Supplement to: "Individualized Treatment Effects with Censored Data via Fully Nonparametric Bayesian Accelerated Failure Time Models"

A Posterior Computation

In the description of the Gibbs sampler, we use z_i to denote a latent variable that represents a transformed, imputed survival time, and we let y_i^c denote the "complete-data" survival times for the transformed survival times. That is, $y_i^{c,tr} = y_i^{tr}$ if $\delta_i = 1$ and $y_i^{c,tr} = z_i$ if $\delta_i = 0$. For posterior computation related to the Dirichlet process mixture, we let S_i denote the cluster to which the i^{th} observation has been assigned.

An outline of the steps used in a *single iteration* of our Gibbs sampler is provided below.

1. Using $\log y_i^{c,tr} - \tau_{S_i}$ as the responses, update trees $\mathcal{T}_1, \ldots, \mathcal{T}_J$ and node parameters $B_1, \ldots, B_J$ using the Bayesian backfitting approach of Chipman *and others* (2010). Using the updated $\mathcal{T}_1, \ldots, \mathcal{T}_J$ and $B_1, \ldots, B_J$, one may directly update $m(A_i, \mathbf{x}_i)$, for $i = 1, \ldots, n$.

2. Update cluster labels $S_1, \ldots, S_n$ by sampling with probabilities

$$P(S_i = h) \propto \pi_h \phi\Big(\frac{\log y_i^{c,tr} - m(A_i, \mathbf{x}_i) - \tau_h}{\sigma}\Big),$$

   and tabulate cluster membership counts $n_h = \sum_i \mathbf{1}\{S_i = h\}$.

3. Sample stick-breaking weights V_h, $h = 1, \ldots, H-1$ as $V_h \sim \text{Beta}(\alpha_h, \beta_h)$ where $\alpha_h = 1 + n_h$ and $\beta_h = M + \sum_{k=h+1}^{H} n_k$. Set $V_H = 1$. The updated mixture proportions are then determined by $\pi_h = V_h \prod_{k<h}(1 - V_k)$, for $h = 1, \ldots, H$.

4. Sample unconstrained cluster locations τ_h^*

$$\tau_h^* \sim \text{Normal}\Big(\frac{\sigma_\tau^2}{n_h\sigma_\tau^2 + \sigma^2}\sum_{i=1}^{n}\{\log y_i^{c,tr} - m(A_i, \mathbf{x}_i)\}\mathbf{1}\{S_i = h\}, \frac{\sigma_\tau^2\sigma^2}{n_h\sigma_\tau^2 + \sigma^2}\Big),$$

   and update constrained cluster locations $\tau_h = \tau_h^* - \mu_{G^*}$, where $\mu_{G^*} = \sum_{h=1}^{H} \pi_h \tau_h^*$.

5. Update mass parameter $M \sim \text{Gamma}\Big(\psi_1 + H - 1, \psi_2 - \sum_{h=1}^{H-1} \log(1 - V_h)\Big)$ and scale parameter $\sigma^2 \sim \text{Inverse-Gamma}\big(\frac{\nu+n}{2}, \frac{\hat{s}^2+\kappa\nu}{2}\big)$, where $\hat{s}^2$ is given by

$$\hat{s}^2 = \sum_{h=1}^{H}\sum_{i=1}^{n}\{\log(y_i^{c,tr}) - m(A_i, \mathbf{x}_i) - \tau_h\}^2 \mathbf{1}\{S_i = h\}.$$

6. For each $i \in \{k : \delta_k = 0\}$, update z_i by sampling

$$\log z_i \sim \text{Truncated-Normal}(m(A_i, \mathbf{x}_i) + \tau_{S_i}, \sigma^2; \log y_i^{tr}),$$

and set $y_i^{c,tr} = z_i$. Here, $X \sim \text{Truncated-Normal}(\mu, \sigma^2; a)$ means that X is distributed as $Z|Z > a$ where $Z \sim \text{Normal}(\mu, \sigma^2)$.

Because we use the transformed responses $\log(y_i^{tr}) = \log(y_i) - \hat{\mu}_{AFT}$ in posterior computation, we add $\hat{\mu}_{AFT}$ to the posterior draws of $m(A, \mathbf{x})$ in the final output.

B Additional Inferential Targets for HTE Analysis

B.1 Individual-level Survival Functions

In terms of the quantities of the non-parametric AFT model described in the main paper, individual-specific survival curves are defined by

$$P\{T > t|A, \mathbf{x}, m, G, \sigma\} = 1 - \int \Phi\Big(\frac{\log t - m(A, \mathbf{x}) - \tau}{\sigma}\Big) dG(\tau).$$

Using the truncated distribution G_H as an approximation in posterior computation, the survival curves are given by

$$P\{T > t|A, \mathbf{x}, m, G_H, \sigma\} = 1 - \sum_{h=1}^{H} \Phi\Big(\frac{\log t - m(A, \mathbf{x}) - \tau_h}{\sigma}\Big)\pi_h, \qquad (1)$$

which may be directly estimated using posterior draws of the regression function and τ_h, π_h.

Figure S1 shows estimated survival curves for randomly selected patients from the SOLVD treatment trial. Averages of these individual-level survival curves are computed for each treatment arm and compared with the corresponding Kaplan-Meier estimates of survival. It is apparent from Figure S1 that considerable heterogeneity in patient risk is present. Indeed, in the control arm, 20% percent of patients had an estimated median survival time less than 500 days, 54% had between 500 and 1500 days, and 26% had an estimated median survival time of more than 1500 days.

B.2 Treatment Allocation

Individualized treatment recommendations may be directly obtained by combining a fit of the non-parametric AFT model with a procedure minimizing the posterior risk associated with a chosen loss function. For instance, when trying to minimize the proportion of treatment misclassifications, one would assign treatment based on whether or not the posterior probability of the event $\{\theta(\mathbf{x}) > 0\}$ was greater than 0.5. Alternatively, one could optimize a weighted mis-classification loss where mis-classifications are weighted by the corresponding magnitude $|\theta(\mathbf{x})|$ of the treatment effect, in which case the optimal treatment decision would depend on the posterior mean of $\mathbf{1}\{\theta(\mathbf{x}) > 0\} \times |\theta(\mathbf{x})|$. Though we do not explore the issue here, such approaches to individualized treatment allocation could potentially be used, for example, in the development of adaptive randomization strategies for clinical trials.

C Simulation Study Using Friedman's Randomly Generated Functions

In this extra simulation study, we further evaluate the performance of the NP-AFTree using randomly generated nonlinear regression functions. To generate these random functions, we use a similar approach to that used in Friedman (2001) to assess the performance of gradient boosted regression trees. This approach allows us to test our approach on a wide range of difficult nonlinear regression functions that have higher-order interactions. For these simulations, we generated random regression functions $m(A, \mathbf{x})$ via

$$m(A, \mathbf{x}_i) = F_0(\mathbf{x}_i) + A_i\theta(\mathbf{x}_i),$$

where the functions $F_0(\mathbf{x})$ and $\theta(\mathbf{x})$ are defined as

$$F_0(\mathbf{x}_i) = \sum_{l=1}^{10} a_{1l} g_{1l}(\mathbf{z}_{1l}) \qquad \text{and} \qquad \theta(\mathbf{x}_i) = \sum_{l=1}^{5} a_{2l} g_{2l}(\mathbf{z}_{2l}). \tag{2}$$

The coefficients in (2) are generated as $a_{1l} \sim \text{Uniform}(-1, 1)$ and $a_{2l} \sim \text{Uniform}(-0.2, 0.3)$. The vector $\mathbf{z}_{jl}^{i}$ is a subset of $\mathbf{x}_i$ of length n_{jl} where the randomly selected indices used to construct the subset of $\mathbf{x}_i$ are the same for each i. The subset sizes are generated as

$n_{jl} = \min(\lfloor r_l + 1.5 \rfloor, 10)$ where $r_{jl} \sim \text{Exponential}(1/2)$.

$$g_{jl}(\mathbf{z}_{jl}) = \exp\Big\{ -\frac{1}{2}(\mathbf{z}_{jl} - \mu_{jl})^T \mathbf{V}_{jl}(\mathbf{z}_{jl} - \mu_{jl}) \Big\}.$$

The elements μ_{jlk} of the vector $\boldsymbol{\mu}_{jl}$ are generated as $\mu_{jlk} \sim \text{Normal}(0, 1)$, and the random matrix $\mathbf{V}_{jl}$ is generated as $\mathbf{V}_{jl} = \mathbf{U}_{jl}\mathbf{D}_{jl}\mathbf{U}_{jl}^T$, where $\mathbf{D}_{jl} = \text{diag}\{d_{jl,1}, \dots, d_{jl,n_{jl}}\}$ with $\sqrt{d_{jl,k}} \sim \text{Uniform}(0.1, 2)$ and where $\mathbf{U}_{jl}$ is a random orthogonal matrix. We generated the covariate vectors $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,20})^T$ of length 20 independently with $x_{i,k} \sim \text{Normal}(0, 1)$. Treatment assignments A_i were generated randomly with $P(A_i = 1) = 1/2$. These simulation settings imply that $\theta(\mathbf{x})$ is positive for roughly 87% of individuals. The parameters of the residual distributions were chosen so that the variances of each distribution were approximately equal.

Figure S2 shows simulation results for NP-AFTree, SP-AFTree, and the parametric AFT model. In this figure, we observe that root-mean squared error is broadly the same for the NP-AFTree and SP-AFTree methods with each of the tree methods exhibiting much better performance than Param-AFT. This similarity in RMSE of SP-AFTree and NP-AFTree seems attributable to the difficulty of estimating these regression functions which seems to overwhelm most of the advantages of more flexible modeling of the residual distribution. Compared to SP-AFTree, NP-AFTree shows modestly better classification performance, particularly for settings that have non-Gaussian residual distributions and for settings with the larger ($n = 1,000$) sample size. However, as in the simulations of Section 4.1 of the main paper, there seems to be no advantage here of either NP-AFTree, SP-AFTree, or Param-AFT over the naive treatment allocation approach when the sample size is $n = 200$. These results suggest that fairly large sample sizes may be needed for there to be any advantage over the naive approach which simply allocates individuals to the treatment having the more beneficial overall treatment effect. For NP-AFTree the average coverage is consistently a few percentage points below the desired 95% level suggesting that modest under-coverage can occur in certain settings. Numerical values corresponding to Figure S2 are shown in Table S3.

D Choice of Prior over Splitting Values

As described in Chipman *and others* (1998) and Chipman *and others* (2010), the prior on the splitting values c used at each internal node is uniform over the finite set of available splitting values for the chosen splitting variable. In implementations of BART, the number of possible available splitting values is typically truncated so that it cannot exceed a prespecified maximum value. The default setting used in the `BayesTree` package (Chipman and McCulloch (2016)) has a maximum of 100 possible split points for each covariate, and the default is to assign a uniform prior over potential split points that are equally spaced over the range of the covariate. An alternative option offered in `BayesTree` is to, for each covariate, assign a uniform prior over the observed quantiles of the covariate rather than the uniform prior over the observed range of the covariate. Our default choice is to use the uniform prior over covariate quantiles for the split point prior rather than the uniform prior over equally spaced points. With this quantile-based prior, we found, in many simulations, improved performance in terms of coverage.

E Approximate Distribution of the Residual Variance

As discussed in Section 2.4 of the main paper, the variance of the residual term may be expressed as

$$\mathrm{Var}(W|G,\sigma) = \sigma^2 + \sigma_\tau^2 \sum_{h=1}^{\infty} \frac{\pi_h}{\sigma_\tau^2}(\tau_h^* - \mu_{G^*})^2 \tag{3}$$

When assuming (as we do) that $\sigma_\tau^2 = \kappa$, this becomes

$$\mathrm{Var}(W|G,\sigma) = \sigma_\tau^2 \Big[\sigma^2/\kappa + \sum_{h=1}^{\infty} \frac{\pi_h}{\sigma_\tau^2}(\tau_h^* - \mu_{G^*})^2\Big] \tag{4}$$

Because we assume that $G \sim CDP(M, G_0)$ with G_0 as a Normal$(0, \sigma_\tau^2)$ distribution, the term $[(\tau_h^* - \mu_{G^*})^2]/\sigma_\tau^2$ has a standard normal distribution.

In Section 2.4 of the main paper, it is stated that the prior distribution of $\mathrm{Var}(W|G,\sigma)$ is approximated with the following distribution

$$\sigma_\tau^2\big[\nu/\chi_\nu^2 + \mathrm{Normal}(1, \{2(M+1)\}^{-1})\big]. \tag{5}$$

The above approximation relies on the fact that $\sum_{h=1}^{\infty} \frac{\pi_h}{\sigma_\tau^2}(\tau_h^* - \mu_{G^*})^2$ has an approximate Normal$(0, \{2(M+1)\}^{-1})$ distribution in the sense described by Yamato (1984). A histogram of simulated values of $\sum_{h=1}^{\infty} \frac{\pi_h}{\sigma_\tau^2}(\tau_h^* - \mu_{G^*})^2$ along with a plot of the approximating Normal$(0, \{2(M+1)\}^{-1})$ density is shown in Figure S3. In this figure, histograms are shown for the cases of $M = 25$ and $M = 50$.

In Figure S4, we display a quantile-quantile plot of simulated values from the distribution of $\text{Var}(W|G,\sigma)$ vs. the approximate theoretical quantiles obtained from the approximate prior distribution stated in (5).

F Cross-Validation across Hyperparameter Settings

When fitting the NP-AFT model with the SOLVD data, we considered several settings for the hyperparameters, and for each setting of the hyperparameters, we computed cross-validation scores to evaluate performance in terms of predicting patient outcomes and in terms of characterizing HTE. For evaluating predictions of patient outcomes, we utilize, as in Tian *and others* (2014) and Tian *and others* (2007), a direct measure of absolute prediction error. In particular, for the k^{th} test set $\mathcal{D}_k$, we compute the following cross-validation score

$$CV_k^{abs} = \frac{1}{n_k} \sum_{i \in \mathcal{D}_k} \frac{\delta_i}{\hat{V}(Y_i|A_i, \mathbf{x}_i)} \Big| \log Y_i - \hat{m}_{-\mathcal{D}_k}(A_i, \mathbf{x}_i) \Big|, \tag{6}$$

where n_k is the number of patients in $\mathcal{D}_k$ and $\hat{m}_{-\mathcal{D}_k}(A, \mathbf{x})$ is the regression function estimated from the k^{th} training set. The weights used in (6) $\hat{V}(Y_i|A_i, \mathbf{x}_i)$ are estimates of the censoring probability $V(t|A, \mathbf{x}) = P(C > t|A, \mathbf{x})$. The total K-fold cross-validation error is computed as $K^{-1}\sum_{k=1}^{K} CV_k^{abs}$.

Figure S5 shows results from applying cross-validation to the SOLVD trials with 36 different settings of the hyperparameters. The censoring probabilities $\hat{V}(Y_i|A, \mathbf{x}_i)$ used as weights in (6) were estimated using a Cox model. The 36 hyperparameter settings were generated by varying the hyperparameter q which determines the parameters of the base distribution G_0, the hyperparameter k that determines the prior variance of the node values, and the number of trees J. We varied q across the four levels, $q = 0.25, 0.5, 0.90, 0.99$; k across the three levels, $k = 1, 2, 3$; and the number of trees J across the three levels, $J = 50, 200, 400$. Ten-fold cross-validation was used for each setting of the hyperparameters. As shown in Figure S5, the hyperparameter q appears to play the most important role in driving the differences in cross-validation performance while larger values of the shrinkage parameter k seem to have a modest beneficial effect in the $q = 0.25$ and $q = 0.5$ settings. The settings with the very conservative choice of $q = 0.99$ exhibit poor performance giving similar cross-validation scores as a parametric AFT model with an assumed linear model for the regression function. The setting with the best cross-validation score was $q = 0.5, k = 3, J = 400$. This cross-validation score, however, was not notably different than many of the settings with either $q = 0.25$ and $q = 0.5$. For this reason, we continued to use the default setting of $q = 0.5, k = 2$, and $J = 200$ in our analysis of the SOLVD trials.

G Simulation Results

More detailed simulation results from the simulation study described in Section 4.1 are shown in Table S1. Table S1 corresponds to Figure 1 in the main paper.

H Handling Missing Covariates

Currently, our software does not support analyses where missing values of the patient covariates are present. Nevertheless, a number of missing-data models could be directly incorporated into our nonparametric AFT model. We describe here two missing-data approaches and how they would fit into our BART-based AFT model. Each of these approaches could be directly integrated into our posterior sampling scheme described in Section A by adding an additional step (or series of extra steps). In each case, one would need to include an extra Gibbs step to sample (for each patient i that has missing covariate values) from the conditional distribution $\mathbf{x}_{i,mis}|\mathbf{x}_{i,obs}$ where $\mathbf{x}_{i,mis}$ denotes the missing covariate values of patient i and $\mathbf{x}_{i,obs}$ denotes the set of observed covariate values for patient i. If there are additional parameters (e.g., $\boldsymbol{\alpha}$) governing the missing-data model, these would be updated after first sampling from $\mathbf{x}_{i,mis}|\mathbf{x}_{i,obs}, \boldsymbol{\alpha}$.

One approach is to use the parametric class of covariate models described in Ibrahim *and others* (1999). Here, the joint distribution of the covariate vector for patient i, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, is factored as

$$p(x_{i1}, \ldots, x_{ip}|\boldsymbol{\alpha}) = p(x_{ip}|x_{i,p-1}, \ldots, x_{i1}, \boldsymbol{\alpha}_p)p(x_{ip-1}|x_{i,p-2}, \ldots, x_{i1}, \boldsymbol{\alpha}_{p-1}) \cdots p(x_{i2}|x_{i1}, \boldsymbol{\alpha}_1), \tag{7}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_p^T)^T$. Each one of the above p conditional distributions could be modeled using a regression model with the form of the regression depending on the type of covariate (i.e., continuous, binary, or categorical). For example, if x_{ij} is a continuous covariate, one could assume that $x_{ij}|x_{i,j-1}, \ldots, x_{i1}, \boldsymbol{\alpha}_j \sim \text{Normal}(\alpha_{0j} + \sum_{k=1}^{j-1} \alpha_{kj} x_{ik}, \sigma_{\alpha,j}^2)$ where is $\boldsymbol{\alpha}_j$ is the vector of parameters $\boldsymbol{\alpha}_j = (\alpha_{0j}, \ldots, \alpha_{jj}, \sigma_{\alpha,j}^2)^T$. Similarly, if x_{ij} is a binary covariate, one could assume that $x_{ij}|x_{i,j-1}, \ldots, x_{i1}, \boldsymbol{\alpha}_j \sim \text{Bernoulli}\big(\text{logit}^{-1}(\alpha_{0j} + \sum_{k=1}^{j-1} \alpha_{kj} x_{ik})\big)$ where in this case $\boldsymbol{\alpha}_j$ would be the parameter vector $\boldsymbol{\alpha}_j = (\alpha_{0j}, \ldots, \alpha_{jj})^T$. See Ibrahim *and others* (2001) for further discussion of posterior computation and choice of prior distributions for the parameters $\boldsymbol{\alpha}$ governing the covariate distribution (7).

Another class of missing-data models is the nonparametric sequential model described in Xu *and others* (2016). With this approach, the conditional distribution of each covariate x_{ij} given the remaining covariates is modeled with BART. For example, if x_{ij} is continuous, it is assumed that the distribution of x_{ij} given the remaining covariates is Gaussian with mean function $\mu_j(\cdot)$ and variance σ_j^2. This mean function $\mu_j(\cdot)$ is then modeled using BART. Likewise, if x_{ij} is binary, one instead assumes that x_{ij} has a Bernoulli distribution with success probability $h_j(\cdot)$ where $h_j(\cdot)$ is again modeled using BART. With this approach, the imputed covariate values are sampled one covariate at-a-time and then the BART parameters (for the missing data model) are updated directly after imputing the missing covariate values. This missing-data model would fit quite naturally into our posterior sampling scheme as it is based on a collection of BART models. However, because it involves updating the parameters of p separate BART models within each Gibbs iteration, this approach may be very computationally intensive for even moderately large values of p.

References

Chipman, H. and McCulloch, R. (2016). *BayesTree: Bayesian Additive Regression Trees*. R package version 0.3-1.4.

Chipman, H. A., George, E. I. and McCulloch, R. E. (1998). Bayesian CART model search (with discussion and a rejoinder by the authors). *Journal of the American Statistical Association* **93**(443), 935–960.

Chipman, H. A., George, E. I. and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **4**(1), 266–298.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**(5), 1189–1232.

Ibrahim, J. G., Chen, M.H. and Lipsitz, S. R. (1999). Monte carlo EM for missing covariates in parametric regression models. *Biometrics* **55**(453), 591–596.

Ibrahim, Joseph G., Chen, Ming Hui and Sinha, Debajyoti. (2001). *Bayesian Survival Analysis*. New York, NY: Springer-Verlag.

Tian, L., Cai, T., Goetghebeur, E. and Wei, L.J. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* **94**(2), 297–311.

Tian, L., Zhao, L. and Wei, L.J. (2014). Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics* **15**(2), 222–233.

Xu, Dandan, , Daniels, Michael J. and Winterstein, Almut J. (2016). Sequential BART for imputation of missing covariates. *Biostatistics* **17**(3), 589–602.

Yamato, H. (1984). Characteristic functions of means of distributions chosen from a Dirichlet process. *The Annals of Probability* **12**(1), 262–267.

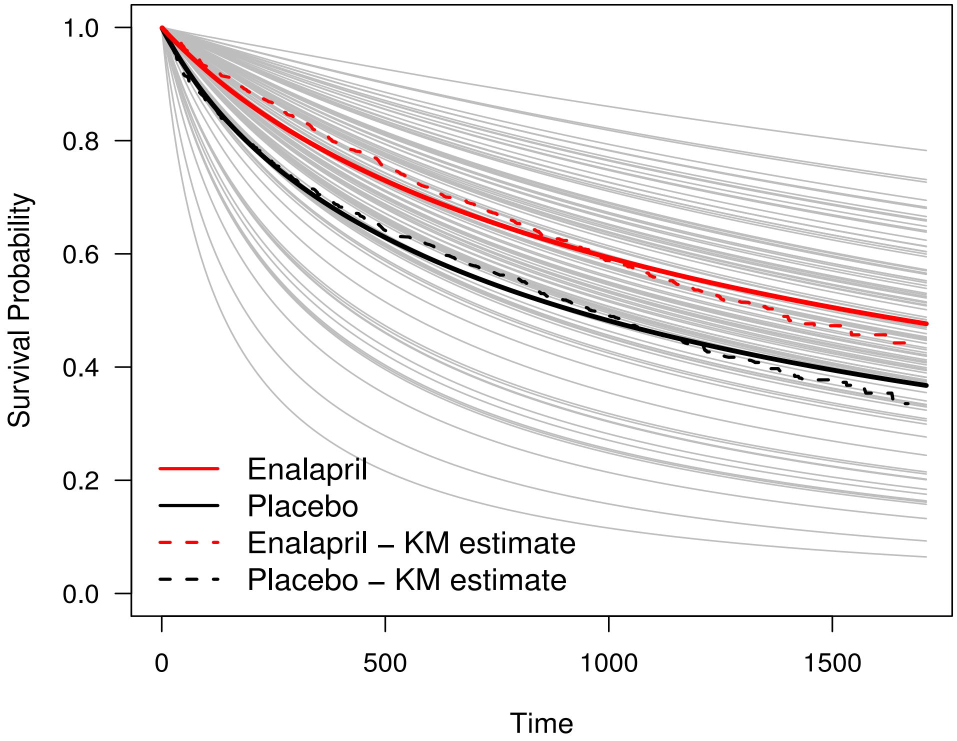

Figure S1: Estimates of individual-specific survival curves for selected patients from the SOLVD treatment trial. For each patient, the posterior mean of the survival functions $P\{T > t|A, \mathbf{x}, m, G_H, \sigma\}$ as defined in (1) are plotted. The solid black and red survival curves are the average by treatment group of these estimated individual-specific surves. The dashed survival curves are the Kaplan-Meier estimates for each treatment group.

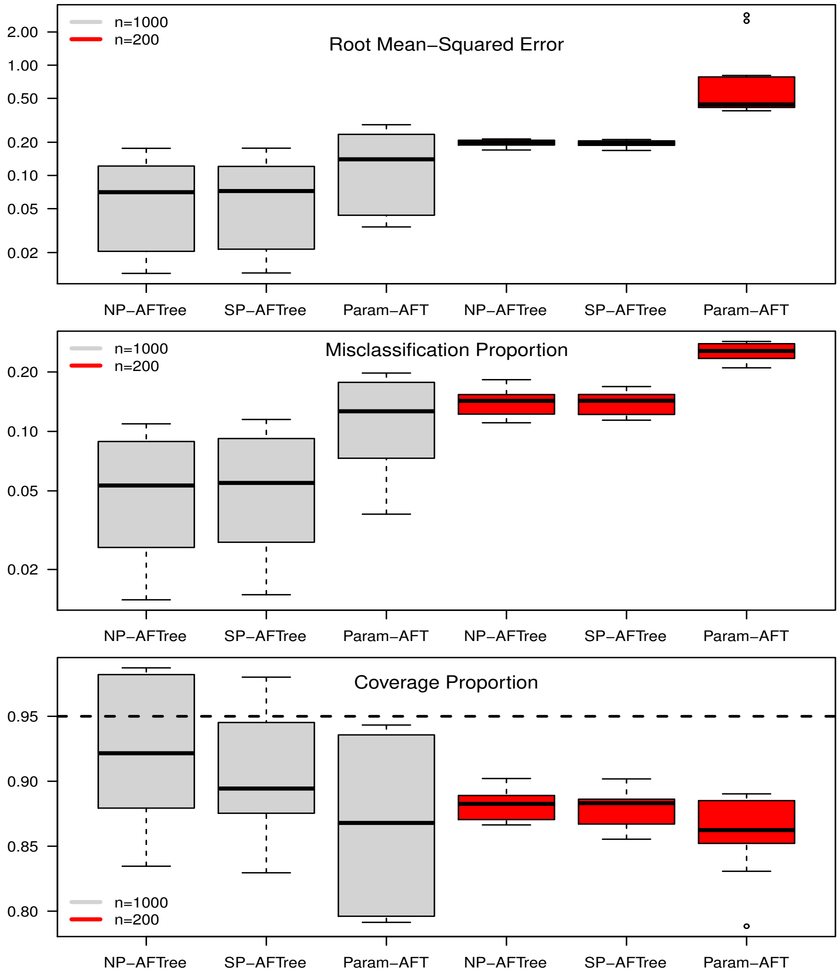

Figure S2: Simulations for AFT models with randomly generated regression functions. Results are based on 50 simulation replications. Root mean-squared error, misclassification proportion, and empirical coverage are shown for each method. Performance measures are shown for the non-parametric tree-based AFT (NP-AFTree) method, the semi-parametric tree-based AFT (SP-AFTree), and the parametric, linear regression - based AFT (Param-AFT) approach. Four different choices of the residual distribution were chosen: a Gaussian distribution, a Gumbel distribution with mean zero, a "standardized" Gamma distribution with mean zero, and a mixture of three t-distributions with 3 degrees of freedom for each mixture component.

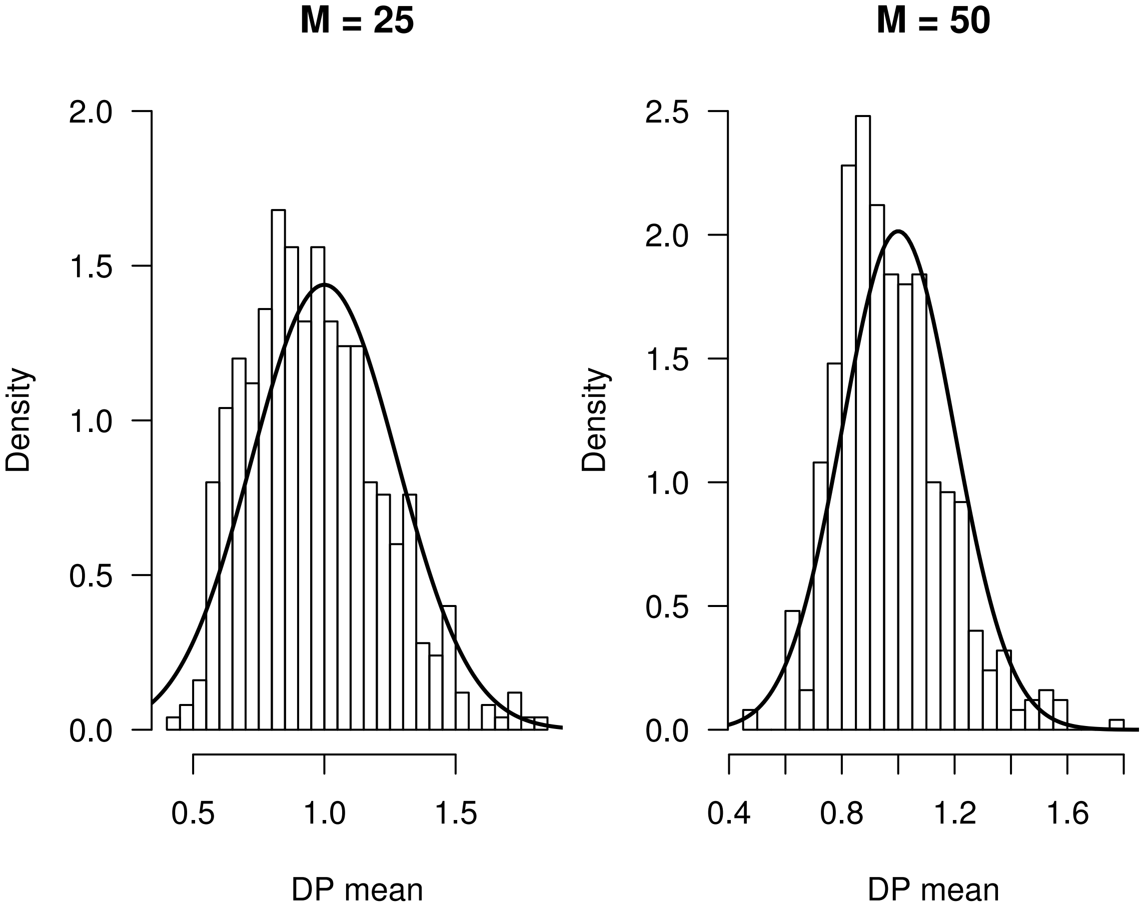

Figure S3: Histogram of simulated values of $\sum_{h=1}^{\infty} \frac{\pi_h}{\sigma_\tau^2}(\tau_h^* - \mu_{G^*})^2$ along with the approximating Normal$\{1, 2/(M+1)\}$ density. Simulations were performed with $M = 25$ and $M = 50$ for the mass parameter. In each case, 500 simulated values of a were drawn.

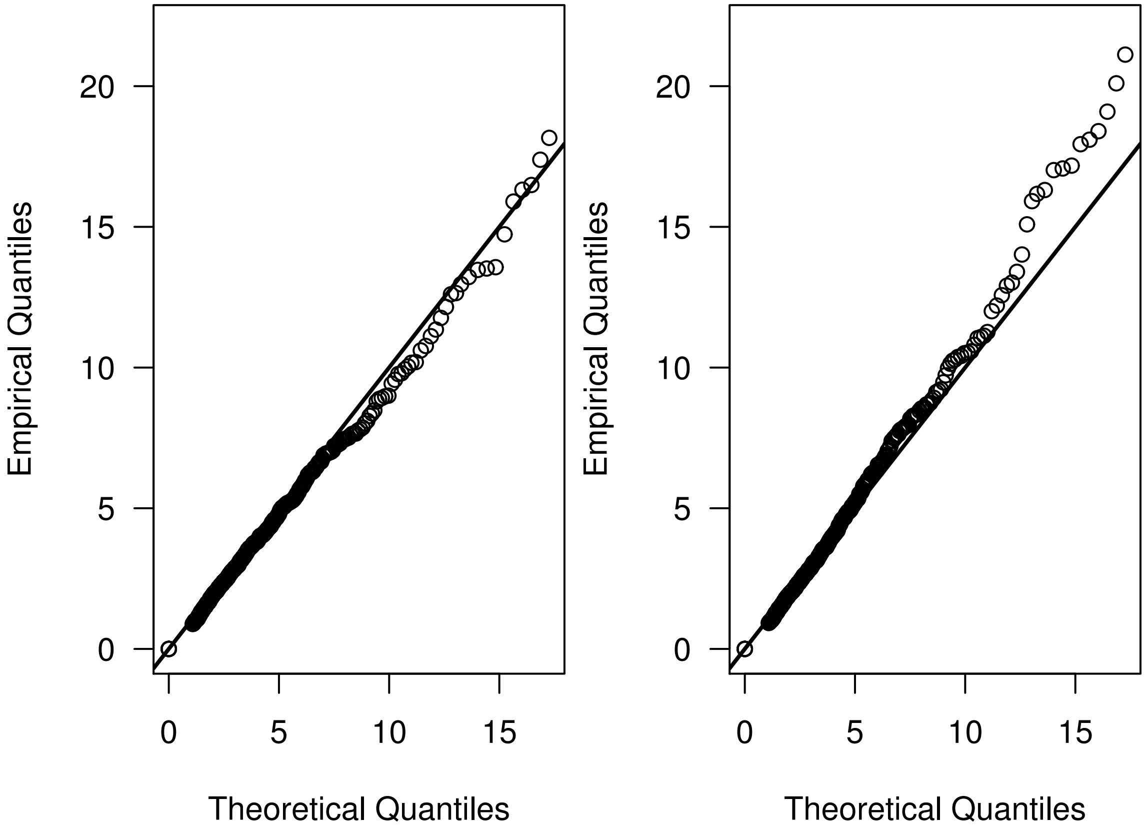

Figure S4: Quantile-Quantile plot with simulated values of $\mathrm{Var}(W|G, \sigma)$ (using equation).

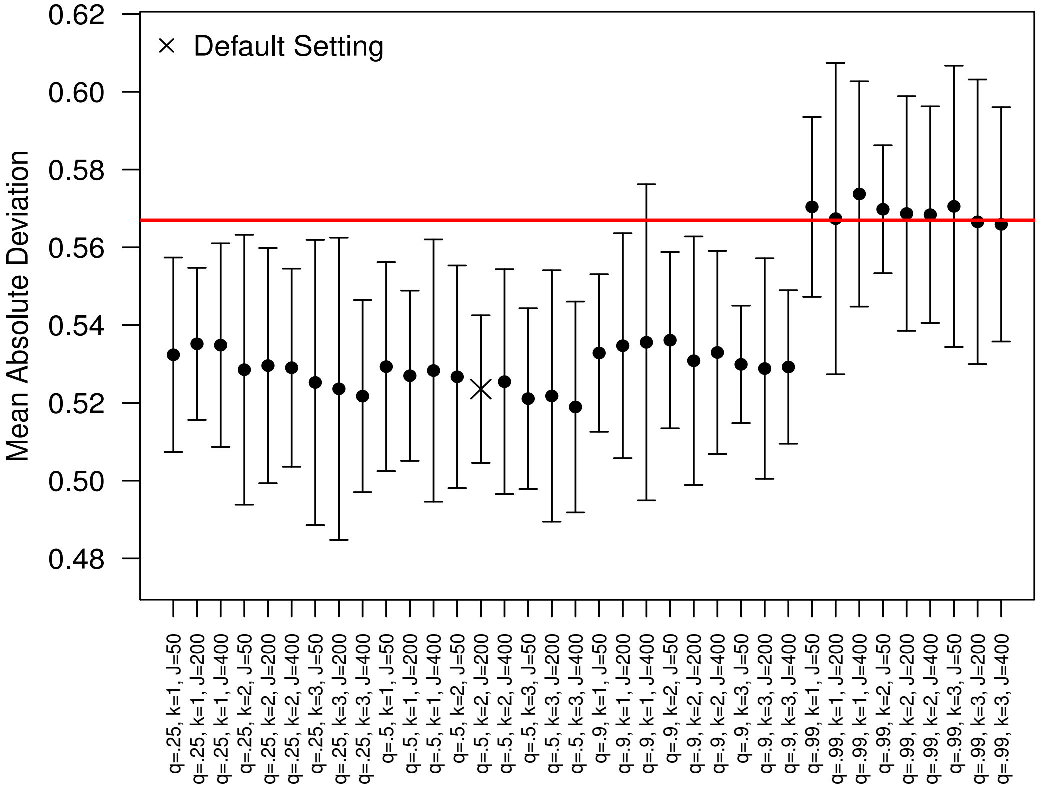

Figure S5: Ten-fold cross-validation for the SOLVD-T and SOLVD-P trials using the mean absolute deviation estimate defined in (6). Twenty seven settings of the hyperparameters are considered. The cross-validation score for the default setting of the hyperparameters is marked with an ×. The horizontal red line denotes the ten-fold cross-validation score of a parametric AFT model with log-normal errors where a linear regression model with treatment-covariate interactions was assumed.

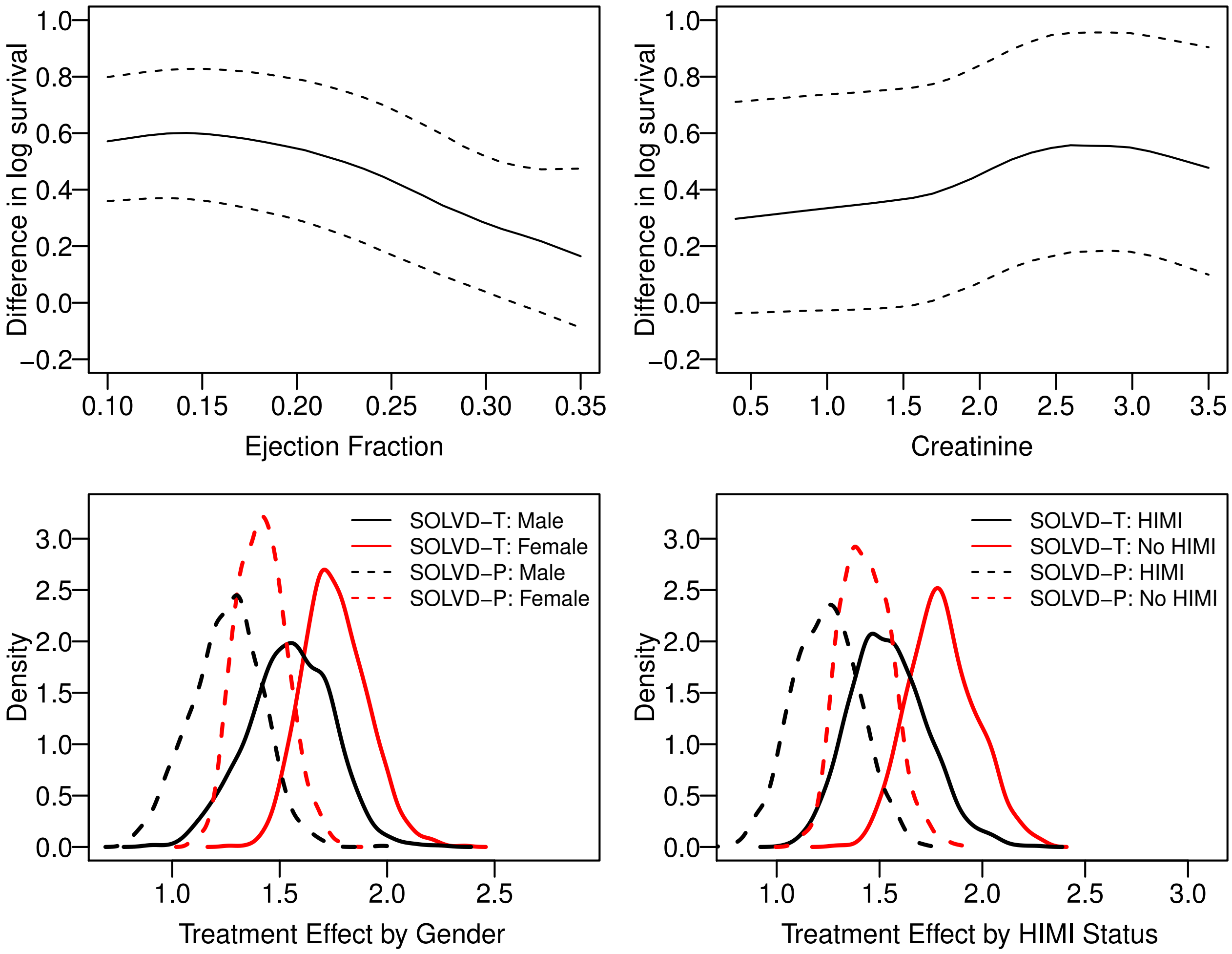

Figure S6: Smoothed partial dependence plots for ejection fraction and creatinine levels, and posterior distributions of treatment effect for men vs. women and for those with a history of myorcardial infarction vs. those with no history of myocardial infarction (HIMI vs. No HIMI).

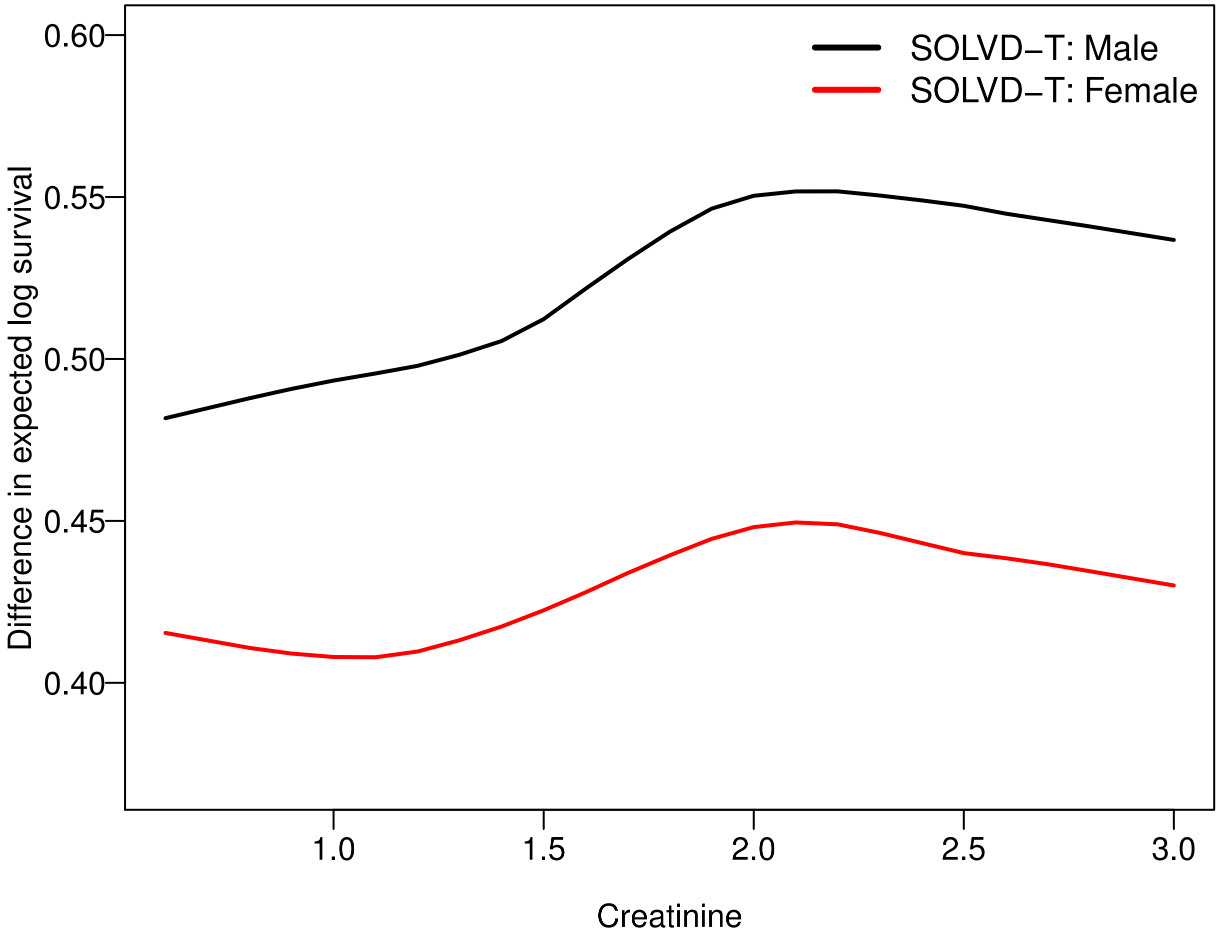

Figure S7: Estimated partial dependence function for baseline creatinine - separated by gender. These were estimated only using data from the SOLVD-T trial. The gender-subsetted partial dependence functions are defined by only averaging over those patients with a specific gender rather than averaging over all the patients in the study. Specifically, the male partial dependence function (for the l^{th} covariate) is defined as $\rho_l^{male}(z) = \frac{1}{n_{male}} \sum_{i=1}^{n} a_{i,male}\theta(z, \mathbf{x}_{i,-l})$, where $a_{i,male} = 1$ if patient i is male, $a_{i,male} = 0$ otherwise, and $n_{male} = \sum_{i=1}^{n} a_{i,male}$. Similarly, the female partial dependence function is defined as $\rho_l^{female}(z) = \frac{1}{n_{female}} \sum_{i=1}^{n} a_{i,female}\theta(z, \mathbf{x}_{i,-l})$.

Table S1: Simulation Results for the regression functions based on the SOLVD trial data. Root mean-squared error (RMSE), mis-classification proportion (MCprop), and empirical coverage are shown for each of the methods. Performance measures are shown for the non-parametric tree-based AFT (NP-AFTree) method, the semi-parametric tree-based AFT (SP-AFTree), and the parametric, linear regression - based AFT (Param-AFT) approach. Four different choices of the residual distribution were chosen: a Gaussian distribution, a Gumbel distribution with mean zero, a "standardized" Gamma distribution with mean zero, and a mixture of three t-distributions with 3 degrees of freedom for each mixture component.

			NP-AFTree			SP-AFTree			Param-AFT			Naive
Distribution	n	Censoring	RMSE	MCprop	Cover	RMSE	MCprop	Cover	RMSE	MCprop	Cover	MCProp
Normal	200	none	0.053	0.144	0.995	0.060	0.155	0.986	0.148	0.294	0.898	0.115
	200	light	0.054	0.129	0.998	0.059	0.138	0.986	0.151	0.290	0.904	0.115
	200	heavy	0.058	0.142	1.000	0.068	0.163	0.994	0.186	0.338	0.897	0.115
	1000	none	0.013	0.014	0.951	0.013	0.014	0.925	0.034	0.036	0.919	0.559
	1000	light	0.013	0.016	0.965	0.014	0.017	0.924	0.036	0.044	0.921	0.559
	1000	heavy	0.016	0.019	0.974	0.018	0.023	0.914	0.044	0.077	0.937	0.559
Gumbel	200	none	0.055	0.132	0.996	0.061	0.140	0.988	0.147	0.282	0.901	0.115
	200	light	0.055	0.137	0.998	0.061	0.142	0.992	0.157	0.298	0.909	0.115
	200	heavy	0.063	0.182	1.000	0.075	0.208	0.995	0.198	0.376	0.877	0.115
	1000	none	0.013	0.015	0.959	0.013	0.015	0.925	0.034	0.038	0.917	0.559
	1000	light	0.014	0.016	0.955	0.014	0.017	0.908	0.036	0.043	0.924	0.559
	1000	heavy	0.017	0.018	0.977	0.018	0.021	0.910	0.044	0.078	0.937	0.559
Std-Gamma	200	none	0.050	0.138	0.997	0.054	0.143	0.997	0.146	0.289	0.904	0.115
	200	light	0.051	0.134	1.000	0.057	0.148	0.996	0.165	0.308	0.919	0.115
	200	heavy	0.058	0.145	0.999	0.070	0.165	0.990	0.363	0.334	0.892	0.115
	1000	none	0.013	0.014	0.947	0.013	0.015	0.918	0.034	0.038	0.911	0.559
	1000	light	0.014	0.016	0.952	0.014	0.016	0.906	0.036	0.044	0.915	0.559
	1000	heavy	0.016	0.018	0.977	0.018	0.022	0.912	0.043	0.074	0.943	0.559
T-mixture	200	none	0.040	0.107	0.999	0.048	0.117	0.990	0.123	0.256	0.909	0.115
	200	light	0.045	0.122	0.999	0.052	0.135	0.991	0.133	0.275	0.909	0.115
	200	heavy	0.050	0.122	1.000	0.060	0.149	0.989	0.186	0.311	0.895	0.115
	1000	none	0.025	0.038	0.977	0.025	0.039	0.973	0.044	0.072	0.934	0.559
	1000	light	0.026	0.033	0.982	0.027	0.033	0.980	0.047	0.079	0.937	0.559
	1000	heavy	0.031	0.046	0.979	0.032	0.048	0.974	0.059	0.113	0.938	0.559

Table S2: Simulation results for AFT models with randomly generated regression functions. Root mean-squared error (RMSE), mis-classification proportion (MCprop), and empirical coverage are shown for each of the methods. Performance measures are shown for the non-parametric tree-based AFT (NP-AFTree) method, the semi-parametric tree-based AFT (SP-AFTree), and the parametric, linear regression - based AFT (Param-AFT) approach.

			NP-AFTree			SP-AFTree			Param-AFT			Naive
Distribution	n	Censoring	RMSE	MCprop	Cover	RMSE	MCprop	Cover	RMSE	MCprop	Cover	MCProp
Normal	200	none	0.170	0.163	0.902	0.169	0.157	0.902	0.429	0.234	0.884	0.101
	200	light	0.196	0.142	0.868	0.195	0.141	0.867	0.439	0.244	0.862	0.151
	200	heavy	0.213	0.129	0.883	0.212	0.126	0.882	0.583	0.240	0.849	0.108
	1000	none	0.116	0.084	0.872	0.115	0.089	0.877	0.221	0.171	0.825	0.126
	1000	light	0.114	0.094	0.876	0.114	0.096	0.882	0.237	0.171	0.793	0.116
	1000	heavy	0.146	0.096	0.863	0.146	0.097	0.862	0.283	0.194	0.815	0.154
Gumbel	200	none	0.199	0.130	0.885	0.198	0.130	0.884	0.408	0.280	0.890	0.155
	200	light	0.188	0.164	0.889	0.187	0.160	0.888	0.759	0.224	0.857	0.122
	200	heavy	0.210	0.183	0.865	0.206	0.169	0.855	2.509	0.269	0.788	0.124
	1000	none	0.110	0.060	0.886	0.113	0.061	0.881	0.228	0.140	0.791	0.114
	1000	light	0.127	0.066	0.877	0.126	0.069	0.874	0.235	0.198	0.797	0.149
	1000	heavy	0.176	0.109	0.830	0.177	0.115	0.830	0.289	0.183	0.795	0.118
Std-Gamma	200	none	0.182	0.112	0.871	0.181	0.114	0.873	0.386	0.235	0.886	0.091
	200	light	0.207	0.145	0.873	0.206	0.151	0.867	0.443	0.276	0.863	0.134
	200	heavy	0.200	0.145	0.888	0.198	0.145	0.886	2.848	0.281	0.869	0.160
	1000	none	0.013	0.014	0.947	0.013	0.015	0.918	0.034	0.038	0.911	0.559
	1000	light	0.014	0.016	0.952	0.014	0.016	0.906	0.036	0.044	0.915	0.559
	1000	heavy	0.016	0.018	0.977	0.018	0.022	0.912	0.043	0.074	0.943	0.559
T-mixture	200	none	0.195	0.116	0.882	0.195	0.117	0.884	0.395	0.285	0.887	0.211
	200	light	0.189	0.111	0.886	0.188	0.116	0.886	0.419	0.210	0.855	0.112
	200	heavy	0.214	0.144	0.866	0.212	0.150	0.857	0.807	0.267	0.831	0.138
	1000	none	0.025	0.038	0.977	0.025	0.039	0.973	0.044	0.072	0.934	0.559
	1000	light	0.026	0.033	0.982	0.027	0.033	0.980	0.047	0.079	0.937	0.559
	1000	heavy	0.031	0.046	0.979	0.032	0.048	0.974	0.059	0.113	0.938	0.559

Table S3: Simulation results for AFT models with randomly generated regression functions. Comparison of two different approaches for treatment inclusion. Root mean-squared error (RMSE), mis-classification proportion (MCprop), and empirical coverage are shown for each of the methods. Performance measures are shown for two approaches: (1) the original non-parametric tree-based AFT (NP-AFTree) method and (2) the non-parametric tree-based AFT model applied to the two treatment arms separately (NP-AFTree Split Sample). With the second approach, we split the sample by treatment group, apply BART to each, then use the two models to generate an ITE estimate (and uncertainty estimate) for each patient.

			NP-AFTree			NP-AFTree (Split Sample)		
Distribution	n	Censoring	RMSE	MCprop	Cover	RMSE	MCprop	Cover
Normal	200	none	0.185	0.127	0.863	0.246	0.158	0.998
	200	light	0.180	0.116	0.895	0.244	0.203	0.994
	200	heavy	0.203	0.159	0.860	0.276	0.258	0.993
	1000	none	0.116	0.072	0.868	0.223	0.156	0.983
	1000	light	0.144	0.071	0.862	0.233	0.140	0.973
	1000	heavy	0.157	0.062	0.844	0.284	0.172	0.954
Gumbel	200	none	0.195	0.098	0.856	0.255	0.130	0.996
	200	light	0.191	0.126	0.869	0.257	0.186	0.987
	200	heavy	0.212	0.123	0.866	0.329	0.245	0.937
	1000	none	0.130	0.066	0.850	0.248	0.134	0.980
	1000	light	0.140	0.080	0.867	0.260	0.165	0.963
	1000	heavy	0.151	0.104	0.834	0.281	0.250	0.944