout this proof, it is assumed at each stage of the algorithm that one is working with some finite partition $\{\tau_j, j = 1, \ldots, J\}$ of $\mathcal{S}$ and that $\psi(W) = \sum_{j=1}^J I\{W \in \tau_j\}\psi_j$ is the corresponding piecewise constant predictor. In this case, for any subset τ_j,

$$\sum_{i=1}^n I\{W_i \in \tau_j\} L_2(O_i, \psi(W_i); Q) = \sum_{i=1}^n I\{W_i \in \tau_j\}\left[\psi_j^2 + H(O_i)\psi_j + Q(O_i)\right]$$

is uniquely minimized at $\hat{\psi}_j = \sum_{i=1}^n -H(O_i)/(2n_j)$ for $n_j = \sum_{i=1}^n I\{W_i \in \tau_j\}$.

Now we show that all three steps of the *CART* algorithm used in connection with $L_2(O, \psi(W); Q)$ involve decisions that are invariant to the specification of $Q(O)$.

S.7.1 Growing the tree

The first stage of the tree building process is to grow a very large tree. Three elements are required to accomplish this step: (i) developing the candidate set of binary splits; (ii) specifying the node splitting rule; and, (iii) specifying the rule to stop splitting nodes. Only step (ii) depends on the specification of the loss function; hence, we focus on this step below. The following lemma is critical.

Lemma S.7.1. *Suppose $L_2(O, \psi(W); Q) = \psi(W)^2 + H(O)\psi(W) + Q(O)$ is used to evaluate the loss. Let $R(\tau)$ denote the loss within a given subset $\tau \subset \mathcal{S}$; that is, $R(\tau) = \sum_{i=1}^n I\{W_i \in \tau\} L_2(O_i, \widehat{\psi}_\tau; Q)$, where $\widehat{\psi}_\tau$ minimizes the loss function using the data falling into τ. If τ is then split into $L \geq 2$ mutually exclusive subsets $\tau_1, \ldots, \tau_L$ and $\tau_1 \cup \tau_2 \cup \ldots \cup \tau_L = \tau$, the corresponding change in total loss is given by*

$$R(\tau) - \sum_{\ell=1}^L R(\tau_\ell) = \sum_{\ell=1}^L \sum_{i=1}^n I\{W_i \in \tau_\ell\} \left[(\widehat{\psi}_\tau^2 - \widehat{\psi}_{\tau_\ell}^2) + H(O_i)(\widehat{\psi}_\tau - \widehat{\psi}_{\tau_\ell}) \right],$$

where $\widehat{\psi}_{\tau_\ell}$ is the value which minimizes the loss function using the data from the ℓth subset.

Proof. We have

$$R(\tau) = \sum_{i=1}^n I\{W_i \in \tau\} \left[\widehat{\psi}_\tau^2 + H(O_i)\widehat{\psi}_\tau + Q(O_i) \right]$$

and

$$\sum_{\ell=1}^L R(\tau_\ell) = \sum_{\ell=1}^L \sum_{i=1}^n I\{W_i \in \tau_\ell\} \left[\widehat{\psi}_{\tau_\ell}^2 + H(O_i)\widehat{\psi}_{\tau_\ell} + Q(O_i) \right].$$

Subtracting the second from the first, algebra shows that the change in total loss reduces to

$$R(\tau) - \sum_{\ell=1}^L R(\tau_\ell) \quad = \quad \sum_{\ell=1}^L \sum_{i=1}^n I\{W_i \in \tau_\ell\} \left[(\widehat{\psi}_\tau^2 - \widehat{\psi}_{\tau_\ell}^2) + H(O_i)(\widehat{\psi}_\tau - \widehat{\psi}_{\tau_\ell}) \right].$$

□

In the process of growing a tree, *CART* considers at each step all possible candidate splits of a given parent node τ into left and right child nodes, say τ_L and τ_R, and then chooses the (covariate, split) combination that maximizes the decrease $R(\tau) - R(\tau_L) - R(\tau_R)$. This process continues until the stop-splitting rule used in (iii) takes effect, generating a maximally-sized tree $\mathcal{T}_{max}$. Lemma S.7.1 shows that the reduction in loss is independent of $Q(O_i), i = 1 \dots n$ regardless of the stage of partitioning; hence, all splitting decisions made while growing the tree to its maximal size are invariant to the values of $Q(O_i), i = 1 \dots n$.

S.7.2 Pruning

Once a maximally-sized tree $\mathcal{T}_{max}$ is obtained, the second stage of the *CART* algorithm involves generating a sequence of candidate trees from which a final tree can be selected. The indicated sequence of candidate trees is generated using minimal cost-complexity pruning (Breiman et al., 1984, Sec. 3.3, 8.5).

For a given tree $\mathcal{T}$, let $\tilde{\mathcal{T}}$ and $N(\mathcal{T}) = \#(\tilde{\mathcal{T}})$ respectively denote the set and number of terminal nodes. Define the loss of the tree $\mathcal{T}$ as total loss in all terminal nodes: $R(\mathcal{T}) = \sum_{\tau \in \tilde{\mathcal{T}}} R(\tau)$. Finally, let the cost-complexity of a tree $\mathcal{T}$ be defined as $R_\alpha(\mathcal{T}) = R(\mathcal{T}) + \alpha N(\mathcal{T})$, where α is a non-negative real number called the complexity parameter.

Paraphrasing Breiman et al. (1984, Sec. 8.5), minimal cost complexity pruning generates a decreasing sequence of subtrees $\mathcal{T}_{max} \succ \mathcal{T}_1 \succ \mathcal{T}_2 \succ \cdots \succ \{\tau_1\}$ and an increasing sequence of complexity parameters $\alpha_1 < \alpha_2 < \cdots$ such that $\mathcal{T}_k$ is the smallest subtree of $\mathcal{T}_{max}$ for $\alpha_k \leq \alpha < \alpha_{k+1}$ that minimizes $R_\alpha(\mathcal{T})$. Breiman et al. (1984, Sec. 3.3) provide a detailed description of the process by which the sequence of subtrees is generated. Briefly, beginning with the smallest subtree $\mathcal{T}_1$ of $\mathcal{T}_{max}$ such that $R(\mathcal{T}_1) = R(\mathcal{T}_{max})$, *CART* begins the pruning process by considering all nodes τ from the tree T_1 and computing

$$g_1(\tau) = \begin{cases} \frac{R(\tau) - R(\mathcal{T}_{1,\tau})}{N(\mathcal{T}_{1,\tau}) - 1}, & \tau \notin \tilde{\mathcal{T}}_{1,\tau} \\ +\infty, & \tau \in \tilde{\mathcal{T}}_{1,\tau} \end{cases}$$

where $\tilde{\mathcal{T}}_{1,\tau}$ denotes the subtree of $\mathcal{T}_1$ with root node τ. The node(s) minimizing this function are pruned, yielding the next tree in the sequence $\mathcal{T}_2$. This process is repeated until the root node of $\mathcal{T}_1$ is reached.

Critically, the process for pruning any $\mathcal{T}_k$ and hence generating $\mathcal{T}_{k+1}$ depends on mimimizing

$$g_k(\tau) = \begin{cases} \frac{R(\tau)-R(\mathcal{T}_{k,\tau})}{N(\mathcal{T}_{k,\tau})-1}, & \tau \notin \tilde{\mathcal{T}}_k \\ +\infty, & \tau \in \tilde{\mathcal{T}}_k \end{cases}$$

for each $\tau \in \mathcal{T}_k$. Evidently, the function $g_k(\tau)$ depends on the loss function only through $R(\tau) - R(\mathcal{T}_{k,\tau})$; applying Lemma S.7.1 shows this quantity does not depend on $Q(O_i), i = 1 \ldots n$. As a result, the decision made to prune away any subtree and consequently the sequence of candidate trees generated by this process will be invariant to $Q(O_i), i = 1 \ldots n$.

S.7.3 Choosing the best candidate tree via cross-validation

Selection of the optimally sized tree from the sequence of candidate trees is done using V-fold cross validation. Specifically, suppose a given data set $\mathcal{O} = (O_1, \ldots, O_n)$ is divided into V mutually exclusive subsets $\mathcal{O}_1, \ldots, \mathcal{O}_V$. Suppose that the procedure of Section S.7.2 generated D trees with complexity parameters $\alpha_1, \ldots, \alpha_D$ using the loss function $L_2(O, \psi(W); Q) = \psi(W)^2 + H(O)\psi(W) + Q(O)$. Define $\gamma_1 = 0$, $\gamma_j = \sqrt{\alpha_j \alpha_{j+1}}, j = 2, \ldots, D-1$, and $\gamma_D = \infty$; see Breiman et al. (1984, Sec. 3.4 & 8.5.2) for discussion. For each $v = 1, \ldots, V$ let $\mathcal{T}_m(\mathcal{L}_{-v}), m = 1 \ldots D$ be a sequence of trees built using the learning set $\mathcal{L}_{-v} = \mathcal{O} - \mathcal{O}_v$ as follows: (i) growing a tree $\mathcal{T}_{max,v}$ as described in Section S.7.1; (ii) determining the associated sequence of pruned trees using minimal cost complexity pruning as described in Section S.7.2; and, (iii) identifying the sequence elements that correspond to using the complexity parameters $\gamma_1, \ldots, \gamma_D$. For $m = 1, \ldots, D$, let $\widehat{\psi}(W; \mathcal{T}_m(\mathcal{L}_{-v}))$ be the prediction for a subject with covariate information W that is obtained using the tree $\mathcal{T}_m(\mathcal{L}_{-v})$.

Then, the cross-validation error associated with γ_m is $C_m/(nV)$ where

$$\begin{aligned} C_m &= \sum_{v=1}^{V}\sum_{i=1}^{n} I(i \in \mathcal{O}_v) L_2(O_i, \widehat{\psi}(W_i; \mathcal{T}_m(\mathcal{L}_{-v})); Q) \ , \\ &= \sum_{v=1}^{V}\sum_{i=1}^{n} I(i \in \mathcal{O}_v) \left[Q(O_i) + H(O_i)\widehat{\psi}(W_i; \mathcal{T}_m(\mathcal{L}_{-v})) + \widehat{\psi}(W_i; \mathcal{T}_m(\mathcal{L}_{-v}))^2 \right] , \\ &= C^* + \sum_{v=1}^{V}\sum_{i=1}^{n} I(i \in \mathcal{O}_v) \left[H(O_i)\widehat{\psi}(W_i; \mathcal{T}_m(\mathcal{L}_{-v})) + \widehat{\psi}(W_i; \mathcal{T}_m(\mathcal{L}_{-v}))^2 \right] , \end{aligned} \tag{S-10}$$

for $m = 1, \ldots, D$ and $C^* = \sum_{v=1}^{V} \sum_{i=1}^{n} I(i \in \mathcal{O}_v) Q(O_i)$. The optimal tree is now given by $\mathcal{T}_{m^*}(\mathcal{O})$, where $m^* = \text{argmin}_{m \in \{1,\ldots,D\}} C_m$ and $\mathcal{T}_m(\mathcal{O})$ is the m^{th} candidate tree built using the full dataset $\mathcal{O}$ (i.e., that corresponding to α_m). Evidently, the constant C^* plays no role in selecting the member of the sequence that minimizes $C_m, m = 1, \ldots, D$ and hence selection of the optimally sized tree is also invariant to $Q(O_i), i = 1 \ldots n$.

References

Breiman, L. “Random forests.” *Machine Learning*, 45(1):5–32 (2001).

Breiman, L., Friedman, J. H., Stone, C. J., and Olshen, R. A. *Classification and Regression Trees*. Chapman and Hall/CRC (1984).

Brooks, A. G., Posch, P. E., Scorzelli, C. J., Borrego, F., and Coligan, J. E. “NKG2A complexed with CD94 defines a novel inhibitory natural killer cell receptor.” *The Journal of Experimental Medicine*, 185(4):795–800 (1997).

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. “Assessment and comparison of prognostic classification schemes for survival data.” *Statistics in Medicine*, 18(17-18):2529–2545 (1999).

Last, G. and Brandt, A. *Marked Point Processes on the Real Line: The Dynamical Approach*. Probability and Its Applications. New York: Springer-Verlag (1995).

LeBlanc, M. and Crowley, J. “Relative risk trees for censored survival data.” *Biometrics*, 48:411–425 (1992).

Liaw, A. and Wiener, M. “Classification and regression by randomForest.” *R News*, 2(3):18–22 (2002).
URL `http://CRAN.R-project.org/doc/Rnews/`

Lostritto, K., Strawderman, R. L., and Molinaro, A. M. “A partitioning deletion/substitution/addition algorithm for creating survival risk groups.” *Biometrics*, 68(4):1146–1156 (2012).

Plonquet, A., Haioun, C., Jais, J., Debard, A., Salles, G., Bene, M., Feugier, P., Rabian, C., Casasnovas, O., Labalette, M., et al. “Peripheral blood natural killer cell count is associated with clinical outcome in patients with aaIPI 2–3 diffuse large B-cell lymphoma.” *Annals of Oncology*, 18(7):1209–1215 (2007).

Rubin, D. and van der Laan, M. J. “A doubly robust censoring unbiased transformation.” *The International Journal of Biostatistics*, 3(1):1–21 (2007).

Steingrimsson, J., Diao, L., Molinaro, A. M., and Strawderman, R. L. “Doubly robust survival trees.” *Statistics in Medicine*, 35(17-18):3595–3612 (2016).

Suzukawa, A. “Unbiased estimation of functionals under random censorship.” *Journal of the Japan Statistical Society*, 34(2):153–172 (2004).

Van De Vijver, M. J., He, Y. D., van’t Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., et al. “A gene-expression signature as a predictor of survival in breast cancer.” *The New England Journal of Medicine*, 347(25):1999–2009 (2002).

Van’t Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., et al. “Gene expression profiling predicts clinical outcome of breast cancer.” *Nature*, 415(6871):530–536 (2002).

Wang, Z. and Wang, C. “Buckley-James boosting for survival analysis with high-dimensional biomarker data.” *Statistical Applications in Genetics and Molecular Biology*, 9(1) (2010).

Weng, C.-S. “On a second-order asymptotic property of the Bayesian bootstrap mean.” *The Annals of Statistics*, 705–710 (1989).