Statistics
in Medicine

Research Article

Received 5 May 2012, Accepted 30 September 2013 Published online 18 October 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6021

The consequences of proportional hazards based model selection

H. Campbell*† and C.B. Dean

For testing the efficacy of a treatment in a clinical trial with survival data, the Cox proportional hazards (PH) model is the well-accepted, conventional tool. When using this model, one typically proceeds by confirming that the required PH assumption holds true. If the PH assumption fails to hold, there are many options available, proposed as alternatives to the Cox PH model. An important question which arises is whether the potential bias introduced by this 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 regard to power, of a simple resampling technique, the permutation adjustment, a natural recourse for addressing such challenges. We also consider a recently proposed two-stage testing strategy (2008) for ameliorating these effects. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: two-stage approach; proportional hazards; model selection bias

1. Introduction

When one wishes to relate event times to one or more covariates, a common and straightforward approach is to employ the Cox proportional hazards (PH) model [2]. The PH model is the most widely used model for 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 *hazard ratio*, as described by the PH model, would therefore be one-half. The standard model (with no time-dependent covariates) requires the assumption that the hazards ratio be constant over the entire follow-up period. In other words, the model assumes that the covariate (i.e., treatment) 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 [3] (chapter 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 result in erroneous inference. 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 [4, 5]. The most popular is a test attributed to Grambsch and Therneau [6], (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.

John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, U.K.
Correspondence to: Harlan Campbell, Journals Production Department, John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, U.K.
†*E-mail: hcampbel@sfu.ca*

Despite the fact that many adequate and flexible alternatives to the PH model are available, these may be less powerful and considerably more difficult to interpret for the average practitioner. Thus, typical practice would first try to fit the data to a PH model, and only consider alternatives, such as the popular accelerated failure time (AFT) model [7], in the event that a G&T type test indicates a lack of proportionality. Putter *et al.* [8] considered a similar situation and described the process of fitting an adjusted Cox PH model with time-varying coefficients to account for the non-PH. This approach seems practical and straightforward to implement; as discussed in Therneau and Grambsh [3] (chapter 6.2). However, the adjusted model has the disadvantage that one must choose a functional form describing how the effect of the treatment changes over time. Perperoglou *et al.* [9] considered the adjusted Cox PH model with time-varying coefficients and suggested using reduced rank regression to overcome the drawback of choosing a functional form for the treatment effect. Their approach attempts to remove some subjectivity 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. In addition to reviewing several methods for comparing hazard rates when there is evidence of non-PH, Liu *et al.* [10] proposed a new model in which a modified Box–Cox transformation is employed to cover a wide range of functional forms.

A situation that often occurs when hazards are found to be non-proportional, is that, the Kaplan–Meier (KM) survival curves appear to cross. Logan *et al.* [11] recommended a number of simple tests for comparing treatments in the presence of non-PH and 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. Although some of the tests considered are relatively insensitive to the choice of t_0, it is often the case that no prior knowledge regarding such a time point is available and the required pre-specification cannot be reasonably made. Mantel and Stablein [12], in a similar study, recognized this inconvenience—'admittedly, this is a difficult situation to envisage'— and considered 'letting the data suggest the crossing point'. Recently, methods for point and interval estimation of the crossing time point have been considered [13].

A common concern with all these approaches is that the uncertainty associated with the staged model selection procedure is not taken into account. Although the unfortunate practice of ignoring model uncertainty is not limited to the analysis of time-to-event data—Breiman [14] deemed 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 and Anderson [15], who considered using the bootstrap to validate the stability of a chosen Cox model, and Sauerbrei and Schumacher [16], who discussed the use of bootstrapping for variable selection. Yet, bootstrapping remains unpopular. Sauerbrei and Royston [17] drew 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.

Although resampling techniques may be useful for variable selection and model validation in prognostic studies [18]—where one may have dozens of possible covariates resulting in thousands of possible models—the question remains, *are they desirable for addressing the situation of checking and correcting for PH?* Shepherd [19] noted

> A highly cited paper recommends bootstrapping the process of examining and correcting possible violations of the PH assumption ([20]). However, I know of no analysis that has bootstrapped the process of checking the PH assumption.

Shepherd [19] drew attention to this issue and studied 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 of no treatment effect is left unaddressed.

In this work, we wish to determine the consequences, in terms of obtaining correct type-I error, 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 treatment effect in the event that the PH assumption is suspect. We investigate the merits of different two-stage testing strategies by simulation and discuss the results and implications in Section 2. In Section 3, we investigate the possibility of using permutation adjustment to overcome the bias encountered and consider the effects such an adjustment may have on power. We also consider the advantages and disadvantages of an alternative two-stage procedure proposed by Qiu and Sheng [1].

2. The common two-stage approach under the null

2.1. *The common two-stage approach*

As discussed in Section 1, prevailing practice for the analysis of time-to-event data would first try to fit the data to a Cox PH model and only consider alternatives in the event that a G&T type test indicates a lack of proportionality. We designate this procedure the 'common two-stage approach'. The most general hypothesis test under consideration is

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

where X is a binary covariate and t is elapsed time; $h(t)$ and $S(t)$ are the hazard and survivor functions, respectively.

Having fit the data to a Cox PH model, one typically tests for significant covariate effect by either a Wald test or a likelihood ratio test (LRT). Bangdiwala [21] summarized the difference between the Wald test and LRT. Although the likelihood ratio and Wald statistics are asymptotically equivalent and most often result in similar, if not identical conclusions, 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 paper.

One of the main advantages of the Cox model is that, in addition to evaluating the statistical significance of a covariate effect, one obtains a point estimate of the magnitude of this effect as well as a corresponding confidence interval. However, it must be cautioned that the validity of this inference rests on the PH assumption. A common test to verify this assumption is the G&T test, a χ^2 test based on how the scaled Schoenfeld residuals, s_k^*, compare with a function of time, $g(t)$. Therneau and Grambsh [3] noted that 'for long-tailed survival distributions [...] log(t) is often a good choice'. However, setting $g(t)$ as the left-continuous version of the KM curve (without covariates) 'tends to spread the residuals s_k^* fairly evenly across the plot from left to right, avoiding potential problems with outliers'. For the purpose of our simulation studies, because we are simulating event-times from a long-tailed distribution, we will set $g(t) = \log(t)$.

In the event that a G&T type test indicates a lack of proportionality, one of several possible alternative methods is typically employed. Three common 'second-stage' methods will be investigated in simulations throughout this paper: (i) the adjusted Cox PH model with time-varying coefficients; (ii) the AFT model; and (iii) the post-cross-point log-rank test. A brief review of these follows.

The adjusted Cox PH model with time-varying coefficients considers the following model:

$$h(t|X_i) = h_0(t)\exp(X_i\beta(t)) = h_0(t)\exp[\beta_0 X_i + \beta_1(X_i f(t))], \tag{2}$$

for $i = 1, \ldots, n$ observations and where $f(t)$ incorporates the time-varying treatment effect. As mentioned in the Section 1, a disadvantage with this model is that one must choose $f(t)$ from many possible functional forms. Three common choices for $f(t)$ are $\log(t)$, $\sqrt{t}$, and t. Putter *et al.* [8] recommended considering goodness of fit with a plot of the scaled Schoenfeld residuals to guide one's choice. Because using a subjective graphical evaluation like this is not possible in a large simulation study such as ours, we will instead select $f(t)$ based on the BIC measure of goodness of fit.

The parametric AFT model, although not appropriate if there are any crossovers in the survival function, is a straightforward and rather interpretable alternative to the PH model [22]. The AFT model can be thought of as a linear model for the logarithm of the survival time:

$$\log(T_i) = X_i^{'}\beta + \sigma\epsilon_i, \tag{3}$$

where β and σ are parameters to be estimated. One must specify 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; as such, we will use this parametric choice in all simulation studies.

The common log-rank (also known as the Mantel–Cox) statistic forms the basis for a post-cross-point log-rank test considered by Logan *et al.* [11]. Assume that a time point t_0 can be 'pre-specified' (before one obtains the data), such that survival beyond t_0 is considered long-term with t_0 (the 'cross-point') chosen such that the survival curves are likely to cross prior to that time point, if at all. Then a *post-*t_0 log-rank test statistic can be defined as

$$Z_{LR}(t_0) = \frac{X_{LR}(t_0)}{\hat{\sigma}_{LR}(t_0)} \sim N(0,1),$$

where

$$X_{LR}(t_0) = \sum_{t_j > t_0} \frac{Y_{1j}Y_{0j}}{Y_j}\left(\frac{d_{1j}}{Y_{1j}} - \frac{d_{0j}}{Y_{0j}}\right), \qquad \hat{\sigma}^2_{LR} = \sum_{t_j > t_0} \frac{Y_{1j}Y_{0j}}{Y_j^2}\left(\frac{Y_j - d_{1j}}{Y_j - 1}\right) d_j,$$

where Y_{kj}, Y_j denote the number at risk at t_j in the k^{th} treatment 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. Several other log-rank type tests have been proposed. These include the Peto–Peto [23] 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 Harrington and Fleming [24], and a weighted log-rank test that emphasizes early and/or late differences studied by Wu and Gilbert [25]. These have been found to be powerful in detecting many non-PH alternatives.

It is worth noting that doing an additional test or fitting an additional model once PH has been rejected is somewhat redundant, because rejection of PH implies that the hazards cannot be equal and thus invalidates the null hypothesis (Eq. 1). In practice, however, this is a moot point. Because the motivation behind conducting a G&T type test is not to determine treatment effect, but rather to validate an assumption at the basis of the Cox PH model, it is not considered sufficient for establishing significant treatment effect. Further research is needed to determine whether or not such a practice (using only the Cox PH model and a G&T type test) has any merit and what type of adjustment would be needed to account for the two tests being performed in conjunction.

2.2. *Simulation study I*

In order to assess the potential bias of the common two-stage approach, we investigated its performance under the null model by simulation study. Consider the probability density function of the Weibull distribution: $f(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} exp\left(\left(-\frac{t}{\lambda}\right)^{k}\right)$. We simulated event-times from the Weibull distribution with parameters $k = 0.6$ and $\lambda = 83$, such that the probability of an event exceeding t =72 is about 0.4. While we will focus our discussion on results obtained from simulations from this distribution, we also performed the identical simulation study with event-times simulated from two alternative Weibull distributions with (i) $k = 1.2$ and $\lambda = 77.5$ and with (ii) $k = 0.2$ and $\lambda = 112$. These additional experiments will provide support to the results.

Event-times were simulated in each experiment such that half are attributed to 'a control group' ($X = 0$) and half to a 'treatment group' ($X = 1$). Three censoring scenarios are considered: (i) right censoring at $t = 72$, when survival probability equals 0.4; (ii) right censoring at $t = 72$ with additional independent exponential censoring generated such that approximately 15% of individuals are censored by $t = 24$; and (iii) right censoring at $t = 72$ with additional independent exponential censoring generated such that approximately 30% of individuals are censored by $t = 24$. The settings used were set to mimic the simulation studies of Logan *et al.* [11].

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

(1) **AdjCox**: the adjusted Cox PH model with time varying coefficients and $f(t) = \log(t)$, as described by Putter *et al.* [8];
(2) **AdjCoxChoice**: the adjusted Cox PH model with time varying coefficients and $f(t)$ chosen as 'best' fit (best BIC value) among $f(t) = \log(t)$, $f(t) = \sqrt{t}$, and $f(t) = t$, as described by Putter *et al.* [8];
(3) **LogRank**(t_0): the log-rank test with $t_0 = 24$, as described by Logan *et al.* [11]; and
(4) **AFT**: the AFT model, as described by Wei [7].

For every model fit, the LRT is employed to calculate a χ^2 statistic and p-value. Sample size was set to be either 50, 100, or 200 event-times. In total, nine datasets of 100,000 event-times and censoring times were generated for each combination sample size and censoring scenario. For each dataset, 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 will compare these numbers to those expected if the sequential tests were fully independent.

2.3. *Results of simulation study I*

Table I summarizes the results of 'simulation study I', listing the percentage of runs for which the null hypothesis was rejected ('significant'), for each of the scenarios under investigation. Figure 1 displays

Table I. Simulation Study I. For each of 100,000 simulation runs, the Grambsch and Therneau test determined whether the data should be analyzed by the Cox proportional hazards model (if $p_{G\&T} > 0.05$) or by the chosen alternative (if $p_{G\&T} \leqslant 0.05$).

			Cox PH ($p_{G\&T} > 0.05$)		Alternative ($p_{G\&T} \leqslant 0.05$)		Two-stage approach	
	Sample size	%censored at at $t = 24$	Significant ($p_{ph} \leqslant 0.05$)	Not Significant ($p_{ph} > 0.05$)	Significant ($p_{alt} \leqslant 0.05$)	Not Significant ($p_{alt} > 0.05$)	Significant ($p \leqslant 0.05$)	Not Significant ($p > 0.05$)
AdjCox	50	0	4.64	90.70	2.50	2.16	7.14	92.86
	50	15	4.73	90.63	2.53	2.11	7.26	92.74
	50	30	4.93	90.49	2.58	2.01	7.51	92.50
	100	0	4.72	90.46	2.37	2.45	7.09	92.91
	100	15	4.85	90.37	2.34	2.45	7.19	92.82
	100	30	4.89	90.24	2.36	2.51	7.25	92.75
	200	0	4.83	90.21	2.31	2.64	7.14	92.85
	200	15	4.89	90.1	2.23	2.79	7.12	92.89
	200	30	4.86	90.28	2.27	2.59	7.13	92.87
AdjCoxChoice	50	0	4.64	90.70	2.69	1.97	7.33	92.67
	50	15	4.73	90.63	2.72	1.92	7.45	92.55
	50	30	4.93	90.49	2.78	1.81	7.71	92.30
	100	0	4.72	90.46	2.58	2.24	7.30	92.70
	100	15	4.85	90.37	2.5	2.28	7.35	92.65
	100	30	4.89	90.24	2.56	2.32	7.45	92.56
	200	0	4.83	90.21	2.48	2.48	7.31	92.69
	200	15	4.89	90.1	2.4	2.61	7.29	92.71
	200	30	4.86	90.28	2.44	2.42	7.30	92.70
LogRank(t_0)	50	0	4.64	90.70	1.06	3.60	5.70	94.30
	50	15	4.73	90.63	1.08	3.56	5.81	94.19
	50	30	4.93	90.49	0.93	3.65	5.86	94.14
	100	0	4.72	90.46	1.07	3.75	5.79	94.21
	100	15	4.85	90.37	1.01	3.78	5.86	94.15
	100	30	4.89	90.24	0.97	3.9	5.86	94.14
	200	0	4.83	90.21	1.01	3.94	5.84	94.15
	200	15	4.89	90.1	0.99	4.02	5.88	94.12
	200	30	4.86	90.28	0.91	3.95	5.77	94.23
AFT	50	0	4.64	90.70	0.22	4.44	4.86	95.14
	50	15	4.73	90.63	0.24	4.41	4.97	95.04
	50	30	4.93	90.49	0.26	4.32	5.19	94.81
	100	0	4.72	90.46	0.28	4.54	5.00	95.00
	100	15	4.85	90.37	0.24	4.55	5.09	94.92
	100	30	4.89	90.24	0.22	4.66	5.11	94.90
	200	0	4.83	90.21	0.23	4.73	5.06	94.94
	200	15	4.89	90.1	0.24	4.77	5.13	94.87
	200	30	4.86	90.28	0.26	4.6	5.12	94.88
Independent tests			4.75	90.25	0.25	4.75	5.00	95.00

The 'Cox PH' columns display the percentage of runs for which $p_{G\&T} > 0.05$ and subsequently, the Cox PH model determined the treatment effect to be significant ($p_{ph} \leqslant 0.05$) and not significant ($p_{ph} > 0.05$). The 'alternative' columns display the percentage of runs for which $p_{G\&T} \leqslant 0.05$ and subsequently, the alternative test determined the treatment effect to be significant ($p_{alt} \leqslant 0.05$) and not significant ($p_{alt} > 0.05$). Finally, the 'two-stage approach' columns display the the percentage of runs for which the treatment effect was determined to be significant ($p_{G\&T} > 0.05$ and $p_{ph} \leqslant 0.05$, or $p_{G\&T} \leqslant 0.05$ and $p_{alt} \leqslant 0.05$) and not significant ($p_{G\&T} > 0.05$ and $p_{ph} > 0.05$, or $p_{G\&T} \leqslant 0.05$ and $p_{alt} > 0.05$). Finally, 'independent tests' show, for comparison, percentages expected if sequential tests were fully independent.

PH, proportional hazards; AFT, accelerated failure time.

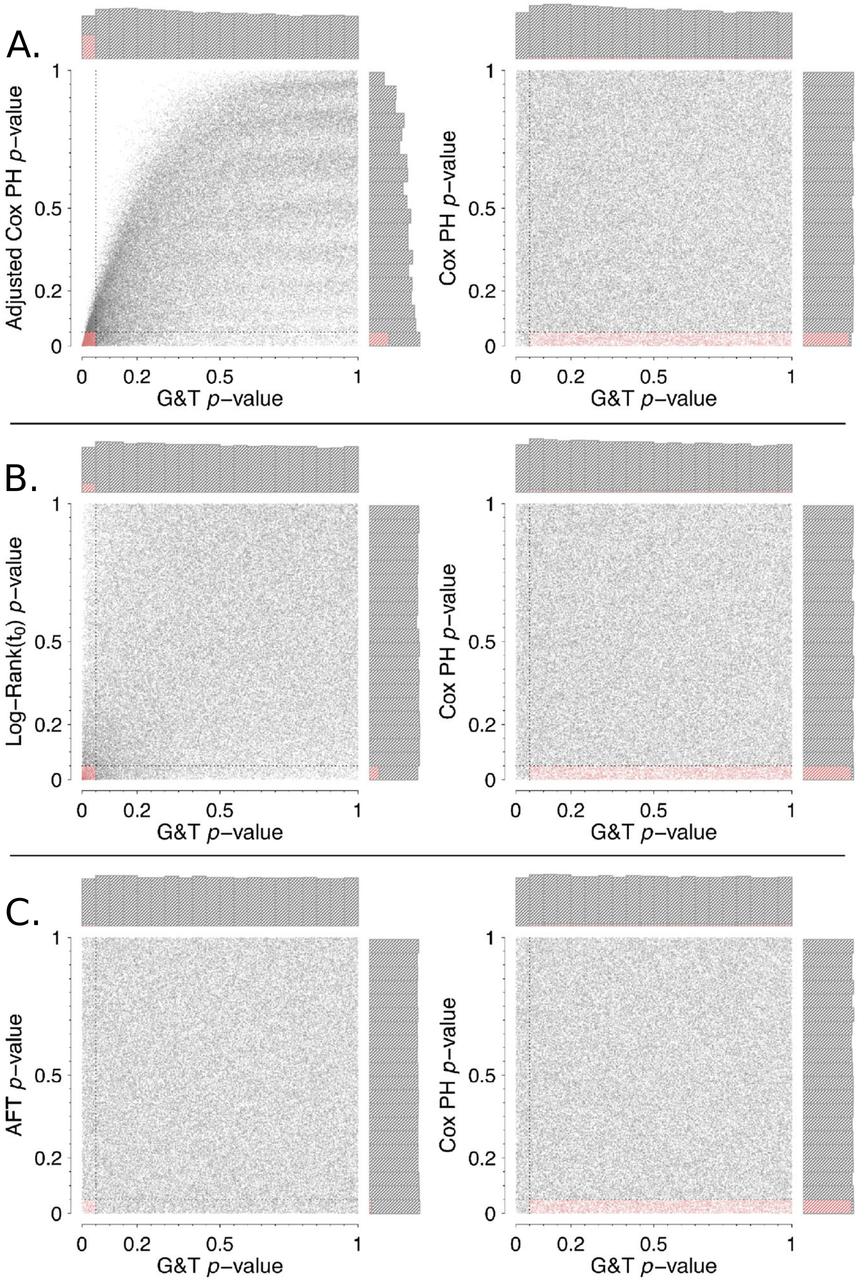


Figure 1. Scatterplots show the joint distribution of p-values recorded in simulation study I, with red points indicating significant outcomes (p-values $\leqslant 0.05$). Histograms show marginal distributions. **A.** With the Adjusted Cox proportional hazards model as second-stage alternative and with sample size of 50 and 0% additional censoring. **B.** With the log-rank(t_0) test as second-stage alternative and with sample size of 100 and 15% additional censoring at $t = 24$. **C.** With accelerated failure time as second-stage alternative and with sample size of 200 and 30% additional censoring at $t = 24$.

three sets of side by side scatterplots with marginal histograms, illustrating the joint distributions of G&T p-values with the p-values obtained from the Cox PH model and the second-stage alternative model. The top row of scatterplots (Figure 1(A)) corresponds to the results summarized in the first row of Table I: with sample size of 50, no additional censoring and the Adjusted Cox PH model with $f(t) = \log(t)$ as the alternative. Figure 1(B) corresponds to the scenario with sample size of 100, 15% additional censoring at $t = 24$ and the log-rank(t_0) test as the second-stage alternative. Finally, Figure 1(C) corresponds to the scenario with sample size of 200, 30% additional censoring at $t = 24$ and the AFT. The results of simulation study I can be summarized as follows:

(1) **Small sample size leads to bias of the G&T test.** Consider the results for when there is no additional censoring. With a sample size of 50, the G&T test failed to reject the null hypothesis of PH for 95.34% (= 4.64% + 90.70%) of the 100,000 simulation runs. Although this is only slightly higher than the expected 95.00%, a careful inspection of a histogram of the G&T p-values (Figure 1(A), top histograms) suggests that this cannot be entirely attributed to sampling error: the G&T p-values do not appear to be uniformly distributed, with fewer values less than or equal to 0.05 than expected. Recall that the G&T test statistic is only asymptotically χ^2_1. Therefore, this may be largely due to the small sample size of 50. In fact, the non-uniformity is less apparent in the simulations with sample size of 100 (95.18% rejection rate) and in simulations with sample size of 200 (95.04% rejection rate). Indeed the equivalent histogram in Figure 1(B) (sample size 100) is substantially less skewed and in Figure 1(C) (sample size of 200) the G&T p-values appear essentially uniform. Results with additional censoring show similar results. Finally, it is worth noting that in additional simulations, not presented here, in which the G&T $g(t)$ was set as the left-continuous version of the KM curve, the small sample bias was also present though to a lesser degree.

(2) **Correlation between the G&T test and the adjusted Cox PH model results in substantial bias.** Once again, consider the results for when there is no additional censoring and a sample size of 50. Among the 4.66% (= 2.50% + 2.16%) of runs for which $p_{G\&T} \leqslant 0.05$, more than half were found to be significant by the adjusted Cox PH model $\left(\frac{2.50\%}{4.66\%} \approx \frac{10.7}{20}\right)$. This represents a rejection rate more than 10 times the expected rate under independent tests $\left(\frac{0.25\%}{5.00\%} = \frac{1}{20}\right)$. Clearly, the adjusted Cox PH model and the G&T tests are not independent. This is confirmed by inspecting the left scatterplot of Figure 1(A). The correlation between of the adjusted Cox PH model and the G&T test results in a type-I error rate of 7.14%, well above the desired 5.00%. The bias appears to diminish with larger sample sizes, albeit not dramatically $\left(\frac{2.37\%}{4.82\%} \approx \frac{9.8}{20}\right.$, with n=100 and $\frac{2.37\%}{4.82\%} \approx \frac{9.3}{20}$, with $\left. n = 200 \right)$. Results with additional censoring show similar results. With the data-driven choice of the $f(t)$ (AdjCoxChoice), the bias is even greater: type-I error rates are recorded between 7.29% and 7.71%.

(3) **Correlation between the G&T test and the Log-rank(t_0) test results in modest bias.** The Log-rank(t_0) test showed bias similar to the adjusted Cox PH model, although to a much lesser degree: type-I error rates ranged from 5.70% and 5.88%. Figure 1(B) left scatterplot shows that the p-values from the log-rank(t_0) test are indeed correlated with the G&T p-values (Pearson correlation = 0.16).

(4) **No bias is observed with the AFT model.** The two-stage approach with the AFT model as the alternative to the Cox PH model appears relatively safe from bias, with type-I error rates recorded between 4.86% and 5.19%. The lack of bias can be attributed to the fact that the p-values obtained from the AFT model are not correlated with those obtained from the G&T test, see Figure 1(C). Rather, the p-values obtained from the AFT model are highly correlated to those obtained from the Cox PH model (Pearson correlation ≈ 0.99). This indicates that for the vast majority of the 100,000 simulation runs, the AFT model is essentially the same test for significance as the Cox PH model. Thus, the main difference between the two methods is not whether a treatment effect is deemed significant. Rather, the difference is in the type of treatment effect inferred: with the AFT model, it need not be constant over the follow-up time. Although this is an encouraging result, recall that the AFT model requires that event-times follow a pre-specified distribution and that, in this simulation study, the Weibull distribution was correctly specified. Simulation studies by Li *et al.* [26] show that model misspecification with the AFT model can lead to inflated type-I error and a substantial reduction in power, particularly when the pre-‘miss’-specified distribution is the Weibull distribution.

The results obtained from the additional simulation studies with event-times simulated from two different Weibull distributions (identical in all other aspects) show very similar outcomes, see Table II.

3. Correcting for bias by incorporating resampling methods

3.1. Three resampling methods

Permutation adjustment is a popular resampling method for obtaining non-biased p-values [27]. In order to overcome the bias encountered by the two-stage procedure, 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 (TDP) adjustment repeats

Table II. Simulation study I: additional simulations.

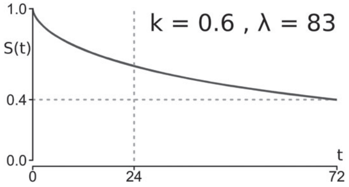
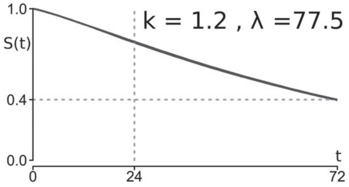
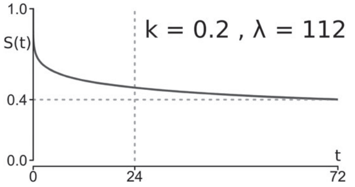
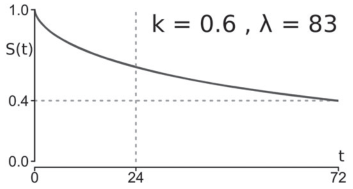
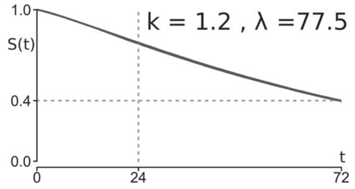
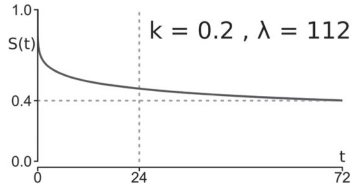
			Distribution 1			Distribution 2			Distribution 3		
											
				Significant by:			Significant by:			Significant by:	
	Sample size	% censored at $t = 24$	Cox PH	Alternative	Two-stage	Cox PH	Alternative	Two-stage	Cox PH	Alternative	Two-stage
AdjCox	50	0	4.64	2.50	7.14	4.63	2.79	7.42	4.63	1.44	6.07
	50	15	4.73	2.53	7.26	4.70	2.79	7.49	4.70	1.42	6.12
	50	30	4.93	2.58	7.51	5.02	2.88	7.90	4.96	1.39	6.35
	100	0	4.72	2.37	7.09	4.73	2.65	7.38	4.73	1.34	6.07
	100	15	4.85	2.34	7.19	4.78	2.65	7.43	4.79	1.23	6.02
	100	30	4.89	2.36	7.25	4.76	2.80	7.56	4.90	1.25	6.15
	200	0	4.83	2.31	7.14	4.83	2.62	7.45	4.83	1.25	6.08
	200	15	4.89	2.23	7.12	4.78	2.56	7.34	4.87	1.17	6.04
	200	30	4.86	2.27	7.13	4.81	2.53	7.34	4.78	1.10	5.88
AdjCoxChoice	50	0	4.64	2.69	7.33	4.63	2.99	7.62	4.63	1.60	6.23
	50	15	4.73	2.72	7.45	4.70	3.01	7.71	4.70	1.59	6.29
	50	30	4.93	2.78	7.71	5.02	3.08	8.10	4.96	1.57	6.53
	100	0	4.72	2.58	7.30	4.73	2.90	7.63	4.73	1.49	6.22
	100	15	4.85	2.50	7.35	4.78	2.90	7.68	4.79	1.36	6.15
	100	30	4.89	2.56	7.45	4.76	3.05	7.81	4.90	1.40	6.30
	200	0	4.83	2.48	7.31	4.83	2.90	7.73	4.83	1.37	6.20
	200	15	4.89	2.40	7.29	4.78	2.84	7.62	4.87	1.30	6.17
	200	30	4.86	2.44	7.30	4.81	2.80	7.61	4.78	1.25	6.03
LogRank(t_0)	50	0	4.64	1.06	5.70	4.63	0.94	5.57	4.63	0.55	5.18
	50	15	4.73	1.08	5.81	4.70	1.09	5.79	4.70	0.34	5.04
	50	30	4.93	0.93	5.86	5.02	1.27	6.29	4.96	0.19	5.15
	100	0	4.72	1.07	5.79	4.73	0.93	5.66	4.73	0.67	5.40
	100	15	4.85	1.01	5.86	4.78	1.06	5.84	4.79	0.46	5.25
	100	30	4.89	0.97	5.86	4.76	1.17	5.93	4.9	0.23	5.13
	200	0	4.83	1.01	5.84	4.83	0.92	5.75	4.83	0.64	5.47
	200	15	4.89	0.99	5.88	4.78	0.99	5.77	4.87	0.52	5.39
	200	30	4.86	0.91	5.77	4.81	1.05	5.86	4.78	0.34	5.12

Table II. *Continued.*

				Distribution 1			Distribution 2			Distribution 3	
											
				Significant by:			Significant by:			Significant by:	
	Sample size	% censored at $t = 24$	Cox PH	Alternative	Two-stage	Cox PH	Alternative	Two-stage	Cox PH	Alternative	Two-stage
AFT	50	0	4.64	0.22	4.86	4.63	0.22	4.85	4.63	0.22	4.85
	50	15	4.73	0.24	4.97	4.70	0.24	4.94	4.70	0.24	4.94
	50	30	4.93	0.26	5.19	5.02	0.28	5.30	4.96	0.22	5.18
	100	0	4.72	0.28	5.00	4.73	0.27	5.00	4.73	0.27	5.00
	100	15	4.85	0.24	5.09	4.78	0.28	5.06	4.79	0.22	5.01
	100	30	4.89	0.22	5.11	4.76	0.24	5.00	4.90	0.23	5.13
	200	0	4.83	0.23	5.06	4.83	0.23	5.06	4.83	0.23	5.06
	200	15	4.89	0.24	5.13	4.78	0.25	5.03	4.87	0.27	5.14
	200	30	4.86	0.26	5.12	4.81	0.25	5.06	4.78	0.24	5.02
Independent tests			4.75	0.25	5.00	4.75	0.25	5.00	4.75	0.25	5.00

Results from simulation study I ('distribution 1') and two identical simulation studies with event-times simulated from alternative Weibull distributions: with $k = 1.2$ and $\lambda = 77.5$ ('distribution 2'), and with $k = 0.2$ and $\lambda = 112$ ('distribution 3'). 'Cox PH' column displays percentage of runs for which the treatment effect was determined to be significant by Cox PH model ($p_{G\&T} > 0.05$ and $p_{ph} \leqslant 0.05$). 'alternative' column displays percentage of runs for which the treatment effect was determined to be significant by the alternative test ($p_{G\&T} \leqslant 0.05$ and $p_{alt} > 0.05$). 'Two-stage 'column displays the percentage of runs for which the treatment effect was determined to be significant ($p_{G\&T} > 0.05$ and $p_{ph} \leqslant 0.05$, or $p_{G\&T} \leqslant 0.05$ and $p_{alt} \leqslant 0.05$). Finally, 'independent tests' show, for comparison, percentages expected if sequential tests were fully independent.
PH, proportional hazards; AFT, accelerated failure time.

the entire two-stage procedure. In these two methods, significance of treatment effect is determined by a permutation-adjusted p-value, regardless of whether the assumption of PH is rejected. In addition, we consider an alternative two-stage approach, proposed by Qui and Sheng [1], which incorporates a bootstrap adjustment. In this method, no test for PH is administered and the interpretation of a significant treatment effect must be carefully considered. What is more, the method does not provide a point estimate for the magnitude of the treatment effect nor does it allow for the possibility of including secondary additional covariates in the inference. The three methodologies are described in the three algorithms in the succeeding text.

TDP adjustment

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

(1) Permute the covariate (treatment) labels of the original data, to yield permuted assignments to each of the responses; denote the permuted data as $\tilde{D}_j$.
(2) Apply the common two-stage approach to $\tilde{D}_j$, testing for PH with $\alpha_{G\&T}$ significance level.
(3) Obtain a p-value for a test of no covariate (treatment) effect; denote this as $\tilde{p}_j$.

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

$$p_{TDP} = \sum_{j=1}^{J} \frac{\mathbb{1}\left(\tilde{p}_j \leqslant p\right)}{J}. \tag{4}$$

Calibration: Modifying the significance level of the G&T test, $\alpha_{G\&T}$, may allow one to calibrate the method to favor power under different alternatives.

Conditional Permutation (CP) adjustment

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

(1) Permute the covariate (treatment) labels of the original data, to yield permuted 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: $\mathbb{1}(G\&T)_j = \mathbb{1}(\tilde{p}(G\&T)_j \leqslant \alpha_{G\&T})$.

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 on the basis of an analysis of the original data under the common two-stage approach, and $\mathbb{1}(G\&T)$ be an indicator for the rejection of the PH assumption (by the G&T test with significance level $\alpha_{G\&T}$). Then,

$$p_{CP} = \sum_{j:\mathbb{1}(G\&T)_j = \mathbb{1}(G\&T)} \frac{\mathbb{1}(\tilde{p}_j \leqslant p)}{\#\{j : \mathbb{1}(G\&T)_j = \mathbb{1}(G\&T)\}}. \tag{5}$$

Calibration: Modifying the significance level of the G&T test, $\alpha_{G\&T}$, may allow one to calibrate the method to favor power under different alternatives.

3.2. *Simulation study II*

In order to assess the performance of the three methods considered, we simulated data under the null model, under a model of PH, and a model of non-PH, as we wish to evaluate both size and power of the procedures. Event-times are simulated from the Weibull distribution as in simulation study I. For both the PH and and non-PH scenarios, parameters of the Weibull were set such that the odds ratio of survival at $t = 72$ between groups is 1.75. A crossing of survival curves in the non-PH scenario occurs at $t = 24$. In total, 100 event-times were simulated in each experiment, half of which are attributed to a 'control group' ($X = 0$) and half to a 'treatment group' ($X = 1$). The three sample sizes and three censoring scenarios employed in simulation study I are once again considered. One thousand 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 adjusted Cox PH model with time-varying coefficients with $f(t) = \log(t)$, and with $\alpha_{G\&T} = 0.05$ and $\alpha_{G\&T} = 0.02$. For implementing the Q&S method, we ran 500 bootstrap resamples with tuning parameter $\epsilon = 0.1$. We proceeded with equally distributed significance levels for each stage, $\alpha_1 = \alpha_2$, (**Q&S**$_1$), as well as unequally distributed levels, $\alpha_1 = 0.04$, (**Q&S**$_2$).

Qiu & Sheng's (2008) two-stage approach (Q&S)

(1) Test for evidence against the null by unweighted log-rank test. (The unweighted log-rank test is equivalent to a LRT test for covariate effect using the Cox PH model except in cases when equal event-times have been observed. In this event, the tests are asymptotically equivalent.) 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 employing the Q&S-weighted log-rank test. The weights are negative prior to the supposed crossing point of the hazards and positive afterwards and derived such that the test statistic is asymptotically independent of the first stage log-rank test statistic. Because 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 the 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 [1] for details. The critical value of the null distribution of this *maximal* test statistic is estimated by bootstrapping the second stage. If significant at the α_2 level, H_0 is rejected. Otherwise, one fails to reject H_0. Because of the asymptotic independence of the tests at the first and second stages, Q&S define the overall significance level α as

$$\begin{aligned}\alpha &= \alpha_1 + Pr_{H_0}(\text{reject in stage 2}|\text{fail to reject in stage 1})(1-\alpha_1)\\ &= \alpha_1 + \alpha_2(1-\alpha_1).\end{aligned} \tag{6}$$

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

$$p - \text{value} = \begin{cases} p_1 & \text{, if} p_1 \leqslant \alpha_1, \\ \alpha_1 + p_2(1-\alpha_1) & \text{, otherwise.}\end{cases} \tag{7}$$

Calibration: Modifying the significance levels of the first and second stages, α_1 and α_2, may allow one to calibrate the method to favor power under different alternatives.

3.3. *Results of simulation study II*

Table III summarizes the results of 'simulation study II', listing the percentage of runs for which the null hypothesis was rejected ('significant'), in each of the scenarios under investigation. The results can be summarized as follows:

(1) **All methods of adjustment show correct type-I error.** With no adjustment, type-I error was recorded between 7.46% and 7.77%. All re-sampling methods of adjustment recorded type-I error much closer to the desired level of 5%, between 4.79% and 5.87%.

(2) **Permutation methods show high yet inconsistent power.** Comparing the two permutation-adjusted methods, with $\alpha_{G\&T} = 0.05$, CP adjustment has somewhat higher power when the true model is PH (a significant result was established about 10% more often). On the other hand, under the alternative of non-PH, the TDP adjustment has substantially higher power (in 5/9 cases, a significant result was established more than twice as often). Therefore, we cannot conclude that one method consistently outperforms the other. Calibration by means of adjusting the $\alpha_{G\&T}$ level from

Table III. Simulation study II.

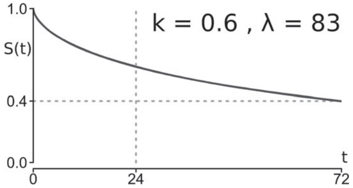
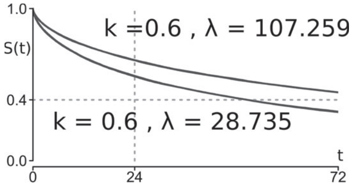
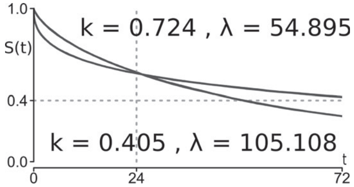
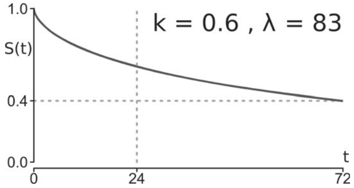
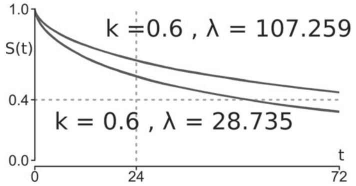
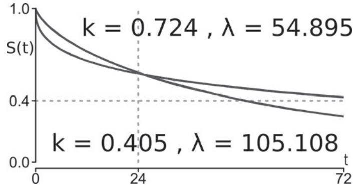
			Null		Proportional hazards		Non-proportional hazards	
								
	Sample size	% censored at $t = 24$	Significant		Significant		Significant	
No adjustment	50	0	7.62		19.66		36.96	
	50	15	7.67		17.16		29.25	
	50	30	7.77		16.14		24.70	
	100	0	7.54		31.68		64.60	
	100	15	7.48		27.46		54.88	
	100	30	7.46		23.86		47.23	
	200	0	7.53		53.17		92.48	
	200	15	7.51		46.73		86.86	
	200	30	7.64		39.94		78.77	
			$\alpha_{G\&T} = 0.05$	$\alpha_{G\&T} = 0.02$	$\alpha_{G\&T} = 0.05$	$\alpha_{G\&T} = 0.02$	$\alpha_{G\&T} = 0.05$	$\alpha_{G\&T} = 0.02$
Conditional permutation	50	0	5.25	5.01	16.97	17.18	13.73	11.70
	50	15	4.96	4.81	14.28	14.28	9.41	7.99
	50	30	4.96	4.80	13.12	13.25	7.86	6.80
	100	0	4.95	4.99	28.97	29.30	30.89	24.40
	100	15	5.15	5.10	24.41	24.59	21.97	16.15
	100	30	5.01	5.17	20.91	21.10	16.42	11.68
	200	0	5.20	5.14	50.86	51.29	70.33	60.01
	200	15	5.31	5.25	44.10	44.39	56.53	45.33
	200	30	5.02	5.13	37.42	37.86	44.23	33.75

Table III. *Continued.*

	Sample size	% censored at $t = 24$	Null		Proportional hazards		Non-proportional hazards	
								
			Significant		Significant		Significant	
			$\alpha_{G\&T} = 0.05$	$\alpha_{G\&T} = 0.02$	$\alpha_{G\&T} =0.05$	$\alpha_{G\&T} =0.02$	$\alpha_{G\&T} =0.05$	$\alpha_{G\&T} =0.02$
Top-down permutation	50	0	5.24	5.36	14.84	15.44	30.34	27.85
	50	15	4.94	4.86	12.54	12.82	23.44	21.28
	50	30	4.98	5.05	11.84	11.94	19.40	16.15
	100	0	5.14	5.14	25.51	25.94	58.44	55.88
	100	15	5.02	4.87	21.78	21.99	47.80	46.40
	100	30	5.06	4.88	18.29	18.90	40.10	37.33
	200	0	5.30	5.32	46.62	47.17	90.34	89.44
	200	15	5.02	5.01	39.82	40.56	82.86	82.10
	200	30	5.06	5.03	32.91	34.00	73.49	72.61
			$\alpha_1 = \alpha_2$	$\alpha_1 =0.04$	$\alpha_1 = \alpha_2$	$\alpha_1 =0.04$	$\alpha_1 = \alpha_2$	$\alpha_1 =0.04$
Q&S	50	0	5.87	5.62	14.16	16.35	34.31	26.74
	50	15	5.51	5.19	12.13	14.12	27.21	20.20
	50	30	5.01	5.04	10.55	12.58	20.79	14.59
	100	0	5.40	5.18	23.07	27.18	57.70	48.50
	100	15	5.46	5.08	19.72	23.19	51.02	39.59
	100	30	5.19	5.00	16.35	19.63	42.15	31.31
	200	0	5.04	5.15	42.40	48.31	86.13	78.47
	200	15	4.79	5.14	35.50	41.68	80.18	70.42
	200	30	4.92	4.97	29.20	34.96	73.15	61.24

For each sample size, censoring scenario, and distribution of event-times, the table displays the percentage of runs (out of 10,000) for which the two-stage approach determined the treatment effect to be significant. The ‘no adjustment’ rows display the results when no adjustment to control bias was employed.

0.05 to 0.02 yields only an incremental increase in power under the alternative of PH while resulting in a substantial reduction of power under the non-PH alternative.

(3) **Properly tuned, the Q&S method achieves high and consistent power.** When $\alpha_1 = \alpha_2$, the Q&S two-stage method is not quite as powerful as the common two-stage approach with TDP. When available power is partitioned such that $\alpha_1 = 0.04$, the Q&S method is somewhat more powerful than the TDP in detecting a treatment effect under the PH alternative (a significant result was established about 4–9% more often), whereas somewhat less powerful under the non-PH alternative. With regard to how the Q&S method compares to the CP method: greater power is achieved under PH with CP, whereas under non-PH, the CP method is substantially less powerful. Despite the fact that, given either the PH or non-PH alternative, the Q&S method fails to achieve greater power than the best performing permutation method, it arguably achieves a better balance when α_1 and α_2 are appropriately tuned. When calibrated such that $\alpha_1 = \alpha_2$, the method appears less favorable. This suggests that proper tuning of α_1 and α_2, based on a priori expectations about the likelihood of the alternative, is crucial for maximizing power.

4. Conclusion

Although 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* with 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 a significant inflation of type-I error, which should not be ignored.

Although only a handful of alternatives were investigated, these serve as examples of the methods advocated for use in precisely the situation we examined [3, 8, 9]. Although the AFT appeared non-problematic, we stress that there are two main drawbacks to consider with this method. First, the AFT cannot accommodate data with crossing survivor curves. Second, an appropriate distribution for the event-times must be specified prior to fitting the model. It would be very useful to investigate the alternative of the accelerated hazards model, which has the ability to properly account for both non-PH and crossing survival survivor curves [28].

The importance of accounting for model uncertainty, which, as Shepherd [19] wrote ‘has been known for years, yet [remains] largely ignored’, clearly deserves further attention in the survival analysis literature. Although the issue has been somewhat addressed in the clinical trials literature, appropriate remedies are misunderstood; see for example Proschan and Waclawiw [29] who suggest that, when assumptions underlying the original method are found to be suspect, investigators demonstrate significant treatment effect by several alternative methods. 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. For even if the observed data are not found to be in violation of the PH assumption, a modification (of some form or another) must be made to the significance test that accounts for what would have been carried out had PH been suspect. The simulation studies presented in this paper demonstrate that this is a necessary action to control type-I error. 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. However, if researchers are a priori convinced that the data will display PH, the protocol might simply specify that the only significance test considered will be that of the Cox PH significance test for treatment effect. If one is confident that the data will exhibit non-PH, a TDP adjustment is not unreasonable. Alternatively, the protocol might specify that only one significance test, robust to departures from PH, will be will be considered.

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 entirely favorable. Recent developments such as the alternative two-stage approach of Q&S, while not without their limitations, suggest the way forward.

References

1. Qiu P, Sheng J. A two-stage procedure for comparing hazard rate functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2008; **70**(1):191–208. DOI: 10.1111/j.1467-9868.2007.00622.x.
2. Cox D. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 1972; **34**(2):187–220.
3. Therneau T, Grambsch P. *Modeling Survival Data: Extending the Cox Model*. Springer Verlag: New York, 2000.
4. Kraus D. Data-driven smooth tests of the proportional hazards assumption. *Lifetime Data Analysis* 2007; **13**(1):1–16. DOI: 10.1007/s10985-006-9027-8.
5. Kvaløy J, Reiersølmoen Neef L. Tests for the proportional intensity assumption based on the score process. *Lifetime Data Analysis* 2004; **10**(2):139–157.
6. Grambsch P, Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; **81**(3):515–526.
7. Wei L. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* 1992; **11**(14-15):1871–1879. DOI: 10.1002/sim.4780111409.
8. Putter H, Sasako M, Hartgrink H, Van De Velde C, Van Houwelingen J. Long-term survival with non-proportional hazards: results from the Dutch Gastric Cancer Trial. *Statistics in Medicine* 2005; **24**(18):2807–2821. DOI: 10.1002/sim.2143.
9. Perperoglou A, Keramopoullos A, van Houwelingen H. Approaches in modelling long-term survival: an application to breast cancer. *Statistics in Medicine* 2007; **26**(13):2666–2685. DOI: 10.1002/sim.2729.
10. Liu K, Qiu P, Sheng J. Comparing two crossing hazard rates by Cox proportional hazards modelling. *Statistics in Medicine* 2007; **26**(2):375–391. DOI: 10.1002/sim.2544.
11. Logan B, Klein J, Zhang M. Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. *Biometrics* 2008; **64**(3):733–740. DOI: 10.1111/j.1541-0420.2007.00975.x.
12. Mantel N, Stablein DM. The crossing hazard function problem. *Journal of the Royal Statistical Society. Series D (The Statistician)* 1988; **37**(1):59–64.
13. Cheng M, Qiu P, Tan X, Tu D. Confidence intervals for the first crossing point of two hazard functions. *Lifetime Data Analysis* 2009; **15**(4):441–454.
14. Breiman L. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* 1992; **87**(419):738–754. DOI: 10.1080/01621459.1992.10475276.
15. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine* 1989; **8**(7):771–783. DOI: 10.1002/sim.4780080702.
16. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine* 1992; **11**(16):2093–2109. DOI: 10.1002/sim.4780111607.
17. Sauerbrei W, Royston P. Modelling to extract more information from clinical trials data: on some roles for the bootstrap. *Statistics in Medicine* 2007; **26**(27):4989–5001. DOI: 10.1002/sim.2954.
18. Augustin N, Sauerbrei W, Schumacher M. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling* 2005; **5**(2):95. DOI: 10.1191/1471082X05st089oa.
19. Shepherd BE. The cost of checking proportional hazards. *Statistics in Medicine* 2008; **27**(8):1248–1260. DOI: 10.1002/sim.3020.
20. Harrell Jr F, Lee K, Mark D. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**(4):361–387.
21. Bangdiwala SI. The wald statistic in proportional hazards hypothesis testing. *Biometrical Journal* 1989; **31**(2):203–211. DOI: 10.1002/bimj.4710310209.
22. Chiou S, Kim J, Yan J. Semiparametric multivariate accelerated failure time model with generalized estimating equations, 2012. *arXiv preprint arXiv:1204.0285*.
23. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)* 1972; **135**:185–207.
24. Harrington D, Fleming T. A class of rank test procedures for censored survival data. *Biometrika* 1982; **69**(3):553–566. DOI: 10.1093/biomet/69.3.553.
25. Wu L, Gilbert P. Flexible weighted log-rank tests optimal for detecting early and/or late survival differences. *Biometrics* 2002; **58**(4):997–1004. DOI: 10.1111/j.0006-341X.2002.00997.x.
26. Li Y, Klein J, Moeschberger M. Effects of model misspecification in estimating covariate effects in survival analysis for small sample sizes. *Computational Statistics & Data Analysis* 1996; **22**(2):177–192. DOI: 10.1016/0167-9473(96)88029-7.
27. Routledge R. P-values from permutation and F-tests. *Computational Statistics & Data Analysis* 1997; **24**(4):379–386. DOI: 10.1016/S0167-9473(96)00073-4.
28. Chen Y, Wang M. Analysis of accelerated hazards models. *Journal of the American Statistical Association* 2000; **95**(450):608–618. DOI: 10.1080/01621459.2000.10474236.
29. Proschan M, Waclawiw M. Practical guidelines for multiplicity adjustment in clinical trials. *Controlled Clinical Trials* 2000; **21**(6):527–539. DOI: 10.1016/S0197-2456(00)00106-9.