8-8-2008

Estimation for Arbitrary Functionals of Survival

Kyle Rudser
University of Minnesota, rudser@umn.edu

Michael L. LeBlanc
Fred Hutchinson Cancer Research Center, mleblanc@fhcrc.org

Scott S. Emerson
University of Washington, semerson@u.washington.edu

Estimation for Arbitrary Functionals of Survival

Kyle D. Rudser[†], Michael L. LeBlanc[‡], Scott S. Emerson[‡]

[†]Division of Biostatistics, University of Minnesota,
717 Delaware St. SE, Room 219, Minneapolis, Minnesota, 55414, U.S.A.
[‡]Department of Biostatistics, Box 357232, University of Washington,
Seattle, Washington, 98195, U.S.A.

Abstract

The hazard ratio is commonly used for comparing survival distributions across groups. While easily estimated in the presence of censored data, by itself it does not allow physicians, patients, and regulatory agencies to easily judge the clinical relevance of any difference in survival across groups. We consider an estimation approach for clinically meaningful functionals of a survivor distribution (e.g., restricted mean, quantiles). In this approach we use different regression models to borrow information across sparse data than to form statistical contrasts of an estimated functional of interest.

In the context of three data models, each analyzed by three different statistical models, we examine the ability to form accurate estimates. Specifically, we compare a nonparametric predictive model based on recursive partitioning of a multivariate predictor space to the two semi-parametric approaches of Cox's proportional hazards and Buckley-James' linear regression with censored data. First order linear contrasts across three estimated functionals of interest from each predictive model approach are compared using root mean squared error. In the examples covered we demonstrate that an adaptive nonparametric predictive model could prove markedly superior to the use of semi-parametric predictive models.

Keywords: Nonparametric, Semi-parametric, Variable importance, Predictive model, Survival analysis, Distribution free

1 INTRODUCTION

In analysis of censored time to event data, the primary interest is often to detect differences in survival across multiple groups. To this end, a parameter θ is usually identified as a function of covariates for adjustment. When using common parametric statistical models, θ will typically correspond to some summary measure, e.g., the mean, geometric mean, median, or hazard. In the

absence of censoring, maximum likelihood estimates of θ will often have a corresponding nonparametric interpretation. For example, linear regression estimates based on normal theory correspond to the sample mean, and linear regression with log transformed data (as for a lognormal model) corresponds to the sample geometric mean.

But with censored data, the efficient score function generally will depend more heavily on the shape of the survivor function, and maximum likelihood estimates will not have a straightforward nonparametric interpretation. Perhaps because of this sub-optimal robustness to distributional assumptions, parametric methods are not as widely used for censored data. Broader adherence to distributional assumptions was achieved with the introduction of semi-parametric models. For example, the methods of Buckley and James (Buckley and James 1979) applied to log transformed survival times can be viewed as an extension of linear regression models to the censored data setting in such a way as to maintain a correspondence between parameter estimates and the geometric means and/or quantiles of the distributions. Similarly, Cox's semi-parametric proportional hazards regression model (Cox 1972) can be viewed as an extension of the exponential or Weibull parametric regression models in such a way as to maintain a correspondence between the parameter estimates and the hazard ratio.

However, while semi-parametric models are more flexible than parametric ones, violations of corresponding semi-parametric assumptions may still lead to poor performance. As one considers different statistical models to use for an analysis and how the restrictions imposed by those models may affect performance in different scenarios, one goal may be to avoid approaches that have strong assumptions on the relationships and structure in the data. Indeed, when detecting differences in survival is the goal, there is something to be gained from being able to distinguish different survival curves with respect to one summary measure when other summary measures of those curves are not different (e.g., when some average hazard ratio estimated between two curves is 1.0, but the mean or median survival time is different). In this manuscript we explore the degree to which sensitivity to distributional assumptions may lead to poor performance of semi-parametric statistical models and include in comparison nonparametric methods where $\hat{\theta}$ is a nonparametric estimate of a chosen

θ.

In addition to avoiding strong and restrictive parametric and semi-parametric assumptions, a second motivation for the approach presented here is to facilitate the use of more meaningful summary measures than are typically used to date. While the ease of estimating the hazard ratio in the presence of censored data makes it attractive as a summary measure, it lacks an interpretation useful to clinicians, patients, and regulatory agencies. More meaningful functionals of a survivor distribution include various quantiles (e.g., median, 75^{th} percentile), or restricted means (e.g., 6 months, 3 years). As stated by Amna Ibrahim, M.D. of the Food and Drug Administration Center for Drug Evaluation and Research at an Onocologic Drugs Advisory Committee meeting on September 13, 2004:

> Hazard ratios give only an incomplete picture. Hazard ratios may represent statistical significance, however, clinical relevance as the benefit provided to the patient is not captured. For example, hazard ratios will treat the improvement from three days to six days the same as improvement from three years to six years.

Without censored observations, inference on more useful summary measures would be straightforward, but survival data commonly includes censoring.

2 BACKGROUND

When comparing differences in survival across multiple groups is the goal, we may have different functionals of interest depending on the scientific context, or if we are interested only in demonstrating qualitative differences in survival, we might be comfortable with any of a variety of choices for θ. In the case of the latter, the choice of statistical model will depend on the degree to which we are willing to assume that the underlying structure of the data follows parametric or semi-parametric relationships. If we would like to avoid presuming a particular distributional structure, a nonparametric approach may be chosen. For example, if prior knowledge suggests a constant hazard function, an exponential regression model may be selected, and the differences between curves described by differences in the mean. Alternatively, the broader semiparametric

proportional hazards model (which includes the exponential distribution as a special case) could be used to provide inference on constant relationships between time varying hazards. A nonparametric approach modeling the median could also be considered if imposing parametric or semi-parametric structure is undesirable. As we posited no *a priori* preference in this hypothetical setting for the mean, hazard, or median, the three approaches might be equally viable, and the choice of statistical model would likely be driven entirely by one's belief in the underlying distribution of the data. Furthermore, in the case of common parametric and semiparametric models, implicit stochastic ordering of distributions across covariate groups argues that similar qualitative results are obtained for several candidate summary measures, thereby allowing the focus to be on the canonical summary measure that is most "natural" for the probability model.

Other times there is interest in estimating a quantitative difference in survival as measured by a specific summary measure. Justification for a particular choice of summary measure (perhaps formalized via a statistical loss function) may pertain to its scientific importance and ease of interpretation, to the likelihood that it would be affected by the covariates of interest, to the statistical precision with which it is estimated, or to the ease of estimation. For instance, the proportion of patients exceeding some threshold might be selected because it is clinically most relevant for a particular disease. Alternatively, if it were equally important to detect all departures from equality, the mean might be selected due to its sensitivity to a greater variety of ways that survival curves might differ. And despite its more difficult interpretation, the hazard ratio might be selected due to the ease with which the hazard is estimated in the presence of censoring. Inference on the selected summary measure can then proceed according to the available or presumed knowledge about the true distribution of the data. If one knows the underlying distribution of the data, the corresponding parametric statistical model may be used and the functional of interest evaluated afterwards. For instance, if the scientific interest is in the median, but the data is known to be lognormally distributed, an efficient analysis would be a parametric lognormal regression model of the log geometric mean, with the parametric estimator for the median derived from the knowledge that the median for lognormal data is also the geometric mean. Otherwise an approach can be

chosen that corresponds to the functional of interest with a degree of robust performance over a range of possible unknown underlying structure (e.g., using the sample median when primary interest is in the population median).

Performance of the statistical model used will depend on adherence to the distributional assumptions (e.g., parametric or semi-parametric assumptions), as well as any associated restrictions of the approach (e.g., trees are generally implemented as step functions while most other regression models consider effects that are on some scale linear in the modeled covariates). Models presuming particular relationships will force those relationships on the estimated survival curves, whether or not they hold for the real data. This means that, for instance, the true survival curves cannot be estimated consistently when using the proportional hazards model to analyze data exhibiting time-varying hazard ratios. Furthermore, the estimates obtained will be influenced by different patterns of censoring. Similarly, for the Buckley-James approach with log-transformed survival times, it may no longer be looking at the true geometric means if the true probability model does not have the accelerated failure times property. Additionally, if the probability and statistical model do not coincide, there is no guarantee for the results to indicate the desired ordering of survival curves.

We propose an approach to nonparametric inference for clinically meaningful functionals of a survivor distribution (e.g., the restricted mean, quantiles) amenable to avoiding strong parametric or semi-parametric assumptions on the true underlying structure. In this approach, two different models are used: First a predictive model is used to borrow information across sparse data to estimate survival curves for each combination of covariates. Then a second model is used to form contrasts across these groups based on chosen functionals of interest. This general approach is examined for three different predictive models each analyzed on three different data scenarios. The first is nonparametric recursive partitioning (Breiman, Friedman, Olshen, and Stone 1984; Nobel and Olshen 1996) of a multivariate predictor space to derive groups based on differences in their survival distributions. These differences are evaluated using a variety of statistics from the $G^{\rho,\gamma}$ family (Fleming and Harrington 1991), including the Wilcoxon and logrank statistics described in section 2.2. Estimates of the survival distribution within leaves of the tree are then used to

compute functionals, such as the restricted mean, and the 50^{th} and 75^{th} percentile of survival. The second predictive model considered is a semi-parametric approach using Cox proportional hazards models, and represents the current most common form of survival analysis. The third model included in the illustration is a second semi-parametric approach using Buckley-James' linear regression for censored data. In this approach, censored observations are replaced by their conditional mean based on a Kaplan-Meier estimate of the ordered residuals. Since the estimate is not finite whenever the largest residual is censored, the convention is to re-assign it as an event. For both semi-parametric approaches, estimates of the three summary measures of interest are computed from the corresponding adjusted subject-specific survival curves. Then first order linear contrasts are formed to evaluate associations of survival with covariates of interest across groups. The three approaches are compared on root mean squared error (RMSE) of the estimated contrasts. In the examples covered, we demonstrate that an adaptive nonparametric predictive model could prove markedly superior to the use of semi-parametric predictive models.

2.1 Notation

Let T be a continuous random variable, which denotes time to an event of interest. In analysis of survival data, it is often the case that subjects are censored. That is, let the random variable C denote the time to being censored, and $Y = min(T, C)$. Let δ be an indicator if the observed time is an event, i.e., $\delta = 1$ for $Y = T$ and $\delta = 0$ otherwise. Hence a sample of survival data of size n consists of the pairs $\{y_i, \delta_i\}$ for $i = 1, \ldots, n$. For the purposes of this paper, we will focus on scenarios where the censoring mechanism does not depend on the survival time. A survivor function will be denoted as $S(t) = 1 - F(t)$. Other related quantities commonly used in analyses are the hazard, $\lambda(t)$, and cumulative hazard, $\Lambda(t)$, function. It should be noted that knowing any one of $S(t),\ F(t),\ f(t),\ \lambda(t)$, or $\Lambda(t)$ is sufficient. For example, the survivor, hazard, and cumulative hazard functions can all be obtained from a known CDF. Hence they are all defined for *any* random variable with a CDF whether that be one for time to an event or not, e.g., the survivor, hazard and cumulative hazard functions are all defined for a random variable measuring blood pressure. In this sense, the approaches examined here are not restricted to survival analyses

and may be used in the general setting.

2.2 Two Sample Tests

In addition to estimating survival curves between groups, it is often of interest to formally test for differences in survival. One approach is the nonparametric $G^{\rho,\gamma}$ family of weighted rank tests by Fleming and Harrington (Fleming and Harrington 1991). These statistics take the following form:

$$G^{\rho,\gamma} = K^{1/2} \sum_{t \in \mathcal{F}} w(t)[\hat{\lambda}_1(t) - \hat{\lambda}_0(t)], \tag{1}$$

where $K = (M_1 + M_0)/(M_1 M_0)$ with M_i denoting the number of subjects in group i initially at risk, $\mathcal{F}$ denotes the set of unique failure times, $w(t) = [(n_{1t} n_{0t})/(n_{1t} + n_{0t})]\hat{S}(t-)^{\rho}[1 - \hat{S}(t-)]^{\gamma}$, with $\hat{S}(t-)$ denoting the pooled KM estimate of the survival curve (both groups 1 and 0) just prior to time t, and $\hat{\lambda}_i$ denoting the estimated hazard at time t for group i. The $G^{\rho,\gamma}$ statistics can be converted to a Z score by dividing by a consistent estimate of the variance, σ^2, under the null hypothesis $H_0 : S_1(t) = S_0(t)$:

$$\sigma^2 = K \sum_{t \in \mathcal{F}} w(t)^2 \left(\frac{1}{n_{1t}} + \frac{1}{n_{0t}}\right) \left(1 - \frac{(d_{1t} + d_{0t} - 1)}{(n_{1t} + n_{0t} - 1)}\right) \left(\frac{(d_{1t} + d_{0t})}{(n_{1t} + n_{0t})}\right), \tag{2}$$

where d_{1t} and d_{0t} are the number of events at time t for each of group 1 and 2 respectively. This results in a consistent, asymptotically normal statistic to evaluate against a standard normal distribution.

This family of statistics includes the logrank ($(\rho, \gamma) = (0, 0)$) and the Wilcoxon weighted form of the logrank ($(\rho, \gamma) = (1, 0)$). The most common nonparametric two sample test of differences in survival is the logrank test (score statistic in a Cox proportional hazards model). Note that since the logrank statistic is a member of the $G^{\rho,\gamma}$ family, it corresponds to a particular weighting of survival curves. If other weightings were used (e.g., ones that accentuate more heavily early or late differences in survival), these will have more power for other alternatives. In fact, under non-proportional hazards, the logrank test will not be the most efficient. As such, other tests in the

$G^{\rho,\gamma}$ family, which weigh more heavily early or late differences in survival, may be more powerful. Hence, better performance for detecting any differences in survival could be realized if we used a combination of these test statistics, such as a maximum of four of them.

2.3 Overview

The remainder of this paper is organized as follows. Section 3 details the paradigm of separating borrowing information from forming contrasts of interest. This may be viewed as a "two stage" procedure where a predictive model of choice is used to first estimate conditional cumulative distribution functions. In the second stage, a second model is used to define contrasts of functionals that are of scientific importance for a given context. We appeal to linear regression to illustrate how these two concepts occur regularly in standard statistical techniques.

Section 4 describes the setting we use to examine and contrast the performance of three different predictive models when making inference on the mean, median, and 75^{th} percentile. The root mean squared error of an estimated linear contrast is compared across predictive models for three distinct survival settings. In two of the settings, the survival probability models are chosen such that the proportional hazards assumption is violated, while the third satisfies proportional hazards. The first two models are further chosen to examine the behavior of the three predictive models when pairwise hazard ratios are nearly 1 and when pairwise differences in the median survival time are 0. Evaluations under 0%, 20% and 60% censoring are included.

Section 5 presents a summary of our results and offers some discussion of other contexts where the general approach could prove useful.

3 BORROWING INFORMATION AND FORMING CONTRASTS

Typically a dataset will not have information for every possible covariate combination. For this situation of sparse data we typically "borrow information" from the other available sample units to obtain a better estimate and greater understanding of the scientific context under consideration.

While there is no formal definition of borrowing information, the following will serve as the working definition for this paper. Suppose we have a sample of n subjects indexed by $i = 1, \ldots, n$. Information is borrowed for subject i from other sample units $j \neq i$ if the estimate for subject i depends on observations from subjects $j \neq i$, but subjects j do not directly estimate the same quantity as that of subject i. For example, in the setting of linear regression, we borrow information by presuming a linear trend in the means across groups defined by some modeled covariates. The familiar formula to borrow information in linear regression takes the following form:

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} \tag{3}$$

This same formula constitutes a contrast. A contrast, β, across groups $j = 1, \ldots, J$ is defined as follows:

$$\beta = \sum_{j=1}^{J} a_j \mu_j \qquad \text{where} \qquad \sum_{j=1}^{J} a_j = 0 \tag{4}$$

and where μ_j denotes a parameter for group j (e.g., the mean). Estimating the contrast is straight forward providing there is an estimate for μ_j: $\hat{\beta} = \sum_{j=1}^{J} a_j \hat{\mu}_j$. In the equation presented above, y_i can be thought of as being an individual specific estimate of μ_i:

$$\tilde{\beta} = \frac{\sum(x_i - \bar{x})\tilde{\mu}_i}{\sum(x_i - \bar{x})^2} \tag{5}$$

A similar approach is used in ANOVA, where the mean within each group serves as the estimate in forming a contrast. For example, for groups $1, \ldots, J$, the estimate for μ_j is $\bar{y}_{j\cdot} = \frac{\sum y_{ji}}{n_j}$. In the equations above, each individual represents their own group of size one. If a different parameter, θ_i is of interest (e.g., restricted mean or quantile of survival), a contrast can still be formed providing there is an estimate for it.

3.1 The Separation of Borrowing Information and Forming Contrasts

In linear regression, borrowing information and forming contrasts occur simultaneously with the equations presented above. This need not be the case. One method may be used to first borrow information across sparse data and then another approach used to form contrasts across groups. With time to event data, it is often the case to have censored observations where the event of interest is not observed, but it is known not to have occurred up to a certain point in time. In this setting, the survival curve for a given individual will typically depend on his/her covariates, $S_i(t) = S(t|\vec{x}_i)$. Estimation of the survival curve, $\hat{S}(t|\vec{x_i}) = g(t, \vec{x})$, can be done parametrically (e.g., exponential or Weibull), semi-parametrically (e.g., proportional hazards or accelerated failure times with estimated baseline survival), or nonparametrically (e.g., Kaplan-Meier (KM) curves for each group). Then an estimate of an arbitrary functional θ_i can be obtained providing there is an estimate of the survival curve, $\tilde{\theta}_i = h(\hat{S}(t|\vec{x_i}))$. For example, the p-th quantile can be estimated directly from an estimate of the survival curve, $\tilde{\theta}_i = \hat{S}^{-1}(p)$, and the mean (or restricted mean up to time τ) can be estimated by the area under $\hat{S}(t)$, $\tilde{\theta}_i = \int_0^\tau \hat{S}(t)dt$.

3.2 Choices of Predictive Models

The first stage of the proposed approach is to use a predictive model of choice to estimate functionals of interest for each individual. This is accomplished via functionals of estimated conditional distribution functions (survival curves). We do this by ascribing to each individual the survival curve estimated for the group having the same covariates as that individual. Examples of predictive models include:

1. Parametric models: After fitting a regression curve with the canonical parameter, conditional survival curves are then estimated along with functionals of interest.

2. Cox proportional hazards model: After fitting a regression curve with modeled covariates, the baseline survival curve $S_0(t)$ is estimated using the Breslow cumulative hazard estimator. The estimated survival curve for an individual having covariates x is then $S_0(t)^{e^{xb}}$.

3. Buckley-James model: After fitting a regression curve with modeled covariates, the estimated

survival curve for an individual having covariates x is then $S_0(te^{xb})$.

4. Regression tree: A survival regression tree is grown as described by Leblanc and Crowley (LeBlanc and Crowley 1993), and the survival curve estimate for each individual is taken from the KM estimate for the leaf containing that individual's covariates.

5. Multivariate kernel smoothing: A KM estimate incorporates weighted observations according to the 'distance' of each observation from a particular individual.

6. Bagged or boosted estimates for any of the above.

3.3 Linear contrasts

Any standard regression model can be used to define a linear contrast of the functionals computed from the individual survival curves. For greatest utility, the covariates included in the linear contrast must be a subset of those used in the predictive model. One straightforward approach is to use a linear regression model, although any weighted or unweighted GLM or GAM of choice may be implemented. Inference for corresponding contrasts would need to account for the data driven aspect of the predictive model; Bootstrapping the entire two stage procedure would be one possible approach. This manuscript will restrict focus to accuracy of estimation, however, and thus does not investigate the inferential procedures.

4 SIMULATIONS

In order to investigate potential improvements in accuracy, as well as potential loss of efficiency across a variety of choices of predictive models for the two stage approach, we compare a nonparametric tree approach to two analogous semi-parametric approaches: one based on the widely used Cox proportional hazards model (Cox 1972) and a second based on linear regression for censored data by Buckley and James as a representative of an accelerated failure time model (Buckley and James 1979). To highlight situations where the methods might behave differently, we constructed three scenarios in which survival curves differ across rectangular regions of a predictive space: one where pairwise hazard ratios (as estimated by a proportional hazards model in the absence of cen-

soring) are all near 1.0 (referred to as null hazard from now on), one where the median survival is equal across all curves (referred to as null median from now on), and one under proportional hazards. For the first two scenarios, the survival curves were specifically generated to not come from any standard parametric or semi-parametric distribution.

4.1 Predictive Models

Tree formation is a natural nonparametric partitioning algorithm based on splitting criteria. Briefly, at each partitioning step, all possible dichotomous splits are examined in deciding where to split the data. This continues until stopping criteria are met, or there are no more possible splits. After a tree is grown, it is pruned. This involves examining the strength of splits and possibly removing some of them. Many splitting, pruning and tree growing methods have been developed and suggested (Breiman, Friedman, Olshen, and Stone 1984; LeBlanc and Crowley 1992; LeBlanc and Crowley 1993; Alexander and Grimshaw 1996; Breiman 2001; Molinaro, Dudoit, and van der Laan 2004; Hothorn, Hornik, and Zeileis 2006).

For this illustration, the data were split based on comparisons of KM curves in similar fashion to that of LeBlanc and Crowley (LeBlanc and Crowley 1993), but instead using a cocktail of four different statistics from the $G^{\rho\gamma}$ family: (ρ, γ) equaling $(0, 0)$ (the logrank statistic), $(1, 0)$ (Wilcoxon form of the logrank statistic), $(0, 1)$, and $(1, 1)$, with the goal of forming groups that are homogeneous with respect to survival. The maximum of the four different statistics was used to indicate differences in survival as done similarly by Lee (Lee 1996).

Upon termination of the splitting procedure, the tree is pruned according to the Z-statistics for each split. Once groups have been identified through this data enhancement procedure, a KM curve is estimated for each group, which can be regarded as homogeneous or kernel smoothing across individuals in some neighborhood. Estimated functionals, $\tilde{\theta}_i$, from the KM curve for each group defined by the partitioning represent an estimate for each individual in the group. Regression models of these estimates can then be used to define contrasts across covariates of interest.

The Cox and Buckley-James approaches fit stepwise models to capture the flavor of the adaptive tree approach. Inclusion of log transformations and interactions was designed to also capture the

flavor of the tree approach since trees are invariant to monotonic transformations and fit interactions immediately. For each set of 2,000 simulations, data were analyzed by the nonparametric tree based approach and both semi-parametric approaches. In an effort to make comparisons between the three on a 'fair' playing field, each approach was calibrated to 10% experimentwise type I error when the entire sample was drawn from a single distribution (i.e., under a strong null hypothesis). Survival estimates obtained for each individual were used to compute three individual specific functionals for comparison: mean, median, and 75^{th} percentile.

4.2 Underlying Structure

In order to use a realistic predictor distribution, the underlying structure for the simulations used the covariate structure from a frequently used dataset of a clinical trial on primary biliary cirrhosis among 416 patients (Fleming and Harrington 1991). Two moderately correlated prognostic variables were used: serum bilirubin and prothrombin time (time to blood coagulation). The covariate space was partitioned into nine regions based on a survival tree grown on the original data, but the actual survival time data was not otherwise used. For the null hazard and null median scenarios, nine distinct groups based on these two covariates were instead assigned distinct survival curves. Curves were generated piecewise from a power family, with one curve representing uniform survival. Under the null hazard scenario, the hazard ratio with respect to uniform survival was 1.0 for each of the groups, although each one had a different value for its mean, median, and 75^{th} percentile of survival (Figure 1 and 2). Since the hazard ratio (HR) is non-transitive (Gillen and Emerson 2007), i.e., for groups Y, X, and Z, $HR_{YX} = 1.0$ & $HR_{XZ} = 1.0 \not\Rightarrow HR_{YZ} = 1.0$, it is impossible to have a hazard ratio of 1.0 for each pairwise comparison between nine distinct curves. Hazard ratios between non-uniform survival curves (all but curve #5) ranged between 0.83 and 1.12. This results in a setting where survival times are simulated from curves where the truth is *nearly* without signal in hazard ratios. By examining the nine curves in Figure 1, one can observe the difficulty in interpreting the hazard ratio, the non-transitivity, and how failure to distinguish between a weak null hypothesis (i.e., the hypothesis that the time averaged hazard ratio is 1) and

strong null hypothesis (i.e., the hypothesis that two survival curves are exactly the same) can lead to stark differences in statistical inference. In each of the nine plots, the survival curve is displayed with a solid line, which is superimposed over a dotted line representing uniform survival. As such, in each subplot the solid and dotted survival curves have a hazard ratio of 1.0 (by design). Clearly, in each plot (save #5 where the solid and dotted curves are the same) the two survival curves displayed are quite different. They have different values of mean, median and 75^{th} percentile as was seen in Figure 2. This probability model for the data might be expected to illustrate advantages for predictive models that do not assume proportional hazards.

For the null median scenario, all of the curves were designed to have an identical median survival of 0.5. However, there are differences between curves for the other functionals, including hazard ratios (Figure 3). The context of this simulation does not inherently favor one approach over the other *a priori*.

For the proportional hazards scenario, curves were generated by a Cox proportional hazards model with prothrombin time and the log of bilirubin modeled continuously. In this manner, a separate survival curve was used to simulate failure times for each unique combination of bilirubin and prothrombin time in the dataset, of which there are 330. A sample of these curves is shown in Figure 4, where the proportional hazards property can be seen to dictate the relationship between adjusted survival curves, resulting in curves that are stretched vertically from the baseline survival curve. Features of these curves are displayed in Figure 5.

The two stepwise semi-parametric approaches were given seven possible covariates to be added into or removed from the model: bilirubin, prothrombin time (protime), log(bilirubin), log(protime), and interactions bilirubin x protime, log(bilirubin) x protime, and bilirubin x log(protime). After fitting predictive models, contrasts of the estimated functionals were evaluated via an additive first order linear regression model of linear continuous terms for bilirubin and prothrombin time: $E(\tilde{\theta}_j|bili_j,\ protime_j) = \beta_0 + \beta_1\ bili_j + \beta_2\ protime_j$. Accuracy and precision of the approaches were evaluated using the root mean squared error (RMSE) of estimated contrasts $\tilde{\beta}_1$ and $\tilde{\beta}_2$.

4.3 Results: Null Hazard

The context of this simulation does not favor a Cox model because it was specifically designed to have similar hazard ratios between all of the curves. It is not completely without signal in hazard ratios, however, which is reflected in a null stepwise model having been fit only 77.3% of the 2,000 simulations. This is in contrast to the 90% or more it would have had if all pairwise hazard ratios were 1.0: The stepwise model fitting was calibrated to 10% error under the strong null hypothesis.

It should be noted that without any censoring, the Buckley-James approach reduces precisely to linear regression with log-transformed survival times. Hence, that model was able to detect differences in survival probabilities more often, and resulted in null models less frequently.

The tree based approach also regularly avoided null-fits (no splits). Trees with 2, 3, and 4 groups were identified 26.2%, 29.8%, and 23.1% of the time, respectively, and only rarely identified 9 distinct groups, the true number of underlying distinct survival curves. Figure 6 shows a graphical representation of tree predictions using CARTscans (Nason, Emerson, and LeBlanc 2004) to illustrate the underlying structure. Plot(a) depicts the nine true underlying groups, where each distinct color represents one of the nine different groups with shading to indicate the magnitude of the true mean. In addition, a sample of 8 tree fits are presented with shading to represent estimated means for each group. None of the 8 tree fits shown here result in 9 groups being identified (groups range from 3 in plot(b) to 7 in plot(i)), although there is a distinctive major separation in the true structure (plot(a)) in log bilirubin just below 1, which appears in all of the eight sampled simulated times.

Under the null hazard scenario, the tree approach (Tree) shows magnitudes of improvement in RMSE compared to the proportional hazards approach (Cox) and accelerated failure times approach (Buckley-James) for all three functionals (Table 1). When 20% censoring is included, the results are similar with the tree procedure having the best RMSE between the three approaches. Compared to the case of no censoring, the Tree estimates are not noticeably different, as expected since the underlying approach using KM curves is point-wise consistent. This is not the case for the Cox approach where the hazard ratio is not consistently estimated under varying levels or patterns of

censoring in the setting of non-proportional hazards. This aspect is reflected in there being only 57.2% null stepwise fits as opposed to 77.3% without censoring. As a consequence, the Cox approach will estimate differences across bilirubin and protime more often (non-null fits) and perform better than it had without censoring. The Buckley-James approach performed much worse with estimates that converged only 83.5% of the time. This undesirable feature of the Buckley-James procedure has been documented and noted in the original manuscript and others (Buckley and James 1979; Miller and Halpern 1982). In the absence of censoring, the Buckley-James approach can be viewed as nonparametric estimation of geometric means. In the presence of censoring, the Buckley-James approach uses a strong semi-parametric assumption, and in this data (and likely in general) that assumption is invalid, resulting in a failure to converge relatively often in finite samples.

4.4 Results: Null Median

RMSE results for the protime contrast were similar to those observed under the null hazard scenario with the tree approach performing as well as or much better than the other two across all three functionals (Table 2). For the bilirubin contrast, the Cox approach did well for the median but not as well for the other two functionals when compared to the two other approaches.

With the addition of 20% censoring, the tree approach held steady in estimated RMSE compared to the two semi-parametric approaches. Compared to no censoring, the tree approach again did not perform noticeably worse. This again is likely due to the fact that the tree approach is based on KM estimates, which are point-wise consistent for the survival curve, and hence will be consistent for the estimated functionals of the mean, median and 75^{th} percentile. It will thus, in turn, also be consistent for the contrasts of bilirubin and prothrombin time. The Cox approach had some results with censoring that were better and some that were worse compared to no censoring. This arises from a combination of having fit a null model more often (28.9% and 6.9% for with and without censoring respectively) and not being consistent in the presence of censoring without proportional hazards. The Buckley-James approach, as with the previous scenario, has trouble converging again with only 72.0% convergence and worse RMSE for all functionals in the presence of censoring.

4.5 Results: Proportional Hazards

In the third scenario of proportional hazards (Figure 4 and 5), the Cox model was expected to do the best, perhaps by far, since the probability model used to generate the data satisfies assumptions of proportional hazards. As displayed in Table 3, the approach using a Cox model does in fact perform the best across all three functionals, although the nonparametric tree approach did surprisingly well with only moderate loss of precision when compared to the Cox approach. The magnitude between the Tree and Cox approaches is on the order of approximately 2 at most in favor of the Cox predictive model for bilirubin, and essentially indistinguishable for prothrombin time. For the previous two scenarios in which the Tree approach outperformed the Cox approach, the order of magnitude ranged from 1.7-6.2 and 1.2-2.0 for the null hazard and null median scenarios respectively.

With the addition of censoring, the tree again did not change noticeably. The Cox approach did not change much either, save the contrasts for the median functional, where the contrast for bilirubin did more poorly with censoring and the contrast for prothrombin time improved slightly. Although the Cox approach did not have any null stepwise fits under this scenario with censoring either, the distribution of covariates included in the Cox stepwise models across simulations shifted downward towards fewer covariates, reflecting expected lower power with the lower number of events. The Buckley-James approach appears to perform comparably, however, the RMSE presented for this approach does not include 28.9% of simulated sets with censoring where it failed to converge. As such, the performance of the Buckley-James approach is characterized as much worse than the other two predictive models where estimates are always obtained.

4.6 Heavier Censoring

Inclusion of censoring in the evaluation of performance between predictive models is important to consider due to the potential extreme influence it may have. For a given predictive model, the degree of influence will largely depend on assumptions being valid. A case in point is the Buckley-James approach which may not give a solution at all for a given context when in the presence of as little as 20% censoring. In real-life scenarios, it is not uncommon for heavier censoring to be

observed, which may exaggerate the discouraging results observed previously.

Under heavier censoring of 60%, the tree approach continued to dominate in estimated RMSE for the null hazard scenario. Compared to 20% censoring, the RMSE for the tree did not change noticeably. For the Cox approach, on the other hand, RMSE dropped for both contrasts across all three functionals. The explanation of this is the same as why improvement was seen going from 0% to 20% censoring, there are more non-null model fits – 98.2%. This might be expected due to the influence of the censoring pattern on the hazard ratios estimated by the Cox model: With 60% censoring, the estimated hazard ratios between pairs of curves ranged as high as 3.47. Again, while the hazard ratio has changed for every pairwise comparison, the underlying true mean, median and 75^{th} percentile functionals do not change at all.

For the null median scenario with 60% censoring, the Cox approach no longer has the best RMSE performance for any of the functionals. The tree approach now dominates all of the RMSE being compared, as was true in the null hazard scenario. Similar to the null hazard scenario with more censoring, the Buckley-James approach has an increased proportion of non-convergence under the null median scenario from 28.05% with 20% censoring to 58.55% with 60% censoring.

Lastly, examining the proportional hazards scenario with 60% censoring shows the RMSE comparison was similar to that with 20% censoring except that the tree approach now has better performance for the bilirubin contrast of the 75^{th} percentile. The RMSE results for the Buckley-James approach appear comparable; however, that does not include 81.1% of the time which the procedure was non-convergent.

5 SUMMARY

We explored one nonparametric and two semi-parametric approaches for time to event data examining clinically relevant summary measures based on time. While this was in an unrealistic setting, the nonparametric approach was found to have improved ability to find differences in functionals of interest, particularly when such differences did not occur within the context of a readily identified parametric or semi-parametric model.

Parametric models specify the entire shape of the distribution, including areas well beyond the support of the sample. One might imagine this to be analogous to extrapolation with polynomial regression models. Hence, it is not very surprising that many statisticians have found these models do not perform as well as they would like. Some semi-parametric models provide a greater level of flexibility as compared to parametric ones, but are still restrictive with their corresponding assumptions. For example, while the underlying or baseline survival distribution may be unrestricted, everything is presumed to be completely specified with an estimate of the baseline survival and a set of covariates. Also, semi-parametric methods in general are still restrictive in using the shape of the estimated survival curve. In essence, the semi-parametric estimating equations will use the baseline survival estimates to extrapolate the survival curves for groups with sparse data. When the true survival does not follow the associated semi-parametric assumptions, resulting estimates can be thrown off. This results in having the wrong size (type I error) under a weak null hypothesis. As we would like an approach to mimic the decision that would have been made under no censoring, this possibility is undesirable.

As noted, the tree approach considered here is also not without limitations. However, with the nonparametric nature of the approach it avoids strong parametric or semi-parametric distributional assumptions. It can adaptively identify the extent of heterogeneity of distributions over a range of covariate values, and it is robust to non-linearities and interactions of covariate effects. In as much as complex step functions are able to reasonably approximate the true underlying probability distribution, the approach ought to perform well.

A general theme that has emerged from the results here is the problem of underfitting. This was seen with the Cox semi-parametric model under the null hazard scenario where the inability to fit a model at all greatly inhibited performance. In this sense, improvements in estimated contrasts may be obtained by allowing more than the 10% "predictive model error" on the strong null that was specified by design for this manuscript.

The two stage approach examined here that separates the two ideas of borrowing information and forming contrasts can be used with parametric, semi-parametric, or nonparametric predictive

models. Ideally, an approach would borrow information across sampled units as much and as far as to maximize accuracy and precision. Parametric and semi-parametric approaches inherently borrow information across every individual according to the structure imposed by model assumptions. This is particularly noticeable in Figure 4 where all of the estimated individual step-function curves are not just allowed to change at each time an event was observed in a subset of the sample, but indeed forced to change at every observed event time. The nonparametric approach shown here is a flexible, data-adaptive procedure to borrow information only within each identified group in the tree, with potential to allow for borrowing information as far as is appropriate, and not any further. It also avoids strong semi-parametric assumptions, such as proportional hazards, and will replace it with a less restrictive nonparametric assumption of local smoothness within each group. Approaches incorporating flexibility to implement the full range between these two in how information is borrowed warrant further investigation in the contexts presented here; two examples of such approaches are kernel smoothing and bagged trees (Breiman 1996).

The results presented here show the RMSE performs well using a nonparametric, tree-based approach. Future investigations include comparisons in other contexts such as smoothly varying survival curves across underlying groups. Along with appropriate estimation of contrasts, we would also like to have a method of inference with accurately estimated standard errors. To this end, bootstrapping may be used to obtain reliable estimates of variability of the contrasts of interest. Issues arise due to the discrete nature of estimated survival curves. To ameliorate this, bagging may be used to decrease discreteness and bootstrapping residuals to decrease impact of influence at extremes of the predictor distribution.

References

Alexander, W. and S. Grimshaw (1996). Treed regression. *Journal of Computational and Graphical Statistics 5*, 156–175.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

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

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. Wadsworth and Brooks.

Buckley, J. and I. James (1979). Linear regression with censored data. *Biometrika 66*(3), 429–436.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological 34*, 187–220.

Fleming, T. R. and D. P. Harrington (1991). *Counting Processes and Survival Analysis*. Wiley.

Gillen, D. L. and S. S. Emerson (2007). Nontransitivity in a class of weighted logrank statistics under nonproportional hazards. *Statistics and Probability Letters 77*, 123–130.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics 15*(3), 651–674.

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

LeBlanc, M. and J. Crowley (1993). Survival trees by goodness of split. *Journal of the American Statistical Association 88*, 457–467.

Lee, J. W. (1996). Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics 52*(2), 721–725.

Miller, R. and J. Halpern (1982). Regression with censored data. *Biometrika 69*(3), 521–531.

Molinaro, A. M., S. Dudoit, and M. J. van der Laan (2004). Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis 90*, 154–177.

Nason, M., S. Emerson, and M. LeBlanc (2004). Cartscans: A tool for visualizing complex models. *Journal of Computational and Graphical Statistics 13*(4), 1–19.

Nobel, A. B. and R. A. Olshen (1996). Termination and continuity of greedy growing for tree-structured vector quantizers. *IEEE Transactions on Information Theory 42*(1), 191–205.

Table 1: Root mean squared error comparison on the mean, median, and 75^{th} percentile of survival between the three approaches under the null hazard scenario with and without 20% censoring. True values for the contrasts of the mean, median, and 75^{th} percentile are -0.149, -0.277, and -0.296 for bilirubin and 0.354, 0.593, and 0.795 for protime respectively.

		No Censoring		20% Censoring	
θ	Approach	Bilirubin	Protime	Bilirubin	Protime
Mean	Tree	0.052	0.191	0.045	0.194
	Cox	0.154	0.358	0.124	0.353
	Buckley-James*	0.180	0.466	0.190	0.484
Median	Tree	0.118	0.355	0.101	0.353
	Cox	0.280	0.595	0.242	0.585
	Buckley-James*	0.188	0.807	0.571	3.802
75%ile	Tree	0.049	0.381	0.051	0.394
	Cox	0.303	0.793	0.256	0.784
	Buckley-James*	0.065	0.216	0.266	1.583

*With 20% censoring model fit did not converge in 16.5% of simulations.

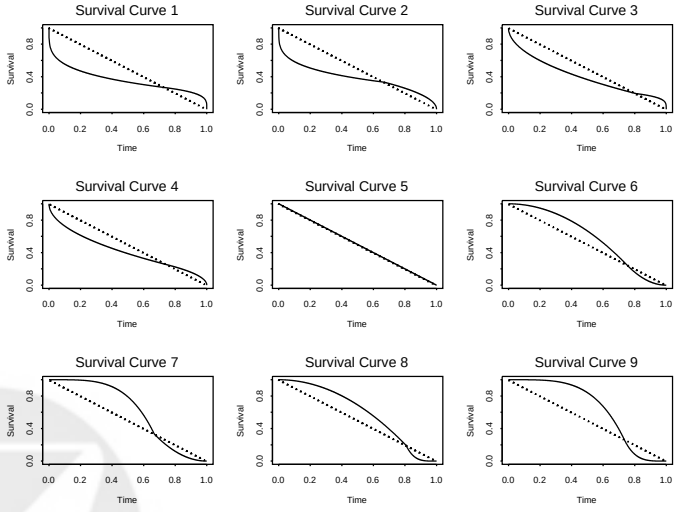


Figure 1: Survival curves used for a null hazard scenario. All curves have a hazard ratio of 1.0 with respect to the uniform survival curve (#5), which is presented as the dotted line in each plot.

Table 2: Root mean squared error comparison on the mean, median, and 75^{th} percentile of survival between the three approaches under the null median scenario with and without 20% censoring. True values for the contrasts of the mean, median, and 75^{th} percentile are -0.110, 0, and -0.254 for bilirubin and 0.358, 0, and 0.273 for protime respectively.

		No Censoring		20% Censoring	
θ	Approach	Bilirubin	Protime	Bilirubin	Protime
Mean	Tree	0.071	0.145	0.064	0.188
	Cox	0.096	0.181	0.093	0.238
	Buckley-James*	0.192	0.283	0.198	0.376
Median	Tree	0.118	0.317	0.124	0.300
	Cox	0.050	0.502	0.059	0.492
	Buckley-James*	0.379	0.698	0.761	2.656
75%ile	Tree	0.118	0.210	0.100	0.225
	Cox	0.231	0.250	0.229	0.218
	Buckley-James*	0.056	0.251	0.413	1.550

*With 20% censoring model fit did not converge in 28.05% of simulations.

Table 3: Root mean squared error comparison on the mean, median, and 75^{th} percentile of survival between the three approaches under the proportional hazards scenario with and without 20% censoring. True values for the contrasts of the mean, median, and 75^{th} percentile are -0.182, -0.173, and -0.111 for bilirubin and 0.117, 0.123, and 0.077 for protime respectively.

		No Censoring		20% Censoring	
θ	Approach	Bilirubin	Protime	Bilirubin	Protime
Mean	Tree	0.036	0.144	0.037	0.145
	Cox	0.019	0.141	0.021	0.153
	Buckley-James*	0.024	0.141	0.030	0.163
Median	Tree	0.034	0.152	0.037	0.153
	Cox	0.021	0.151	0.062	0.130
	Buckley-James*	0.026	0.158	0.030	0.177
75%ile	Tree	0.028	0.093	0.029	0.094
	Cox	0.016	0.093	0.021	0.090
	Buckley-James*	0.016	0.101	0.019	0.112

*With 20% censoring model fit did not converge in 28.85% of simulations.

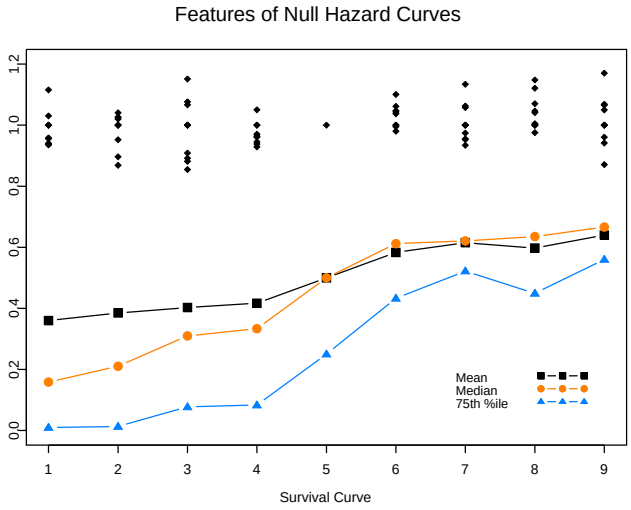


Figure 2: Features of survival curves used for the null hazard scenario: mean, median, and 75^{th} percentile as labeled; black diamonds across the top indicate hazard ratios with other curves.

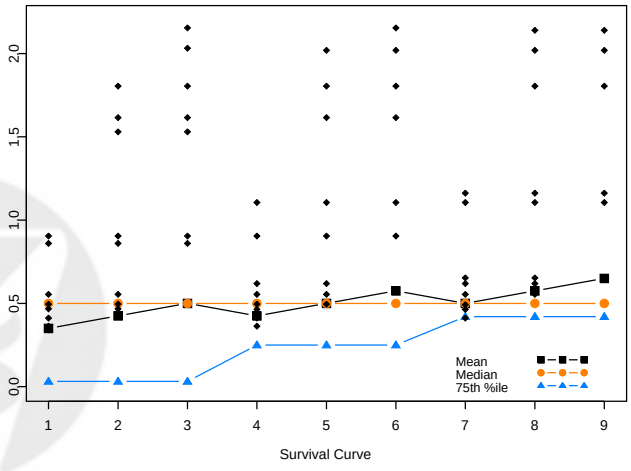


Figure 3: Features of survival curves used for the null median scenario: mean, median, and 75^{th} percentile as labeled; black diamonds indicate hazard ratios with other curves.

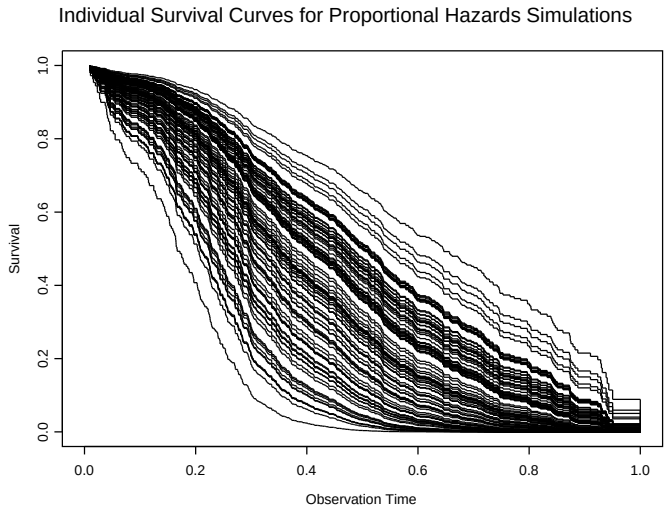


Figure 4: Sample of 100 individual survival curves used for a proportional hazards scenario.

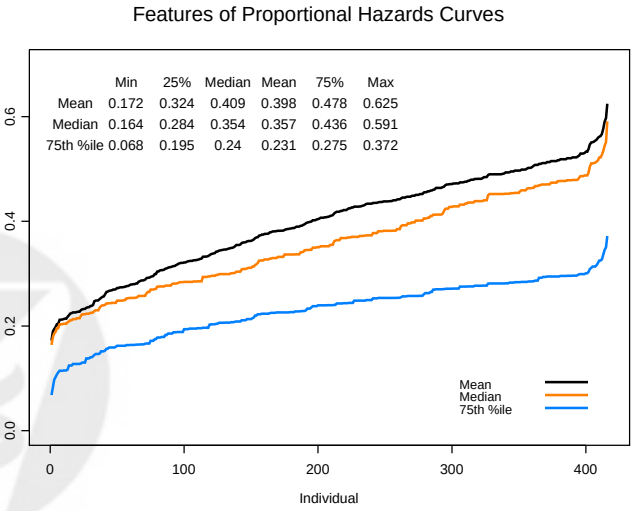
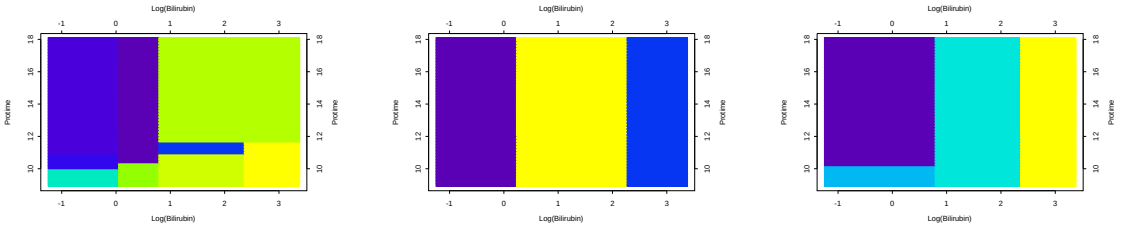
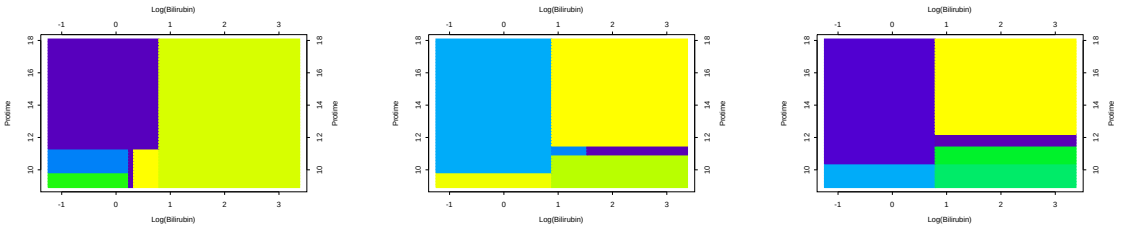


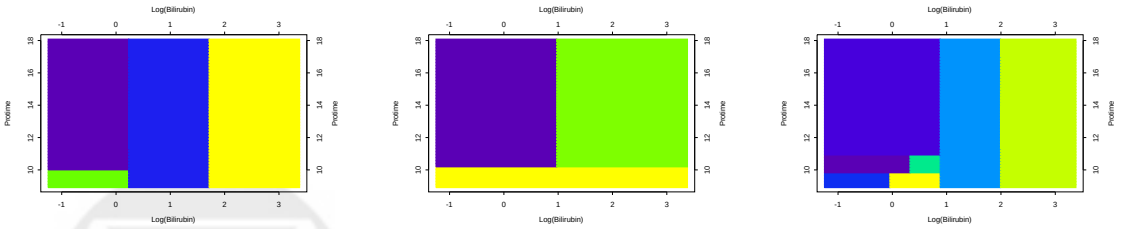
Figure 5: Features of survival curves used for the proportional hazards scenario: mean, median, and 75^{th} percentile as labeled; summary statistics of the three functionals displayed in upper left.



(a) Underlying Groups (b) Simulated Times 1 (c) Simulated Times 2



(d) Simulated Times 3 (e) Simulated Times 4 (f) Simulated Times 5



(g) Simulated Times 6 (h) Simulated Times 7 (i) Simulated Times 8

Figure 6: Plot (a) shows the nine underlying groups defined by bilirubin (logbili) and prothrombin time (protime). Plots (b)-(i) show 8 examples of estimated groups obtained with the tree approach from simulated survival times.