CONSEQUENCES OF THE TWO-STAGE APPROACH: COMPARING TREATMENTS WHEN SURVIVAL CURVES MAY CROSS

by

Harlan Campbell

B.Sc., McGill University, 2008

THESIS PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN THE
DEPARTMENT OF STATISTICS & ACTUARIAL SCIENCE
FACULTY OF SCIENCES

SIMON FRASER UNIVERSITY
Summer 2011

APPROVAL

Name: Harlan Campbell

Degree: Master of Science

Title of Thesis Project: Consequences of the two-stage approach: comparing treatments when survival curves may cross

Examining Committee: Dr. Douglas Woolford (Chair)

Dr. Charmaine Dean
Statistics and Actuarial Science
Simon Fraser University
Senior Supervisor

Dr. Richard Lockhart
Statistics and Actuarial Science
Simon Fraser University

Dr. Lawrence McCandless
Health Sciences Simon Fraser University
External Examiner

Date Approved: August 12, 2011

Abstract

For testing the efficacy of a treatment in a clinical trial (e.g. treatment vs. control), the Cox proportional hazards model is the well-accepted, conventional tool. When using this model, one must confirm that the required proportional hazards (PH) assumption holds true. If the PH assumption fails to hold, it may occur that upon examining a Kaplan-Meier (KM) plot, the survival curves appear to cross, suggesting long-term survival is higher among one group of patients. In this situation –given that the PH assumption does not hold, and given that the KM survival curves are observed to cross– there are options available, proposed as alternatives to the Cox PH model, which are used to test that a treatment yields better long-term survival. An important question which arises is whether the potential bias introduced by such a sequential model fitting procedure merits concern and, if so, what are effective mechanisms for correction. We investigate by means of simulation study and draw attention to the considerable drawbacks, with regards to power, of a simple resampling technique, the permutation adjustment, a natural recourse for addressing such challenges. Finally, we consider the recently proposed two-stage testing strategy of Qiu & Sheng (2008) and a new procedure based on permutation-adjusted bootstrap model averaging, as attractive alternatives.

Acknowledgments

This thesis would not have been possible without the patience and fortitude of my supervisor Dr. Charmaine Dean. She demanded of me a work ethic which elevated my abilities; gave me the academic freedom to pursue new ideas; and provided financial support that made my research possible.

I am also obliged to Dr. Richard Lockhart and (soon to be Dr.) Darby Thompson for encouraging new ideas, dismissing outrageous ones, and assisting me with many challenges encountered throughout my studies at SFU. I am also thankful to the entire Statistics department, who foster a creative and friendly atmosphere; my fellow students, who tolerate my questionability; and to the members of my lab, Alisha, Caroline, Elizabeth and the aforementioned Darby, with whom the days at our desks have flown by at breakneck speed.

To many others I am grateful, those without whom I would not have chosen to pursue my masters in statistics: Dr. Russell Steele, whose enthusiastic teaching of regression first ignited my interest in statistics and is perhaps most responsible for my undergraduate success; Dr JF Williams who wholeheartedly welcomed me to SFU as an undergraduate for a summer during which I first experienced the joys of research.

But there is more than just work, I am grateful to the friends who have made my time in Vancouver magical. Foremost, I must thank Eric and Petr for the outrageous weekend adventures into the mountains -seldom a weekend was missed; Tayu for questioning common assumptions, and common sense; and Robin, for everything.

Last and most I thank my family, for always allowing me, and encouraging me, to follow whatever I have done, with enthusiasm and words of wisdom.

Contents

List of Tables

List of Figures

1. Introduction

> *In the world of cancer research, there is something called a Kaplan-Meier curve, which tracks the health of patients in the trial of an experimental drug. In its simplest version, it consists of two lines. The first follows the patients in the "control arm", the second the patients in the "treatment arm". In most cases, those two lines are virtually identical. That is the sad fact of cancer research: nine times out of ten, there is no difference in survival between those who were given the new drug and those who were not. But every now and again–after millions of dollars have been spent, and tens of thousands of pages of data collected, and patients followed, and toxicological issues examined, and safety issues resolved, and manufacturing processes fine-tuned–the patients in the treatment arm will live longer than the patients in the control arm, and the two lines on the Kaplan-Meier will start to diverge.*
>
> – M. Gladwell, *The New Yorker, 2010*

When one wishes to relate event times to one or more covariates, a common and straightforward approach is to employ the Cox proportional hazards model (Cox, 1972). The proportional hazards (PH) model is the most widely used model in the analysis of censored clinical trial data, where one seeks to relate the time of an individual's death (or other illness related event) to whether or not the individual received a potentially beneficial treatment.

The PH model is formulated such that the covariate effect is multiplicative with respect to the hazard rate, defined as the instantaneous risk of event occurrence. For example, a particular drug treatment may halve the hazard rate of dying for those suffering from cancer, and the *hazards ratio*, as described by the PH model, would therefore be $\frac{1}{2}$. The model requires the assumption that the hazards ratio be constant over the entire follow-up period. In other words, the model assumes that the covariate effect is constant during the entire time the individuals are observed. While this PH assumption may be reasonable in many situations, it may not hold in others. For example, among cancer patients, those receiving treatment requiring elevated doses of chemotherapy may tend to have higher early mortality due to the toxicity of the chemotherapy. However, those who survive the early stages of treatment may benefit from a lower long-term hazard rate if the treatment is effective. See Therneau and Grambsch (2000, ch. 6.6) for a review of different causes of non-proportionality.

In situations when the PH assumption fails to hold, the PH model may not be appropriate as it can produce erroneous results. Therefore, in order to avoid any misleading conclusions, when analyzing time-to-event data, one should first verify that the PH assumption holds before fitting the model to generate estimates about covariate effects. Numerous tests for the validation of the PH assumption have been proposed, see for example Kraus (2007) and Kvaløy & Neef (2004). The most popular is a test attributed to Grambsch & Therneau (1994), (G&T). Typically, if such a test invalidates the PH assumption, one subsequently alters the PH model or employs an altogether different method for the analysis.

Despite the fact that many adequate and flexible alternatives to the PH model are available, these may be less powerful and considerably less interpretable to the average practitioner. Thus, the typical model fitting procedure would first try to fit the data to a PH model, and only consider alternatives, such as the popular accelerated failure time (AFT) model, in the event that a G&T type test indicates a lack of proportionality.

A situation which often occurs when hazards are found to be non-proportional, is that, upon observing the Kaplan-Meier (KM) plot, the survival curves appear to cross. This may suggest that, although initial risk may be higher in one treatment group, long-term survival may be better. Consider a simple case: a potentially beneficial treatment as a binary covariate (treatment vs. control) and survival times as right-censored outcome variables. When a test, such as the G&T test, invalidates the PH assumption, it appears, upon examining the KM plot, that the survival curves of the treatment and control strata cross, suggesting that long-term survival may be higher in the treatment arm. The Cox PH and AFT models are inappropriate in this situation, and other common methods for survival analysis are problematic.

Putter et al. (2005) considers a similar situation and describes the process of fitting an adjusted Cox PH model with time-varying coefficients (PHTVC) to account for the crossing of the survival curves. This approach seems practical and straightforward to implement; see Therneau and Grambsch (2000, ch. 6.2). However, the adjusted model has the disadvantage that one must choose a form describing how the effect of the treatment changes over time, Putter et al. (2005). Perperoglou et al. (2007) consider the PHTVC and suggest using reduced rank regression to overcome the drawback of choosing a functional form for the treatment effect. Their approach attempts to remove some of the subjectivity in the decision by allowing the AIC criteria to guide one's choice among a large number of candidates. The authors also consider the use of frailty models and cure rate models, which are not as straightforward to interpret and more complex.

In an effort to avoid complexity, Logan, Klein & Zhang (2008) recommend a number of simple tests for comparing treatments in the presence of crossing survival curves. These tests compare the long-term survival of patients and require that a time point t_0 be "pre-specified" (before one obtains the data), such that survival beyond t_0 is considered long-term. This value t_0 must be chosen such that the survival curves are likely to cross prior to that time point, if at all. Unfortunately, it is often the case that no such prior knowledge is available and such a pre-specification cannot be reasonably made. Mantel & Stablein (1988), in a similar study, recognize this inconvenience –"admittedly, this is a difficult situation to envisage"– and consider "letting the data suggest the crossing point."

A common concern with these approaches is that the uncertainty associated with the staged model selection procedure is not taken into account. While the unfortunate practice of ignoring model uncertainty is not limited to the analysis of time-to-event data –Breiman (1992) deems this a "quiet scandal in the statistical community"– it has, for the most part, been left unaddressed within the survival analysis literature. Notable exceptions include Altman & Anderson (1989), who consider using the bootstrap to validate the stability of a chosen Cox model, and Sauerbrei & Schumacher (1992), who discuss the use of bootstrapping for variable selection. Yet bootstrapping remains unpopular.

Sauerbrei & Royston (2007) draw attention to the fact that bootstrapping plays an unfortunately negligible role in the analysis of clinical trial data. The authors acknowledge that "well-known problems of data-dependent model building, such as over-fitting the data or biased estimates of parameters, are possible reasons for modeling not playing a more important role in clinical research." They argue that the bootstrap and other resampling techniques could, and should, play an important role to overcome these issues.

While resampling techniques may be useful for variable selection and model validation in prognostic studies (see e.g., Augustin et al., 2005) –where one may have dozens of possible covariates resulting in thousands of possible models– *are they desirable for addressing the situation of crossing survival curves?* Perhaps, but the practice is ignored. Shepherd (2007) notes: "A highly cited paper recommends bootstrapping the process of examining and correcting possible violations of the PH assumption (Harrell, 1996). However, I know of no analysis that has bootstrapped the process of checking the PH assumption." Shepherd (2007) draws attention to this issue and studies the differences one obtains in confidence intervals when properly accounting for model selection by bootstrapping. Unfortunately, the cost of checking the PH assumption under the null hypothesis is left unaddressed.

In this work, we wish to determine the consequences of the common two-stage approach: fitting a Cox PH model if there is no evidence against the PH assumption, while using an alternative test for a treatment effect in the event that the PH assumption is rejected and the KM survival curves appear to cross. We investigate the merits of different two-stage testing strategies by simulation and discuss the results and implications in Section 3. In Section 4, we investigate the possibility of permutation adjustment to overcome the bias encountered by several two-stage procedures, and consider the effects such an adjustment may have on power. Finally, we consider an alternative two-stage procedure proposed by Qiu & Sheng (2008) and a new technique based on permutation-adjusted bootstrap model averaging in Section 5. We begin with a review of the Cox PH model and popular alternatives.

2. Methods

2.1 The Cox proportional hazards model

Definitions

The proportional hazards (PH) model is the most widely used model in the analysis of censored clinical trial data, due to the fact that it is both flexible and powerful. In the simple scenario described in the introduction, we consider a potentially beneficial treatment as a binary covariate (treatment vs. control) and survival times as a right-censored outcome variable. Let X_i, i indexing individuals, be the binary (0/1) covariate, and T_i, be the recorded event or censoring time of the i^{th} individual under study, $i = 1, ..., n$. Also, let $\delta_i = 0$ indicate that the i^{th} individual has been censored, while $\delta_i = 1$ indicates that the i^{th} individual is uncensored. The Cox PH model relates the hazard function, $h(t)$, defined as the instantaneous risk of event occurrence:

$$h(t) = \lim_{\Delta t \to 0} \frac{\text{Pr}(t \leq T_i < t + \Delta t | T_i \geq t)}{\Delta t}, \tag{2.1}$$

to the covariate:

$$h(t|X_i) = h_0(t)\text{exp}(\beta X_i). \tag{2.2}$$

The function $h_0(t)$ is defined as the non-parametric baseline hazard (the hazard function of those individuals for whom $X = 0$) and, with partial likelihood approaches typically employed for inference in this model, estimation of β is not affected by choice of $h_0(t)$. The hazard ratio (HR) is defined as the ratio of the hazard rates of treatment and control groups. Suppose that the i^{th} individual is in the treatment group, $X_i = 1$, while the j^{th} individual is in the control group, $X_j = 0$. Then the hazard ratio is simply:

$$\frac{h(t|X_i = 1)}{h(t|X_j = 0)} = \frac{h_0(t)\text{exp}(\beta X_i)}{h_0(t)\text{exp}(\beta X_j)} = \frac{\text{exp}(\beta X_i)}{\text{exp}(\beta X_j)} = \text{exp}(\beta). \tag{2.3}$$

More generally, for $\bar{X}_i$ a (1 x p) vector of covariates operating on individual i, and β the corresponding (p x 1) coefficients, the form of equation (2.2) is similarly partitioned into the effects of time and the covariates as $h_0(t)\text{exp}(\bar{X}_i\beta)$. In parallel with equation (2.3), the ratio of the hazard functions corresponding to two individuals with different covariate vectors does not depend on t. The maximum partial likelihood estimator, $\hat{\beta}$, is the value that maximizes the partial likelihood function defined as in Cox (1972):

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\text{exp}(\beta' X_i)}{\sum_{j \in R_i} \text{exp}(\beta' X_j)} \right)^{\delta_i}, \tag{2.4}$$

or alternatively the partial *log*-likelihood function:

$$l(\beta) = \sum_{i=1}^{n} \delta_i(\beta' X_i) - \sum_{i=1}^{n} \delta_i ln \left(\sum_{j \in R_i} \text{exp}(\beta' X_j) \right), \tag{2.5}$$

where R_i is the set of indexes identifying individuals who remain at risk just before the i^{th} event time. It can be shown, see Anderson & Gill (1982), that :

$$\frac{1}{\sqrt{n}}(\hat{\beta} - \beta) \rightarrow N(0, \frac{1}{n}I(\beta)), \tag{2.6}$$

where $I(\beta)$ is the expected Fisher Information matrix ($I(\beta) = -\text{E}(\frac{\delta^2 l}{\delta \beta^2})$). This asymptotic result forms the foundation for the Wald and likelihood ratio tests (LRT), for testing the significance of $\hat{\beta}$. The Wald test statistic, X_W, is calculated as, $\hat{\beta}'[I(\hat{\beta})]^{-1}\hat{\beta}$. For large samples, under the null hypothesis, $X_W \sim \chi^2_1$. The likelihood ratio statistic, X_{LRT}, is calculated as $2l(\hat{\beta}) - 2l(\beta_0)$, with $\beta = \beta_0$, $X_{LRT} \sim \chi^2_1$, under the null, for large n.

Bangdiwala (1989) summarizes the difference between the Wald test and LRT: “The likelihood ratio test statistic is well behaved in most situations but may be expensive to calculate. The Wald test statistic is easier to calculate, but has some drawbacks.” The likelihood-ratio and Wald statistics are asymptotically equivalent tests and most often result in similar, if not identical conclusions. However, the LRT statistic is preferable for many practitioners, as it converges more quickly to the χ^2 asymptotic form. As such, we will use the LRT throughout this project to determine significance of $\hat{\beta}$.

2.2 The Grambsch & Therneau (1994) test

The test for PH proposed by Grambsch and Therneau (1994) is based on checking constant PH against the alternative of non-proportional time-varying hazards. In the case of no tied event times (i.e. all event times are unique), the Schoenfeld residual at the i^{th} event time is defined as:

$$s_i = X_i - \bar{x}(\hat{\beta}, t_i) \qquad \text{where,} \qquad \bar{x}(\hat{\beta}, t_i) = \frac{\sum_{j \in R_i} \text{exp}(\hat{\beta}' X_j) X_j'}{\sum_{j \in R_i} \text{exp}(\hat{\beta}' X_j)}, \tag{2.7}$$

so that s_i is the difference between the treatment group indicator for an individual failing at time t_i and an expected value of this indicator for all individuals in the risk set R_i. The Schoenfeld residuals can be scaled by a time-dependent variance matrix, and Grambsch and Therneau (1994) show that these scaled Schoenfeld residuals, s_i^*, have approximate expectation:

$$\text{E}(s_i^*) \approx \beta(t_i) - \hat{\beta}, \tag{2.8}$$

where $\beta(t)$ is the time-varying coefficient for an alternative model where the covariate effects change over time (c.f. Section 2.3 below which discusses the adjusted Cox PH model with hazard $h(t|X_i) = h_0(t)\text{exp}(X_i\beta(t)))$. Plotting $s_i^* + \hat{\beta}$ against t provides a natural way to examine the merits of the PH assumption; usually a smoother is overlaid on the plot to visually assess goodness of fit. Testing for significance of the slope results in the G&T test for PH. Let $\alpha_{G\&T}$ be the significance level of the test, typically set to equal 0.05. Different choices for the functional form of $\beta(t)$ result in somewhat different tests; for a given $\beta(t)$, t is appropriately scaled. Therneau and Grambsch (2000) note that: “for long-tailed survival distributions [...] log(t) is often a good choice.” As such we will use log(t)

in all simulations studies when testing for proportional hazards. For additional details on the G&T test, see also Therneau and Grambsch (2000, ch. 6).

2.3 Alternatives to the Cox PH model

The following subsections review three common approaches, proposed as alternatives to the Cox PH model.

Adjusted Cox PH model with time-varying coefficients (PHTVC)

The adjusted Cox PH model with time-varying coefficients (PHTVC) can accommodate non-proportional hazards as well as crossing survival functions. Consider the following model:

$$h(t|X_i) = h_0(t)\text{exp}(X_i\beta(t)) = h_0(t)\text{exp}[\beta_0 X_i + \beta_1(X_i f(t))], \tag{2.9}$$

where $f(t)$ incorporates the time-varying treatment effect. Three common choices for $f(t)$ are: $\log(t)$, $\sqrt{t}$, and t; the scaled Schoenfeld residuals may guide one in the choice of $f(t)$, see Putter et al. (2005). We can test for overall significance of treatment effect, by employing a LRT which compares the likelihood ratio statistic to the χ^2_2 distribution. As mentioned in the introduction, a disadvantage with this model is that one must choose a functional form for the time-varying treatment effect.

Perperoglou et al. (2007) consider the PHTVC and suggest using reduced rank regression to overcome the drawback of needing to specify a functional form for the treatment effect. The approach attempts to remove some of the subjectivity in the decision by allowing AIC criteria to guide one's choice among a large number of candidates. The choice remains subjective to a certain degree however, as exemplified by the authors' decision in an application to use a more parsimonious model than the one selected by AIC alone.

Many authors have also advocated the use of splines to model the changing covariate effect over time, see for example Thompson (2011), Muggeo & Tagliavia (2010) and Abrahamowicz et al. (1992).

Log-Rank tests

The common Log-Rank (also known as the Mantel-Cox) statistic forms the basis for several tests appropriate in the presence of non-proportional hazards and crossing survival curves.

Logan, Klein & Zhang (2008) assume that a time point t_0 can be "pre-specifed" (before one obtains the data), such that survival beyond t_0 is considered long-term with t_0 chosen such that the survival curves are likely to cross prior to that time point, if at all. They then consider a *post*-t_0 log-rank test:

$$\begin{aligned}
&Z_{LR}(t_0) = \frac{X_{LR}(t_0)}{\hat{\sigma}_{LR}(t_0)} \sim N(0,1),\\
\text{where:}\qquad &X_{LR}(t_0) = \sum_{t_j > t_0} \frac{Y_{1j}Y_{0j}}{Y_j}\left(\frac{d_{1j}}{Y_{1j}} - \frac{d_{0j}}{Y_{0j}}\right),\\
&\hat{\sigma}^2_{LR} = \sum_{t_j > t_0} \frac{Y_{1j}Y_{0j}}{Y_j^2}\left(\frac{Y_j - d_{1j}}{Y_j - 1}\right) d_j, \qquad (2.10)
\end{aligned}$$

where Y_{kj}, Y_j denote the number at risk at t_j in the k^{th} group and in total; d_{kj} and d_j denote the number of events at t_j for the k^{th} group and in total. In addition to the post-t_0 log-rank test, a series of combination tests where $X_{LR}(t_0)$ is calculated in combination with other statistics are also investigated. In a simulation study, among all tests considered, the *post*-t_0 log-rank test is found to have the most power to detect long-term difference when the underlying model is one of crossing survival curves. However, in the case when the true model is of proportional hazards, the *post*-t_0 log-rank test is found to have very low power (less than half the power of the unweighted log-rank test in the scenario considered). Thus, unless one is very confident that the data will exhibit non-proportional hazards, it cannot be recommended. It must be noted that the log-rank test is testing the hypothesis of a difference in hazard functions; therefore, it is most powerful when t_0 is the time-point at which the hazards cross, well before the crossing of survival curves.

Mantel & Stablein (1988) consider a similar log-rank test which requires one to pre-specify a time-point, t_0, where there is a "reversal in the merits of [the] two hazard functions." They define:

$$X_{LR}(t_0) = \sum_{t_j < t_0} \frac{Y_{1j}Y_{0j}}{Y_j}\left(\frac{d_{1j}}{Y_{1j}} - \frac{d_{0j}}{Y_{0j}}\right) \cdot Z(t_j, t_0),$$

$$\begin{aligned} \text{where,} \qquad Z(t_j, t_0) &= 1, \quad t_j < t_0 \\ &= -1, \quad t_j > t_0. \end{aligned} \tag{2.11}$$

Mantel & Stablein (1988) recognize the drawback of having to pre-specify a crossing point –“admittedly, this is a difficult situation to envisage”– and consider “letting the data suggest the crossing point.” Unfortunately, no remedy to the bias encountered due to the data-driven search is discussed.

Several other log-rank type tests have been proposed. These include the Peto-Peto (1972) log-rank test which places more weight on earlier time points, a weighted log-rank test with additional weight placed on later time points proposed by Fleming and Harrington (1982), and a weighted log-rank test that emphasizes early and/or late differences studied by Wu & Gilbert (2002). These have been found to be very poweful in detecting many non-PH alternatives.

It is worth noting that the unweighted log-rank test is asymptotically equivalent to the Cox PH model LRT, and also relies on the assumption of PH. As such, one should expect, and indeed one will find that, weighted-log-rank tests are less powerful than the Cox PH LRT in detecting the alternative of proportional hazards. In section 5, we discuss the work of Qiu & Sheng (2008) who develop another weighted log-rank test with weights derived such that the test is asymptotically independent of the unweighted log-rank test.

Accelerated failure time model

The parametric accelerated failure time model (AFT) is a popular alternative to the PH model, when the PH assumption fails, see Wei (1992). Implementation is straightforward, and results are rather interpretable. The model, however, is not appropriate if there are any crossovers in the survival functions. AFT models can be thought of as linear models for the logarithm of the survival time:

$$\log(T_i) = X_i'\beta + \sigma\epsilon_i, \tag{2.12}$$

where β and σ are parameters to be estimated. One must select a distribution for ϵ_i. A popular choice is the extreme value distribution. The corresponding distribution of the event-times, T_i, is then the Weibull distribution. The popularity of this parametrization can be attributed to the inherent flexibility the Weibull distribution provides, see Co (2010). As such, we will use this parametric choice in all simulation studies, when applying the AFT model.

2.4 The common two-stage approach

As discussed in the introduction, prevailing practice for analysis of time-to-event data would first try to fit the data to a PH model, and only consider alternatives, such as those mentioned above, in the event that a G&T type test indicates a lack of proportionality. We designate this procedure the "common two-stage approach", see Figure 2.1. The most general hypothesis test under consideration is:

$$\begin{aligned} H_0 : \quad & h(t|X=0) = h(t|X=1) \Leftrightarrow S(t|X=0) = S(t|X=1), \text{for all } t; \\ H_1 : \quad & h(t|X=0) \neq h(t|X=1) \Leftrightarrow S(t|X=0) \neq S(t|X=1), \text{for some } t. \end{aligned} \tag{2.13}$$

In order to investigate the merits of different second-stage alternative methods, we conducted simulation studies, generating data under the null hypothesis.

The "common two-stage approach"

1. Fit a Cox PH model and test the PH assumption with a G&T type test.
2. - If proportionality holds (G&T p-value $> \alpha_{G\&T}$), the Cox PH LRT determines the strength of evidence against H_0.
   - If proportionality does not hold (G&T p-value $\leq \alpha_{G\&T}$), employ an alternative method to determine the strength of evidence against the H_0.

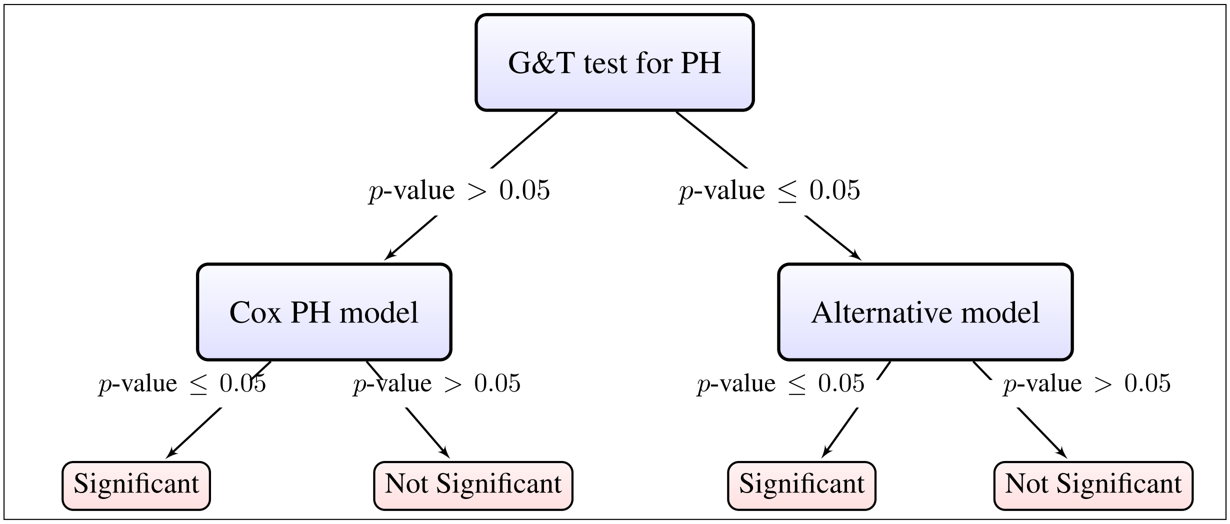

Figure 2.1: The common two-stage approach.

3. The common two-stage approach under the null

3.1 Simulation Study I

In order to assess the appropriateness of the common two-stage approach, we investigate its performance under the null model by simulation study. Recall the probability density function of the Weibull distribution:

$$f(x; \lambda, k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \quad x \geq 0. \tag{3.1}$$

We simulated event-times from the Weibull distribution with parameters $k = 0.6$ and $\lambda = 83.293$; see Figure 3.1. In total, 100 event-times are simulated in each simulation experiment, half of which are attributed to "a control group" ($X = 0$) and half to a "treatment group" ($X = 1$). Two censoring scenarios are considered: (1) right censoring at $t = 72$, when survival probability equals 0.4; and (ii) right censoring at $t = 72$ with additional independent exponential censoring generated such that approximately 15% of individuals are censored by $t = 24$. The settings used were set to mimic the simulation studies of Logan, Klein & Zhang (2008). We evaluate the following hypothesis test:

$$\begin{aligned} H_0: &\quad h(t|X=0) = h(t|X=1) \Leftrightarrow S(t|X=0) = S(t|X=1); \\ H_1: &\quad h(t|X=0) \neq h(t|X=1) \Leftrightarrow S(t|X=0) \neq S(t|X=1). \end{aligned}$$

and consider four different two-stage testing schemes, all of which first test for PH by the G&T test and subsequently fit a Cox PH if the G&T test fails to reject proportionality (i.e. if the p-value from the G&T test is greater than $\alpha_{G\&T}$ = 0.05). In the event that proportionality is suspect (i.e. if the p-value from the G&T test is less than or equal $\alpha_{G\&T}$ = 0.05), the four possibilities are:

- **TVCa**: the adjusted Cox PH model with time varying coefficients and $f(t) = \log(t)$;
- **TVCb**: the adjusted Cox PH model with time varying coefficients and $f(t)$ chosen as best fit among $f(t) = \log(t)$, $f(t) = \sqrt{t}$, and $f(t) = t$;
- **LogRT**: the log-rank test with $t_0 = 24$;
- and **AFT**: the accelerated failure time model.

We generated 100,000 sets of lifetimes and 100,000 sets of censoring times for each of the two censoring scenarios. For each dataset, each of the two censoring scenarios, and each of the four two-stage approaches, we recorded all p-values, and whether or not the null hypothesis was rejected at the $\alpha = 0.05$ significance level, under the two-stage approach. We compare these numbers to those expected if the sequential tests were fully independent.

3.2 Results

Table 3.1 displays the proportion of runs for which the null hypothesis was rejected, while Figure 3.2 displays the joint distribution of the observed p-values and χ^2 test statistics. Let us first consider the results from the simulations when there is no additional censoring. The results show that p-values obtained from the G&T test and those obtained from the adjusted Cox PH with functional form $f(t) = \log(t)$ are strongly correlated, (Figure 3.1). Under the null model, if one rejects proportionality on the basis of a small G&T p-value, the probability that the adjusted Cox PH model will suggest a significant effect is greater than 50%, more than 10 times the expected rate under independent tests (5%), (Table 3.1, TVCa). With the data-driven choice of the $f(t)$, the bias is even worse with a recorded rejection rate of 7.65%, (Table 3.1, TVCb). The $f(t) = \log(t)$ choice was found to be the best for approximately 50 % of the runs while $f(t) = t$ and $f(t) = \sqrt(t)$ were the best fit for approximately 35% and 15% of the runs respectively.

The log-rank test also shows bias, albeit not as strongly, with a rejection rate of 5.75% (Table 3.1, LogRT), which is significantly different from 5% under 100,000 runs conducted. The two-stage approach with the AFT model as alternative to the Cox PH model appears relatively safe from bias, (Table 3.1, AFT). Finally, the G&T rejection rate of 95.232% rather than the expected 95.00% can be attributed to Monte Carlo error. Results of simulations with additional censoring yield similar conclusions.

We note no evidence here that G&T p-values and Cox PH p-values are correlated. Similarly, we see no evidence of correlation between the AFT p-values and the G&T p-values. The possible correlation between G&T p-values, Cox PH p-values and AFT p-values merits further investigation. In all cases, except that of the AFT model, the bias resulting from the two-stage approach is very alarming: we obtain a much greater type-I error than the desired 5.0%, as large as 7.7% in one situation. In Section 4, we consider possible means of correction by permutation adjustment.

No additional censoring					
Result	**TVCa**	**TVCb**	**LogRT**	**AFT**	**Independent Tests**
Significant (CoxPH)	4.751	4.751	4.751	4.751	4.750
Not Significant (CoxPH)	90.481	90.481	90.481	90.481	90.250
Significant (Alternative)	2.577	2.899	1.000	0.224	0.250
Not Significant (Alternative)	2.191	1.869	3.768	4.544	4.750
Significant	7.328	7.650	5.751	4.975	5.000
Not Significant	92.672	92.350	94.249	95.025	95.000
Additional censoring					
Result	**TVCa**	**TVCb**	**LogRT**	**AFT**	**Independent Tests**
Significant (CoxPH)	4.751	4.751	4.751	4.751	4.750
Not Significant (CoxPH)	90.460	90.460	90.460	90.460	90.250
Significant (Alternative)	2.633	2.966	0.995	0.224	0.250
Not Significant (Alternative)	2.156	1.823	3.974	4.565	4.750
Significant	7.384	7.717	5.746	4.975	5.000
Not Significant	92.616	92.283	94.254	95.025	95.000

Table 3.1: **Results of Simulation Study I.** Proportion of runs in which the null hypothesis is rejected out of 100,000 runs, each simulating 100 event-times from a Weibull(k=0.6, λ=83.293). For comparison, ‘Independent Tests’ shows proportions expected if sequential tests were fully independent.

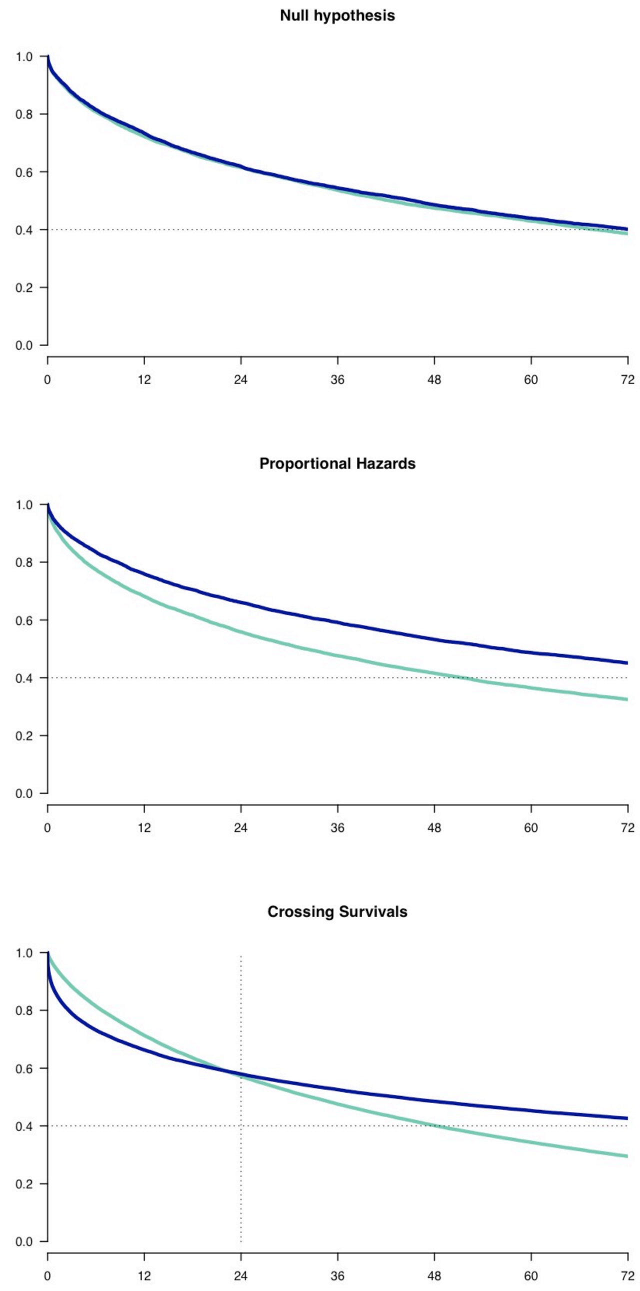

Figure 3.1: Survival curves of the scenarios considered in simulation studies: Null scenario of no treatment effect (top panel); PH scenario with treatment effect, odds ratio of 1.75 at $t = 72$ (centre panel); and CS scenario with treatment effect, odds ratio of 1.75 at $t = 72$, survival curves cross at $t = 24$ (lower panel).

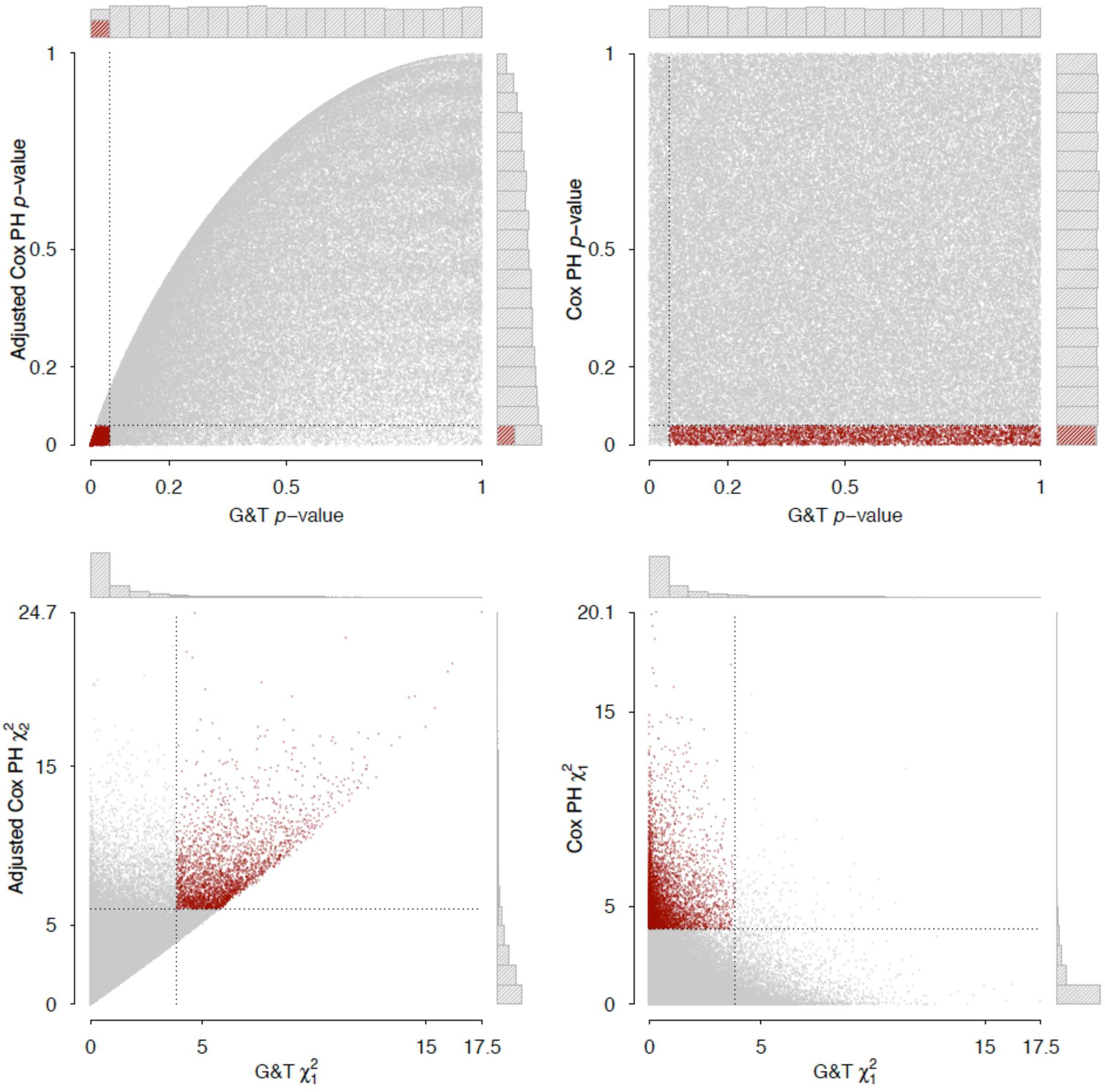

Figure 3.2: Joint distribution of p-values (top panels) and test statistics (lower panels) from G&T and adjusted Cox PH model (**TVCb**). Red points indicate significant outcomes. Non-independence between PHTVC model and G&T test is evident (left).

4. Correcting for bias by permutation adjustment

Permutation adjustment is a popular resampling method for obtaining non-biased p-values, see Routledge (1997). We consider two straightforward permutation-adjustments: *Top-down* and *Conditional*. Intuitively, the conditional permutation adjustment conditions on the outcome of the G&T test while the top-down permutation adjustment repeats the entire two-stage procedure.

Top-down Permutation adjustment (TDP)

For j in 1 in J, where J is large:

1. Permute the treatment labels of the original data, to yield permuted treatment assignments to each of the responses; denote the permuted data as $\tilde{D}_j$.

2. Apply the common two-stage approach to $\tilde{D}_j$.

3. Obtain a p-value for a test of no treatment effect; denote this as $\tilde{p}_j$.

Our conditional-permutation-adjusted p-value, p_{TDP}, is the proportion of those $\tilde{p}_j$s which are smaller or equal to the p-value obtained based on an analysis of the original data under the common two-stage approach, p; p_{TDP} determines the strength of evidence against H_0. Let $\mathbb{1}(A)$ denote the indicator function for event A;

$$p_{TDP} = \sum_{j=1}^{J} \frac{\mathbb{1}(\tilde{p}_j \leq p)}{J}.$$

Conditional Permutation adjustment (CP)

For j in 1 in J, where J is large:

1. Permute the treatment labels of the original data, to yield permuted treatment assignments to each of the responses; denote the permuted data as $\tilde{D}_j$.
2. Apply the common two-stage approach to $\tilde{D}_j$.
3. Obtain a p-value for a test of no covariate (treatment) effect; denote this as $\tilde{p}_j$. Record the p-value from the G&T test; denote this as $\tilde{p}(G\&T)_j$. Let $\mathbb{1}(G\&T)_j$ define rejection by the G&T test: $\mathbb{1}(G\&T)_j = \mathbb{1}(\tilde{p}(G\&T)_j \leq \alpha_{G\&T})$.

The conditional permutation test uses the separate distributions of $\tilde{p}_j$ for which $\mathbb{1}(G\&T)_j = 0$ and $= 1$ to construct a p-value, with the choice of conditional distribution reflecting the result obtained in the original analysis. Let p denote the p-value obtained for a test of treatment effect in the original analysis and $\mathbb{1}(G\&T)$ be an indicator for the rejection of the PH assumption by the G&T test in the original analysis. Then:

$$p_{CP} = \sum_{j:\mathbb{1}(G\&T)_j=\mathbb{1}(G\&T)} \frac{\mathbb{1}(\tilde{p}_j \leq p)}{\#\{j : \mathbb{1}(G\&T)_j = \mathbb{1}(G\&T)\}}.$$

4.0.1 Simulation Study II : Comparison of Permutation Adjustments

In order to assess the performance of the two permutation adjustment methods considered, we simulated data under the null model, under a model of PH, and a model of crossing survival functions (CS), as we wish to evaluate both size and power of the procedures. Event-times are simulated from a Weibull distribution, as discussed previously, see Figure

3.2. For both the PH and CS scenarios, parameters were set such that the odds ratio of survival at $t = 72$ between groups is 1.75. The crossing of survival curves in the CS scenario is at $t = 24$.

In total, 100 event-times are simulated, half of which are from a "control group" ($X = 0$) and half from a "treatment group" ($X = 1$). As in Simulation Study I, two censoring scenarios are considered: (1) right censoring at $t = 72$; and (ii) right censoring at $t = 72$ with additional independent exponential censoring generated such that approximately 15% of individuals are censored by $t = 24$. Five hundred permutation resamples were performed in each of 10,000 simulation runs to evaluate type-I error and power. We applied both TDP and CP adjustment to PHTVC with data-driven choice of $f(t)$ (**TVCb**). For the null scenario, we compare the results obtained by permutation adjustment to those without permutation adjustment from the 100,000 simulation runs of Simulation Study I.

4.0.2 Results & Interpretation

Table 4.1 presents the proportion of runs for which the null hypothesis is rejected under each of the permutation tests. Results from the unadjusted test of Section 3 are also listed (column labeled 'No Adj') for comparison.

The results of the simulation study suggest that correct type-I error is achieved by both permutation methods, with the observed size being close to the nominal level of 0.05. The small differences can be attributed to the Monte Carlo error obtained using only 500 permutation resamples. CP adjustment has somewhat higher power when the true model is PH. On the other hand, under the alternative of CS, the TDP adjustment has substantially higher power.

This phenomenon can be attributed to the fact that p-values obtained from the Cox PH model tend to be larger than those obtained from the adjusted Cox PH model. Therefore, if one obtains a p-value, p^*, from the PHTVC and compares this to a series of p-values, $p_b s$, obtained from only the adjusted Cox PH model, p^* will appear relatively less significant than if one compares it to p-values from both the Cox PH and the adjusted Cox PH models. Contrast the proportion of significant tests under the alternative model when using TDP versus CP tests under the null: 1.82 versus 0.22, respectively. A similar reasoning concerning p-values from the Cox PH model can be made. With this in mind, unless one has strong prior beliefs as to the presence of proportional hazards, neither of these permutation adjustments behave favorably. If the difference in power achieved through TDP and CP under a PH model is considered relatively minor, than TDP may be a preferred choice. However, neither of these methods seem to provide good performance over all segments of the test. Determining how the relative power of the TDP and CP is influenced by sample size and effect size merits further investigation.

	Null			**PH**			**CS**		
	No Adj	TDP	CP	No Adj	TDP	CP	No Adj	TDP	CP.
No additional censoring									
Sign. (CoxPH)	4.751	3.03	4.85	29.30	21.02	27.63	3.75	2.59	3.84
Not Sign. (CoxPH)	90.481	92.37	90.56	71.74	74.28	67.67	26.25	27.41	26.16
Sign. (Alt.)	2.899	1.82	0.22	3.60	2.99	0.83	62.51	56.30	25.84
Not Sign. (Alt.)	1.869	2.78	4.37	1.40	1.72	3.87	7.49	13.70	44.16
Significant	7.650	4.85	5.07	32.90	24.01	28.46	66.26	58.89	29.73
Not Significant	92.350	95.15	94.93	67.10	75.99	71.54	33.74	41.11	70.27
Additional censoring									
Sign. (CoxPH)	4.751	2.81	4.67	23.74	17.68	23.52	2.60	1.70	2.67
Not Sign. (CoxPH)	90.46	92.63	90.77	65.70	77.80	71.96	34.59	35.49	34.52
Sign. (Alt.)	2.966	1.86	0.38	3.50	2.81	0.68	53.71	46.42	18.85
Not Sign. (Alt.)	1.823	2.70	4.18	1.02	1.72	3.84	9.10	16.39	43.96
Significant	7.717	4.67	5.05	27.24	20.49	24.20	56.31	48.12	21.52
Not Significant	92.283	95.33	94.95	72.76	79.51	75.80	43.69	51.88	78.48

Table 4.1: **Results of Simulation Study II.** Proportion of runs in which the null hypothesis is rejected out of 10,000 runs, each simulating 100 event-times from a Weibull under scenarios of no treatment effect (Null), treatment effect with proportional hazards (PH), and treatment effect with crossing survival curves (CS).

5. Alternatives to the common two-stage approach

5.1 Correcting for bias by Qiu & Sheng (2008)'s two-stage approach

Qiu & Sheng (2008) (Q&S) discuss an alternative two-stage approach. In the first stage, one tests for significance of the treatment effect by using the ubiquitous unweighted log-rank test. If the log-rank test detects a significant effect at the α_1 level, one stops and acknowledges a significant effect, thereby rejecting the null hypothesis of identical hazards. If the first stage log-rank test fails to detect a significant result at the α_1 level, one employs a newly conceived weighted log-rank test, with weights derived such that the test statistic is asymptotically independent of the first stage log-rank test statistic. The weights are negative prior to the supposed crossing point of the hazards and positive afterwards. Since the crossing point is unknown, the test statistic is evaluated with every potential crossing point, and the crossing point for which the test statistic is greatest is chosen for implementing the test. The set of potential crossing points can be restricted to a smaller, more reasonable set by assigning the ϵ tuning parameter accordingly, see Qiu & Sheng (2008) for details. The critical value of the null distribution of this *maximal* test statistic is estimated by bootstrapping the second stage. Figure 5.1 illustrates the Q&S alternative two-stage approach.

Due to the asymptotic independence of the tests at the first and second stages, Q&S can easily define the overall significance level α. Let α_1 and α_2 be the significance levels of the first and second stage tests respectively. Then,

$$\begin{aligned}\alpha &= \alpha_1 + Pr_{H_0}(\text{reject in stage 2}|\text{fail to reject in stage 1})(1-\alpha_1) \\ &= \alpha_1 + \alpha_2(1-\alpha_1). \end{aligned} \tag{5.1}$$

For a given α, and $\alpha_1 \leq \alpha$, we take $\alpha_2 = (\alpha - \alpha_1)/(1-\alpha_1)$. The p-value of the entire procedure is then:

$$p - value = \begin{cases} p_1 & \text{, if } p_1 \leq \alpha_1, \\ \alpha_1 + p_2(1-\alpha_1) & \text{, otherwise.} \end{cases} \tag{5.2}$$

In order to attain maximal power, one must prudently select α_1 and α_2 based on prior belief about the plausibility of non-PH. Intuitively, the choice amounts to partitioning the available power between stage 1 and stage 2 tests. While the need to make such an influential decision may be considered a serious drawback of the methodology, a “neutral” choice of $\alpha_1 = \alpha_2 = 1 - \sqrt{(1-\alpha)}$ is available.

Qiu & Sheng (2008)’s two-stage approach

1. Test for evidence against the null by unweighted log-rank test. If significant at the α_1 level, H_0 is rejected. If non-significant, proceed to stage 2.

2. Test for evidence against the null by first evaluating the Q&S-weighted log-rank test statistic with every potential crossing point. The crossing point for which the test statistic is greatest is then chosen for implementing the test. The critical value of the null distribution of this *maximal* test statistic is estimated by bootstrapping the search. If significant at the α_2 level, H_0 is rejected. Otherwise, one fails to reject H_0.

5.2 Correcting for bias by permutation-adjusted bootstrap model averaging

Bayesian Model averaging (BMA) is a helpful procedure to account for uncertainty arising due to model selection procedures and hence is especially attractive for applications in which superior predictive ability is desired. In practice, predictions derived from BMA are found to be consistently more accurate than those derived from a single "best model", see Draper (1995) and Hoeting (1999). BMA considers the posterior probability of a parameter, θ, given the data, D, and equates:

$$Pr(\theta|D) = \int_{\mathcal{M}} Pr(\theta|D, M)Pr(M|D)dM, \tag{5.3}$$

where $\mathcal{M}$ is the entire space of possible models, $M \in \mathcal{M}$. If one restricts $\mathcal{M}$ to a finite number (K) of models, a BMA parameter estimate ($\hat{\theta}_{BMA}$) can be calculated as:

$$\hat{\theta}_{BMA} = \sum_{k=1}^{K} w_k \hat{\theta}_k, \tag{5.4}$$

where $\hat{\theta}_k$ is the parameter estimate under model k (M_k) and w_k is a weight equal to the posterior probability of model k. These *model weights* can be derived by considering:

$$w_k = Pr(M_k|D) = c \cdot \int Pr(D|\theta_k, M_k)Pr(\theta_k|M_k)Pr(M_k)d\theta_k, \tag{5.5}$$

where c is a constant. In addition to the requirement of defining *parameter priors*, $Pr(\theta|M_k)$, standard in all Bayesian frameworks, one must also define *model priors*, $Pr(M_k)$, which reflect prior belief in the plausibility of the models considered. Practically speaking, the requirement to define *model priors* is not unlike the need to set stage 1 and stage 2 significance levels in the method of Q&S. Similarly, if one has little prior information about the relative plausibility of the models under consideration, taking them all to be equally likely a priori, $Pr(M_k) = Pr(M_l), \forall\, k, l$, would be a "neutral" option.

Volinsky et al. (1997) demonstrate the use of BMA in a PH model as an alternative approach to standard variable selection. While the method requires difficult calculation and careful approximation, its predictive performance is found to be superior to that from standard model selection criteria.

Given the success of BMA, and the reluctance of many practitioners to adopt a Bayesian approach, a non-parametric model averaging procedure has recently been proposed. Originally described by Buckland (1997), the non-Bayesian model averaging scheme derives *model weights* in a way that circumvents the need to establish priors. Buckland (1997) advocates defining model weights as:

$$w_j = \frac{exp(-I_j/2)}{\sum_{k=1}^{K} exp(-I_k/2)}, \tag{5.6}$$

where I_j is the information criterion for model j. One may choose any information criterion such as the AIC or BIC. It has been shown that, using the BIC, one attains weights similar to those attained in equation (5.5) when non-informative priors are used, see Augustin et al. (2005). Buckland (1997) also considers the option of defining the weights by means of nonparametric bootstrapping.

Suppose one samples observations with replacement from the original data such that a "bootstrap data set" ($\tilde{D}_b$) of size equal to the original data series (n) is obtained, on the bth resampling, b = 1, ..., B, with B large. On each occasion, the model selected by the chosen selection criteria is recorded. *Model weights* are then defined as the proportion of resamples in which each model is selected:

$$w_k = \frac{\sum_{b=1}^{B} \mathbb{1}(M_k \text{ is selected}|\tilde{D}_b)}{B}. \tag{5.7}$$

Admittedly, this "bootstrap model averaging" (bMA) is ad-hoc (Augustin et al., 2005) and while it has been found to perform similarly to BMA in terms of predictive performance (see Buchholz, (2008) and Hollander et al. (2006)), $\hat{\theta}_{bMA}$ lacks formal justification and determining a proper estimate for the $var(\hat{\theta}_{bMA})$ has been problematic. For this reason the method has not been widely adopted, despite good potential. However, as we shall argue, the idea seems very reasonable from a frequentist standpoint and most useful to the scenario at hand.

Under a frequentist perspective, a parameter estimate calculated by weighted average, $\hat{\theta} = \sum_{k=1}^{K} w_k\hat{\theta}_k$, is consistent, if, for M_k a correct model, we have:

$$\lim_{n\to\infty} \hat{\theta} = \theta_k = \theta,$$

where $\theta_k (= \theta)$ is the true value of the parameter. (In Appendix A1, a rationale is provided to support this property applying to $\hat{\theta}_{bMA}$.) Consider how bMA may be useful for the hypothesis test considered in Section 2:

$$\begin{aligned} H_0: &\quad h(t|X=0) = h(t|X=1) \Leftrightarrow S(t|X=0) = S(t|X=1); \\ H_1: &\quad h(t|X=0) \neq h(t|X=1) \Leftrightarrow S(t|X=0) \neq S(t|X=1); \end{aligned}$$

specifically when utilizing the Cox PH model (M_0) and the PHTVC model (M_1) as in Section 4. Both of these models provide one with a p-value which evaluates the evidence against H_0 in favor of the chosen model. Let us consider both p-values as statistics, p_0 and p_1, respectively. Despite the fact that, being nested, both models may be "correct", in the common two-stage approach, outlined in Section 2, the G&T test unambiguously selects between the two. The appropriate p-statistic is then compared to the Uniform[0,1] distribution in order for one to reject, or fail to reject, H_0.

We propose the following alternative statistic for evaluating the strength of evidence against H_0:

$$\begin{aligned} p_{bMA} &= w_0 p_0 + w_1 p_1 \\ &\text{where,} \quad w_k = \frac{\sum_{b=1}^{B} \mathbb{1}(M_k \text{ is selected}|\tilde{D}_b)}{B}, \qquad k = 0, 1 \end{aligned} \tag{5.8}$$

While p_{bMA} may not be a conventional choice for hypothesis testing, its behavior seems to be similar to that of an ordinary p-value though this needs formal justification. For instance, we have that, if H_0 is false and H_1 is true:

$$\forall \epsilon > 0, \lim_{n \to \infty} Pr(p_{bMA} \leq \epsilon) = 1. \tag{5.9}$$

see Appendix A2 for details.

The critical value of the null distribution of p_{bMA} can be estimated by permutation test, whereby the entire procedure of calculating p_{bMA} is repeated a large number (P) of times randomly permuting the covariate labels at each instance. The calculation of p_{bMA} and its significance level require substantial computational expense as one must do bootstrap-within-permutation resampling (PB runs, e.g. $10,000 \cdot 10,000$ runs). While this may be a serious disadvantage, the main advantage of the permutation adjusted bMA method is that it does not require any influential decisions to be made regarding prior beliefs in the plausibility of PH. We will study the merits of the procedure in comparison to the procedure of Q&S and to the common two-stage approach by simulation.

Permutation-adjusted bMA (PAbMA)

For b in 1 to B:

1. Sample observations with replacement from the original data such that a "bootstrap data set" ($\tilde{D}_b$) of size equal to the original data (n) is obtained.
2. Apply the "common two-stage approach" to $\tilde{D}_b$.
3. Record the result of the p-value from the G&T test; denote this as $\tilde{p}(G\&T)_b$. Let $\mathbb{1}(G\&T)_b$ define rejection by the G&T test: $\mathbb{1}(G\&T)_b = \mathbb{1}(\tilde{p}(G\&T)_b \leq \alpha_{G\&T})$

We define:

$$\begin{aligned} w_0 &= \sum_{b=1}^{B} \mathbb{1}(\mathbb{1}(G\&T)_b = 0), \\ w_1 &= \sum_{b=1}^{B} \mathbb{1}(\mathbb{1}(G\&T)_b = 1), \\ p_{bMA} &= w_0 p_0 + w_1 p_1 \end{aligned}$$

where p_0 is the Cox PH model LRT p-value and p_1 is the PHTVC LRT p-value. For p in 1 to P:

1. Permute the covariate labels.
2. Repeat the 3 steps above for b in 1 to B, and obtain $\tilde{p}_{bMA_p}$.

We define a permutation bMA p-value, p_{PAbMA}, as equal to the proportion of permutation resamples of $\tilde{p}_{bMA_p}$ less than or equal to p_{bMA}. p_{PAbMA} determines the strength of evidence in favor of rejecting H_0;

$$p_{PAbMA} = \sum_{p=1}^{P} \frac{\mathbb{1}(\tilde{p}_{bMA_p} \leq p_{bMA})}{P}.$$

5.3 Simulation Study III: Q&S and PAbMA alternatives

In order to evaluate the merits of both the Q&S and PAbMA methods, we simulated data under the null, PH, and CS, identical to those considered in earlier simulation studies. As in earlier simulation studies, two censoring scenarios are considered: (1) right censoring at $t = 72$; and (ii) right censoring at $t = 72$ with additional independent exponential censoring generated such that approximately 15% of individuals are censored by $t = 24$. Given the substantial computational expense of the permutation-adjusted bMA, we will run the method with $P = 100$ and $B = 10$ determining bootstrap weights quite roughly. An additional consideration here is the trade-off between increasing B or P, and we expect that relatively low B but high P will be more effective. For implementing Q&S we run 500 bootstrap resamples with tuning parameter $\epsilon = 0.1$.

5.4 Results

Table 5.1 presents the proportion of runs for which the null hypothesis is rejected under each of the permutation tests. Let us first focus on the results from simulations with no additional censoring imposed. The PAbMA and Q&S procedures appear to have correct size, observed close to 0.05. The small differences from the target value can be attributed to Monte Carlo and resampling approximation error. When $\alpha_1 = \alpha_2$, the Q&S two-stage method is not quite as powerful as the common two-stage approach with TDP, (see Table 4.1). Contrast the proportion of significant tests under the alternatives of PH and CS when using Q&S versus TDP: 24.01 versus 22.48 (PH), 58.89 vs. 58.70 (CS), respectively.

In comparison to the PAbMA procedure, the Q&S, with $\alpha_1 = \alpha_2$, in substantially more powerful in detecting a treatment effect under the CS alternative and somewhat less powerful in detecting a treatment effect under the PH alternative.

We adjusted the α_1 level to equal 0.04, such that the power to detect a treatment effect under PH was approximately equal for both Q&S and PAbMA methods. With available power partitioned in this way, the Q&S method was found to be considerably more powerful than the PAbMA method in detecting a treatment effect under the CS alternative. Contrast 48.65 (Q&S, $\alpha_1 = 0.04$) with 39.19 (PAbMA). Similar conclusions can be made from the simulations run with additional censoring.

The PAbMA approach cannot be tuned to balance testing power between competing alternatives in the same way as for the Q&S method. However, further adjustments may be made by either resetting the significance level of the G&T test or employing an altogether different model selection tool, such as the BIC. It would be interesting to see the impact of such tuning on power.

Recall from the results of Section 4 that the CP adjustment has somewhat higher power when the true model is of PH; whereas under the alternative of CS, the TDP adjustment has substantially higher power. The results of Simulation Study III suggest that PAbMA, unlike the other permutation adjusted methods, achieves a "good balance": relatively high power to detect a treatment effect for both PH and CS alternatives. This can be attributed to the fact that, by averaging evidence over two models, the PAbMA method never relies entirely on the Cox PH model which may be incorrect.

The Q&S method also achieves "good balance", when the α_1 and α_2 levels are appropriately set. The Q&S method's superior performance can be attributed to the fact that, regardless of whether or not the PH assumption holds, the log-rank test (or equivalently the Cox PH model) may still be useful in detecting a treatment effect. While "significant" non-proportionality may suggest that the treatment effect changes over time, its impact on a test for *overall* treatment effect may be minimal in situations when the variance of $\beta(t)$ is small relative to its overall magnitude. Therefore, by first employing the log-rank test, regardless of how the underlying assumptions appear, the Q&S method fully exploits the power of the log-rank test. This reasoning is supported by the fact that among the 58.70 significant tests recorded by the Q&S method (with $\alpha_1 = \alpha_2$, no additional censoring), 7.70 were found significant by the first stage log-rank test.

No additional censoring												
	Common two-stage			**Q&S,** $\alpha_1 = \alpha_2$			**Q&S,** $\alpha_1 = 0.04$			**PAbMA**		
	Null	PH	CS	Null	PH	CS	Null	PH	CS	Null	PH	CS
Sign.	7.650	32.90	66.26	5.03	22.48	58.70	5.23	27.01	48.65	4.93	27.10	39.19
Not Sign.	92.350	67.10	33.74	94.97	77.51	41.29	94.77	72.99	51.35	95.07	72.90	60.81
Additional censoring												
	Common two-stage			**Q&S,** $\alpha_1 = \alpha_2$			**Q&S,** $\alpha_1 = 0.04$			**PAbMA**		
	Null	PH	CS	Null	PH	CS	Null	PH	CS	Null	PH	CS
Sign.	7.717	27.24	56.31	5.07	19.30	50.59	4.90	23.67	38.88	5.01	22.94	28.80
Not Sign.	92.283	77.76	43.69	94.93	80.70	49.41	95.10	76.33	61.12	94.99	77.06	71.20

Table 5.1: **Results of Simulation Study III.** Proportion of runs in which the null hypothesis is rejected out of 10,000 runs, each simulating 100 event-times from a Weibull under scenarios of no treatment effect (Null), treatment effect with proportional hazards (PH), and treatment effect with crossing survival curves (CS). Q&S run with 500 bootstrap replicates. PAbMA run with 10 bootsrap within 100 permutation runs.

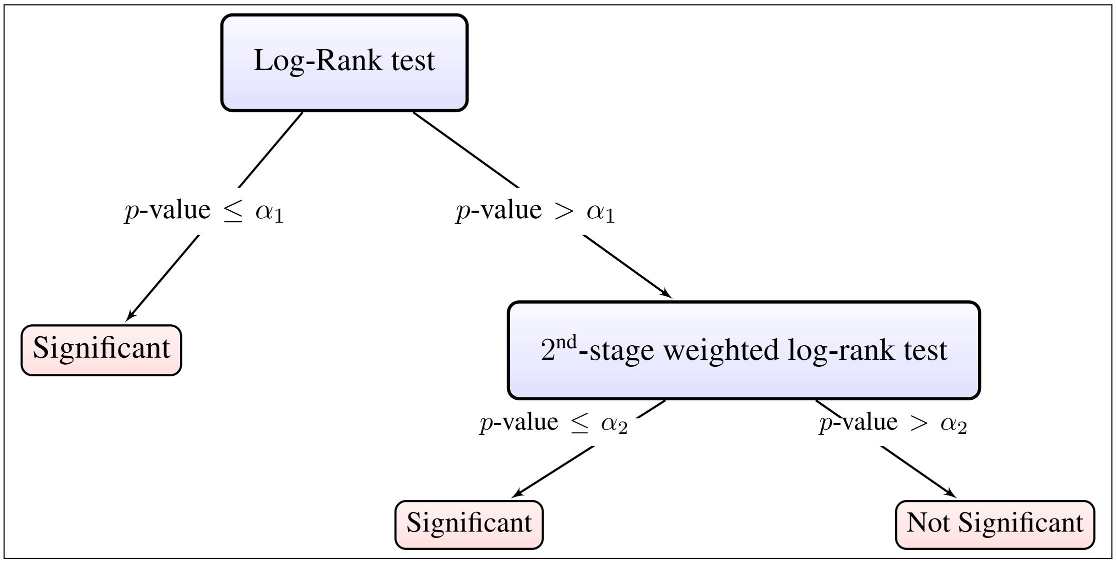

Figure 5.1: Qiu & Sheng (2008)'s two-stage approach.

6. Conclusion

While countless alternatives to the Cox PH model have been thoroughly studied in the survival analysis literature, the prevailing practice of the "common two-stage approach" requires an understanding not only of how these alternatives *compare* to the Cox model, but also of how they act *alongside* the Cox model. As we have demonstrated, employing certain alternatives within the common two-stage approach results in significant inflation of type-I error which should not be ignored.

While only a handful of alternatives were investigated, these serve as examples of the methods advocated for use in precisely the situation we examined; see for example Putter et al. (2005), Perperoglou et al. (2007) and Therneau and Grambsh (ch. 6.5, 2000). While the AFT appeared non-problematic, we stress that this does not provide a complete solution due to the fact that, when survival functions cross, the AFT may not be employed. It would be very useful to investigate the alternative of the accelerated hazards (AH) model, which has the ability to properly account for both non-PH and CS, see Chen & Wang (2000).

The importance of accounting for model uncertainty, which, as Shepherd (2007) writes "has been known for years, yet [remains] largely ignored", clearly deserves further attention in the survival analysis literature. While the issue has been partially addressed in the clinical trials literature, appropriate remedies are misunderstood; see for example Proschan & Waclawiw (2000) who write:

> Another difficult issue is that of multiple analyses methods. Clearly, it is not acceptable to specify four different primary analyses, and then declare a positive result if at least one is statistically significant. But what if the reason for the multiple analyses is that the assumptions underlying the originally intended method are suspect? [...] This is not something that can be corrected with a multiplicity adjustment because it was not preplanned. We believe that

> the onus is on the investigators to show that (1) a substantial violation of assumptions has occurred, and (2) several alternative methods that do not require the assumption demonstrate a significant treatment benefit.

As we have discovered –regardless of the number of "alternative methods" which demonstrate a significant effect– unacceptable bias will occur whenever sequential methods are not fully independent. A more appropriate recommendation would be to *preplan* for any possibility that model assumptions may be violated. In other words, one should establish, before obtaining any data, a hypothesis testing protocol which specifies, and accounts for, any alternative analysis to be employed in the event that model assumptions are suspect. Unfortunately, as we have seen with the permutation adjustment methods, common ways to account for non-independent sequential tests may substantially impact testing power.

Given the immense expense required to obtain data, anyone working to determine efficacy of a new treatment in a clinical trial will no doubt be reluctant to adopt any approach which substantially compromises power. When one is confident that the PH holds, a conditional permutation adjustment is not unreasonable, as power remains high. Alternatively, if one is confident that the data will exhibit non-PH and CS, a top-down permutation adjustment is not unreasonable. The problematic and most likely scenario is when one cannot, with a high degree of confidence, determine the nature of the treatment effect before obtaining the data. In such a situation, neither permutation adjustments were found to be favorable.

Recent developments such as the alternative two-stage approach of Q&S suggest the way forward. Unfortunately, with such an approach, one must partition available power based on prior beliefs in the plausibility of model assumptions. Despite this fact, the Q&S is an attractive alternative to the common two-stage approach as it has superior power to detect a treatment effect with an appropriately chosen partition. The newly proposed approach based on bootstrap model averaging also showed promise and suggests that creative solutions will be needed for addressing the challenges of model selection bias.

A. Appendix

A.1 Arguments for consistency of $\hat{\theta}_{bMA}$

Let $\{\mathcal{M}\}$ be the set of models under consideration and n be the number of observations in the sample of data, D, distributed under the unknown distribution F.

We recognize that in many scenarios, several different models may be "correct". (This is the situation for example with nested models, whereby a single "correct" model may be a special case of a second "correct" model.) Let $\{\mathcal{M}^*\}$ be the set of "correct" models and $\{\mathcal{M}^!\}$ be the set of "incorrect" models.

Two modest assumptions are required:

1. The chosen model selection criterion is such that:

$$\forall k : \text{model}_k \in \{\mathcal{M}^!\}, \quad \lim_{n\to\infty} Pr_F(M_k \text{ is selected } | \ D) = 0,$$

2. Correct models provide consistent estimates such that:

$$\forall k : \text{model}_k \in \{\mathcal{M}^*\}, \quad \lim_{n\to\infty} \hat{\theta}_k = \theta_k = \theta.$$

Then, given that the distribution of the bootstrap resampled data is the empirical distribution of the observed data ($\tilde{D}_b \sim \hat{F}$), and that $\sum_{k=1}^{K} w_k = 1$, we have:

$$\forall k : model\ k \in \{\mathcal{M}^!\}, \quad \lim_{n\to\infty} w_k = \lim_{n\to\infty} \frac{\sum_{b=1}^{B} \mathbb{1}(M_k \text{ is selected}|\tilde{D}_b(n))}{B} = 0$$

$$\Rightarrow \lim_{n\to\infty} \sum_{k\in\mathcal{M}^!} w_k\hat{\theta}_k = 0$$

$$\Rightarrow \lim_{n\to\infty} \hat{\theta}_{bMA} = \lim_{n\to\infty} \left(\sum_{k\in\mathcal{M}^*} w_k\hat{\theta}_k + \sum_{k\in\mathcal{M}^!} w_k\hat{\theta}_k \right) = \theta$$

A.2 Details of (5.9)

Recall that M_0 and M_1 are nested models and therefore, one may encounter two possibilities when H_1 is true:

1. H_1 is true, M_0 and M_1 are "correct".

2. H_1 is true, M_1 is "correct" and M_0 is "incorrect" .

Consider each separately,

Possibility 1 H_1 is true, M_0 and M_1 are "correct". Then we have that, $\forall \epsilon > 0$:

$$\begin{aligned}
\lim_{n\to\infty} Pr(p_0 \leq \epsilon) &= 1 \\
\lim_{n\to\infty} Pr(p_1 \leq \epsilon) &= 1 \\
\Rightarrow Pr(p_{bMA} \leq \epsilon) &= Pr((w_0p_0 + w_1p_1) \leq \epsilon) \\
&\geq Pr(2w_Mp_M \leq \epsilon) \quad , \quad (w_Mp_M = max(w_0p_0, w_0p_1)) \\
&= Pr(w_Mp_M \leq \epsilon/2) \\
&= 1 \qquad \text{(A.1)}
\end{aligned}$$

Possibility 2 H_1 is true, M_1 is "correct" and M_0 is "incorrect" . Then we have that, $\forall \epsilon > 0$:

$$\left.\begin{array}{r} \lim_{n\to\infty} Pr(p_1 \leq \epsilon) = 1 \\ \lim_{n\to\infty} Pr(w_0 \leq \epsilon) = 1 \end{array}\right\} \Rightarrow \lim_{n\to\infty} Pr(w_0p_0 + w_1p_1 \leq \epsilon) = 1 \qquad \text{(A.2)}$$

Bibliography

Abrahamowicz, M., Ciampi, A., and Ramsay, J. (1992). Nonparametric density estimation for censored survival data: Regression-spline approach. *Canadian Journal of Statistics* **20,** 171–185.

Altman, D. and Andersen, P. (1989). Bootstrap investigation of the stability of a cox regression model. *Statistics in Medicine* **8,** 771–783.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10,** 1100–1120.

Augustin, N., Sauerbrei, W., and Schumacher, M. (2005). The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling* **5,** 95.

Bangdiwala, S. (1989). The wald statistic in proportional hazards hypothesis testing. *Biometrical Journal* **31,** 203–211.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* **87,** 738–754.

Buchholz, A., Hollander, N., and Sauerbrei, W. (2008). On properties of predictors derived with a two-step bootstrap model averaging approach–a simulation study in the linear regression model. *Computational Statistics & Data Analysis* **52,** 2778–2793.

Buckland, S., Burnham, K., and Augustin, N. (1997). Model selection: an integral part of inference. *Biometrics* **53,** 603–618.

Chen, Y. and Wang, M. (2000). Analysis of accelerated hazards models. *Journal of the American Statistical Association* **95,** 608–618.

Co, C. (2010). Investigating the use of the accelerated hazards model for survival analysis. Master's thesis, Simon Fraser University.

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

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)* **57,** 45–97.

Grambsch, P. and Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81,** 515–526.

Harrell Jr, F., Lee, K., and Mark, D. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15,** 361–387.

Harrington, D. and Fleming, T. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69,** 553–566.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical science* **14,** 382–401.

Hollander, N., Augustin, N., and Sauerbrei, W. (2006). Investigation on the improvement of prediction by bootstrap model averaging. *Methods of Information in Medicine* **45,** 44–50.

Kraus, D. (2007). Data-driven smooth tests of the proportional hazards assumption. *Lifetime Data Analysis* **13,** 1–16.

Kvaløy, J. and Reiersølmoen Neef, L. (2004). Tests for the proportional intensity assumption based on the score process. *Lifetime Data Analysis* **10,** 139–157.

Logan, B., Klein, J., and Zhang, M. (2008). Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. *Biometrics* **64,** 733–740.

Mantel, N. and Stablein, D. (1988). The crossing hazard function problem. *Journal of the Royal Statistical Society. Series D (The Statistician)* **37,** 59–64.

Muggeo, V. and Tagliavia, M. (2010). A flexible approach to the crossing hazards problem. *Statistics in Medicine* **29,** 1947–1957.

Perperoglou, A., Keramopoullos, A., and van Houwelingen, H. (2007). Approaches in modelling long-term survival: an application to breast cancer. *Statistics in medicine* **26,** 2666–2685.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)* **135,** 185–207.

Proschan, M. and Waclawiw, M. (2000). Practical guidelines for multiplicity adjustment in clinical trials. *Controlled Clinical Trials* **21,** 527–539.

Putter, H., Sasako, M., Hartgrink, H., Van De Velde, C., and Van Houwelingen, J. (2005). Long-term survival with non-proportional hazards: results from the dutch gastric cancer trial. *Statistics in Medicine* **24,** 2807–2821.

Qiu, P. and Sheng, J. (2008). A two-stage procedure for comparing hazard rate functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70,** 191–208.

Routledge, R. (1997). P-values from permutation and f-tests. *Computational Statistics & Data Analysis* **24,** 379–386.

Sauerbrei, W. and Royston, P. (2007). Modelling to extract more information from clinical trials data: On some roles for the bootstrap. *Statistics in Medicine* **26,** 4989–5001.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: application to the cox regression model. *Statistics in Medicine* **11,** 2093–2109.

Shepherd, B. (2008). The cost of checking proportional hazards. *Statistics in Medicine* **27,** 1248–1260.

Therneau, T. and Grambsch, P. (2000). *Modeling survival data: extending the Cox model.* Springer Verlag.

Thompson, D. (2011). *Methods for Analyzing Survival Data with Non-Proportional Hazards and Complex Covariate Effects*. PhD thesis, Simon Fraser University.

Volinsky, C., Madigan, D., Raftery, A., and Kronmal, R. (1997). Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46,** 433–448.

Wei, L. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in Medicine* **11,** 1871–1879.

Wu, L. and Gilbert, P. (2002). Flexible weighted log-rank tests optimal for detecting early and/or late survival differences. *Biometrics* **58,** 997–1004.