2007

A Comparison of Graphical Methods for Assessing the Proportional Hazards Assumptions in the Cox Model

Inger Persson

Harry J. Khamis
Wright State University - Main Campus, harry.khamis@wright.edu

Festschrift in honor of
Distinguished Professor Mir Masoom Ali
On the occasion of his retirement
May 18-19, 2007, pages 23-43

A Comparison of Graphical Methods for Assessing the Proportional Hazards Assumption in the Cox Model

Inger Persson, AstraZeneca R&D, Södertälje, SWEDEN

Harry Khamis, Statistical Consulting Center, Wright State University, Dayton, Ohio 45435, USA

ABSTRACT Six graphical procedures to check the assumption of proportional hazards for the Cox model are described and compared. A new way of comparing the graphical procedures using a Kolmogorov-Smirnov like maximum deviation criterion for rejection is derived for each procedure. The procedures are evaluated in a simulation study under proportional hazards and five different forms of nonproportional hazards: (1) increasing hazards, (2) decreasing hazards, (3) crossing hazards, (4) diverging hazards, and (5) nonmonotonic hazards. The procedures are compared in the two-sample case corresponding to two groups with different hazard functions. None of the procedures under consideration require partitioning of the survival time axis. Results indicate that the Arjas plot, a plot of estimated cumulative hazard versus number of failures, is superior to the other procedures under almost every form of nonproportional hazards, especially crossing and nonmonotonic hazards. For increasing hazards, the smoothed plot of the ratio of log cumulative baseline hazard rates versus time or the smoothed plot of scaled Schoenfeld residuals versus time perform the best. The Andersen plot performs very poorly for increasing, decreasing, and diverging hazards.

1. INTRODUCTION

The relation between the distribution of event times and time-invariant covariates or risk factors $\mathbf{z}$ ($\mathbf{z}$ is a $p \times 1$ vector) can be described in terms of a model according to Cox (1972), in which the hazard rate at time t for an individual is

$$\lambda(t,\mathbf{z}) = \lambda_0(t)e^{\boldsymbol{\beta}'\mathbf{z}}, \qquad (1.1)$$

where $\lambda_0(t)$ is the baseline hazard rate, an unknown (arbitrary) function giving the hazard rate for the standard set of conditions $\mathbf{z} = \mathbf{0}$, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters. The factor $e^{\boldsymbol{\beta}'\mathbf{z}}$ describes the hazard for an individual with covariates $\mathbf{z}$ relative to the hazard at standard conditions $\mathbf{z} = \mathbf{0}$.

The ratio of the hazard functions for two individuals with covariate values $\mathbf{z}$ and $\mathbf{z}^*$ is $\frac{\lambda(t \mid z)}{\lambda(t \mid z^*)} = e^{\beta'(z-z^*)}$, an expression that does not depend on t. Thus, the Cox model in (1.1) is only valid for data consistent with the assumption of proportional hazards.

Since the validity of the Cox regression analysis based on the model in (1.1) relies on the assumption of proportionality of the hazard rates of individuals with distinct covariate values, it is important to be able to reliably determine if the assumption is plausible. This can be done graphically or numerically. A partial review of numerous graphical and analytical methods for checking the adequacy of Cox models was given by Lin and Wei (1991). Some authors recommend using numerical tests for such determinations (e.g., Hosmer and Lemeshow, 1999, p. 207). However, others recommend graphical procedures arguing that the proportional hazards assumption only approximates the correct model for a covariate, and that any formal statistical test, based on a large enough sample size, will reject the null hypothesis of proportionality (Klein and Moeschberger, 1997, p. 354). A comprehensive comparative study of numerical procedures is given elsewhere (Persson, 2002). This paper focuses on the effectiveness of graphical procedures. In section 2, six graphical methods for determining the plausibility of the proportional hazards assumption are described. In section 3, the results of a comparative simulation study are presented. A discussion is given in section 4, an example is presented in section 5, and conclusions are given in section 6.

2. GRAPHICAL METHODS COMPARED

Hess (1995) describes eight graphical methods for detecting violations of the proportional hazards

assumption and demonstrated each on three authentic data sets. Five of those methods are described in sections 2.1 – 2.5 below. The methods not included in this paper are (1) methods that require a partitioning of the time axis, which introduces a certain degree of arbitrariness into the procedure, leading to different conclusions depending on the partition used, or (2) methods that do not allow a comparison with other methods through the use of a maximum deviation criterion proposed in this paper. Section 2.6 describes an additional graphical method not included in the article by Hess (1995), the lesser used Arjas plot (Arjas, 1988). These six graphical methods are compared through a simulation study.

Comparing graphical methods can be somewhat arbitrary since there are no clear guidelines for how to interpret the plots. The conclusions are highly dependent on the subjectivity of the viewer. However, to make it possible to compare the results of the different methods, a criterion for rejection is derived for each method individually using measures described in sections 2.1 – 2.6. In each case, a Kolmogorov-Smirnov like maximum deviation criterion is used. See Lin et al. (1993) for an illustration of this approach.

2.1 Method 1: Plot of Survival Curves Based on the Cox Model and Kaplan-Meier Estimates for Each Group

The survival function, $S(t)$, is related to the cumulative hazards function, $H(t)$, as follows:

$$S(t) = e^{-H(t)} = \exp\left\{-H_0(t)^{\exp(\beta' z)}\right\} = S_0(t)^{\exp(\beta' z)}, \qquad (2.1.1)$$

where $H_0(t)$ is the cumulative baseline hazard and $S_0(t)$ is the baseline survival function. Breslow (1974) gives an estimate for the cumulative baseline hazard based on the Cox proportional hazards model,

$$\hat{H}_0(t) = \sum_{t_i \le t} \frac{\delta_i}{\sum_{j \in R_i} e^{\hat{\beta} z_j}}, \qquad (2.1.2)$$

where δ_i is the event indicator for the i^{th} individual and R_i is the risk set at time t_i, i.e., the set of individuals still under study at a time just prior to t_i. Kalbfleish and Prentice (1980) and Link (1984) provide additional estimates for the cumulative baseline hazard. The baseline survival function can be written $S_0(t) = e^{-H_0(t)}$. Thus, an estimate of the baseline survival function based on the Cox model is given by

$$\hat{S}_0(t) = e^{-\hat{H}_0(t)}. \qquad (2.1.3)$$

It is possible to assess violations of the assumption of proportional hazards by comparing survival estimates based on the Cox model with estimates computed independently of the model, such as the Kaplan-Meier product-limit estimate for each group (Kaplan and Meier, 1958), defined by

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \le t} \left[1 - \frac{d_i}{Y_i}\right] & \text{if } t \ge t_1 \end{cases} \qquad (2.1.4)$$

where d_i is the observed number of events at time t_i and Y_i is the number at risk at time t_i (i.e., the number of individuals who are alive at time t_i or experience the event of interest at time t_i). See Kleinbaum (1996), Chapter 3 for a discussion of the quantitative comparison of the Kaplan-Meier and Cox regression estimates.

Clear departures of the two estimates provide evidence against the assumption of proportional hazards. Figure 2.1.1 shows an example of plots of survival curves based on the Cox model along with Kaplan-Meier estimates for each of two groups of patients.

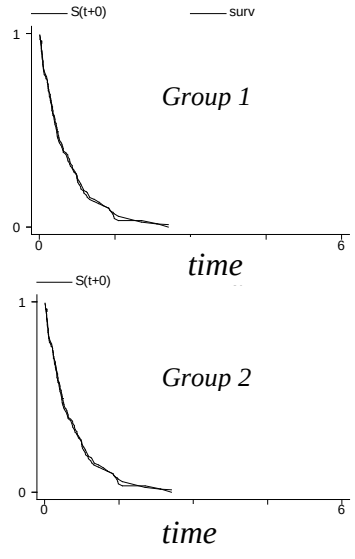


Figure 2.1.1 Survival curves based on the Cox model along withKaplan-Meier estimates for each of two groups of patients

The maximum absolute difference between the curves is used to establish possible deviations from the assumption of proportional hazards. This criterion is used in the Kolmogorov-Smirnov test for goodness-of-fit of two cumulative distribution functions (see, e.g., Sokal and Rohlf, 1995). The larger the absolute difference between the curves, the stronger the indication of violations of the proportional hazards assumption. Let $\text{Diff}_{\text{max1}}$ denote the maximum absolute difference between the curves, then the hazards are proportional if $\text{Diff}_{\text{max1}} = 0$. The larger the value of $\text{Diff}_{\text{max1}}$, the stronger the evidence of nonproportionality. Figure 2.1.2 shows the distribution of 10,000 generated $\text{Diff}_{\text{max1}}$ values under proportional hazards.

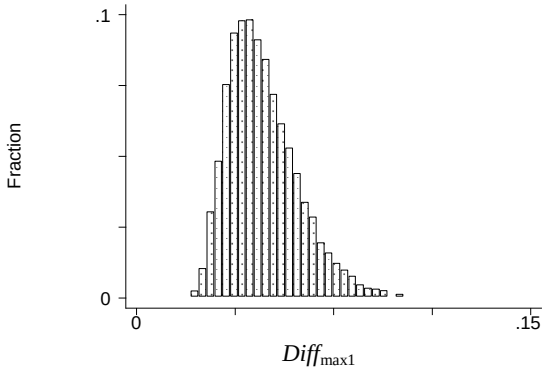


Figure 2.1.2 10,000 $\text{Diff}_{\text{max1}}$ values generated under proportional hazards

This distribution can be used to establish a criterion for determining that the proportional hazards assumption is not plausible. We use the 95th percentile as such a criterion, namely, that value of x for which
$P[\text{Diff}_{\text{max1}} > x] = 0.05$. So, to check the assumption of proportional hazards, $\text{Diff}_{\text{max1}}$ is calculated and it is concluded that the hazards are not proportional if $\text{Diff}_{\text{max1}}$ exceeds x.

2.2 Method 2: Plot of Cumulative Baseline Hazards in Different Groups

Another method to graphically check the assumption of proportional hazards is based on the estimated cumulative baseline hazard rate, namely, the Andersen (1982) plot.

Let $\hat{H}_g(t)$ be the estimated cumulative baseline hazard rate in stratum g, g = 1, 2, …, K. Plot, for all t, $\hat{H}_g(t)$ for g = 2, …, K. If the proportional hazards assumption is true, then these curves should be straight lines through the origin. Figure 2.2.1 shows an example of a plot of the cumulative baseline hazards in two groups of patients.

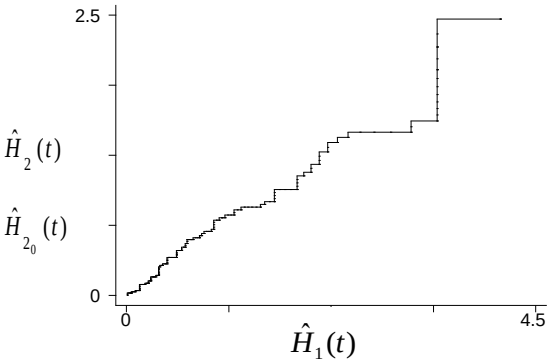


Figure 2.2.1 Estimated cumulative baseline hazard rate in group 2 versus group 1

To determine if this curve follows a straight line through the origin, estimation of a linear regression with no intercept of $\hat{H}_2(t)$ on $\hat{H}_1(t)$ is proposed. Let $\text{Diff}_{\text{max2}}$ denote the maximum absolute difference between $\hat{H}_2(t)$ and the estimated (fitted) values from the regression. Figure 2.2.2 shows the distribution of 10,000 generated $\text{Diff}_{\text{max2}}$ values under proportional hazards.

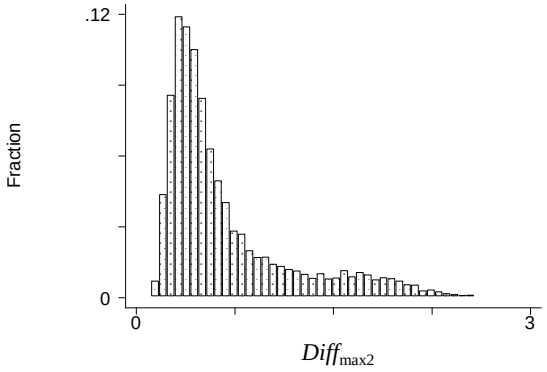


Figure 2.2.2 10,000 $\text{Diff}_{\text{max2}}$ values generated under proportional hazards

The assumption of proportional hazards is concluded to be implausible if the calculated value of $\text{Diff}_{\text{max2}}$ exceeds the 95th percentile of this distribution.

2.3 Method 3: Plot of the Difference of the Log Cumulative Baseline Hazard Versus Time

Schumacher (1990) suggested plotting $\hat{\gamma}(t)$ versus t, where

$$\hat{\gamma}(t) = \log\left[\hat{H}_1(t)\right] - \log\left[\hat{H}_0(t)\right]. \qquad (2.3.1)$$

Under proportional hazards this plot is constant over t, centered around the estimated log hazard ratio $\hat{\beta}$. Figure 2.3.1 shows an example of a plot of the difference of the log cumulative baseline hazard versus time.

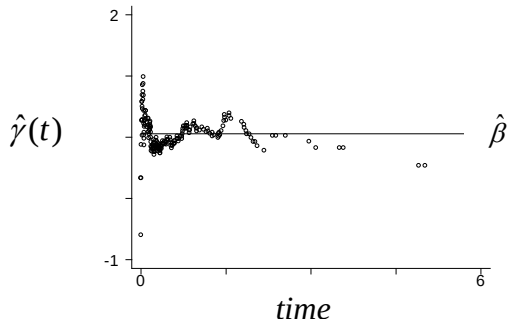


Figure 2.3.1 Plot of $\hat{\gamma}(t)$ versus time

Let $\text{Diff}_{\max 3}$ denote the maximum absolute difference between $\hat{\gamma}(t)$ and $\hat{\beta}$. Figure 2.3.2 shows the distribution of 10,000 generated $\text{Diff}_{\max 3}$ values under proportional hazards.

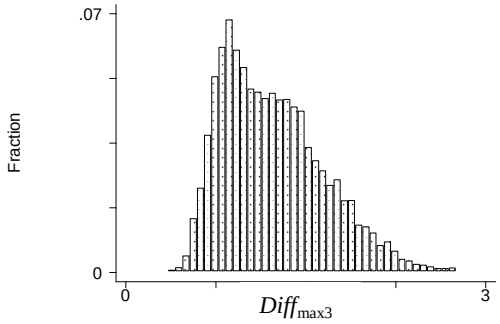


Figure 2.3.2 10,000 $\text{Diff}_{\max 3}$ values generated under proportional hazards

The hazards are concluded to be nonproportional if the calculated value of $\text{Diff}_{\max 3}$ exceeds the 95th percentile of this distribution.

2.4 Method 4: Smoothed Plot of the Ratio of Log Cumulative Baseline Hazard Rates Versus Time

Smoothing helps describe the pattern of dependence, thus making it easier to check the constancy of $\hat{\gamma}(t)$ when plotting it against t as described in subsection 2.3. The choice of smoothing technique is usually not very important as long as the smoother (1) is sensitive to local rather than global features of the data and (2) has an appropriate number of degrees of freedom (Hastie and Tibshirani, 1990). For example, LOWESS (locally-weighted scatter plot smoothing) employs iterated weighted least squares with a robustness feature that identifies and down-weights outliers in successive smoothings. Figure 2.4.1 shows an example of a smoothed plot of the difference of the log cumulative baseline hazard versus time using LOWESS.

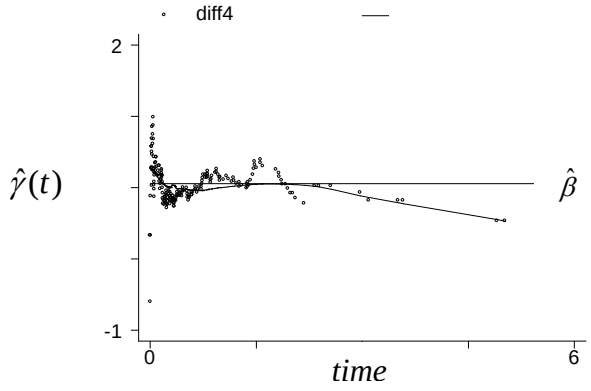


Figure 2.4.1 Smoothed plot of $\hat{\gamma}(t)$ versus time

Let $\text{Diff}_{\max 4}$ denote the maximum absolute difference between the smoothed values of $\hat{\gamma}(t)$ and $\hat{\beta}$. Figure 2.4.2 shows the distribution of 10,000 generated $\text{Diff}_{\max 4}$ values under proportional hazards.

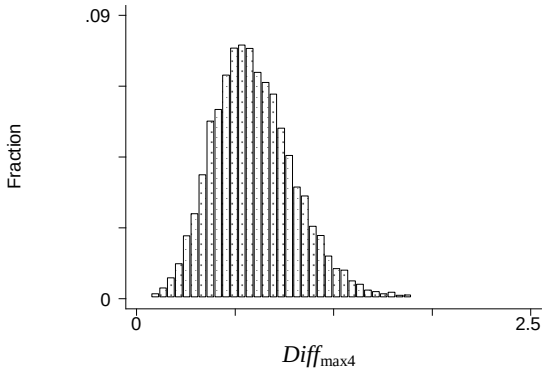


Figure 2.4.2 10,000 $Diff_{max4}$ values generated under proportional hazards

The hazards are concluded to be nonproportional if the calculated value of $Diff_{max4}$ exceeds the 95th percentile of this distribution.

2.5 Method 5: Smoothed Plot of Scaled Schoenfeld Residuals Versus Time

Schoenfeld (1980) defined partial residuals for the Cox model that do not depend on time, so that the j^{th} residual can be plotted against t_j to detect violations of the proportional hazards assumption, where j indexes individuals (j = 1, 2, …, n). The Schoenfeld residuals are defined as

$$\mathbf{r}_i(\boldsymbol{\beta}) = \mathbf{z}_{(i)} - M(\boldsymbol{\beta}, t_i), \qquad (2.5.1)$$

where $\mathbf{z}_{(i)}$ is the covariate vector of the subject with an event at time t_i, where i indexes event times (i = 1, 2, …, D), and $M(\boldsymbol{\beta}, t_i)$ is the conditional weighted mean of the covariate vector at time t_i as described in Persson (2002), section 2.2. Grambsch and Therneau (1994) describe a scale adjustment for Schoenfeld's residuals,

$$\hat{r}_i^*(\hat{\beta}) = \hat{\beta} + \hat{r}_i d\left[\hat{V}(\hat{\beta})\right]^{-1}, \qquad (2.5.2)$$

where $\hat{\beta}$ is the maximum partial likelihood estimate under proportional hazards, $\hat{V}(\hat{\beta})$ is the estimated variance of $\hat{\beta}$, and d is the total number of events where individuals from both groups remain at risk. For a binary covariate coded 0 or 1, a plot of r_i^* versus t_i yields two horizontal bands of residuals. If the proportional hazards assumption holds, then the residuals center around $\hat{\beta}$. Smoothing improves the interpretability of the residual plots, so LOWESS is applied. Figure 2.5.1 shows an example of a smoothed plot of the scaled Schoenfeld residuals versus time.

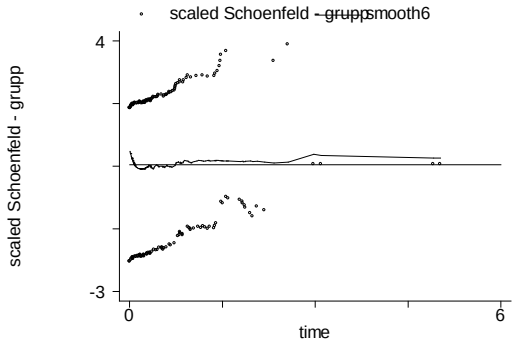


Figure 2.5.1 Smoothed plot of scaled Schoenfeld residuals versus time

Let $Diff_{max5}$ denote the maximum absolute difference between the smoothed residuals and $\hat{\beta}$. Figure 2.5.2 shows the distribution of 10,000 generated $Diff_{max5}$ values under proportional hazards.

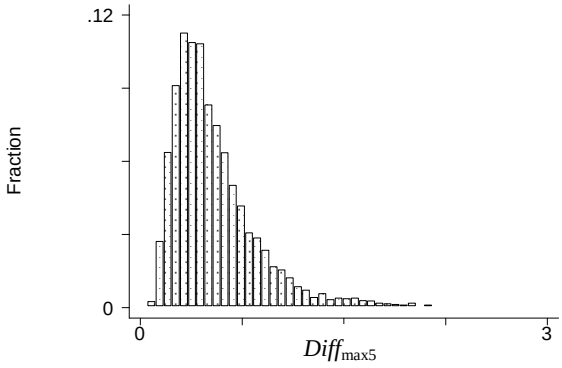


Figure 2.5.2 10,000 $Diff_{max5}$ values generated under proportional hazards

The hazards are concluded to be nonproportional if the calculated value of $Diff_{max5}$ exceeds the 95th percentile of this distribution.

2.6 Method 6: Plot of the Estimated Cumulative Hazard Versus Number of Failures

Arjas (1988) suggested a plot of the estimated cumulative hazard versus the number of failures in each stratum (Arjas plot). See Section

VII.3.4 of Andersen et al. (1993) for the development, of which the following is a brief summary. Consider the differences

$$N_h(t) - \int_0^t \sum_{h(i)=h} p_i(u,\beta_0)dN(u), h = 1,...,k ,$$

where $N_h(t) = \sum_{h(i)=h} N_i(t)$, N is the process counting the observed failures, and $\beta_0 = (\beta_1^0,...,\beta_{p-1}^0,\beta_{p+1}^0,...,\beta_{p+k-1}^0)$ is the true parameter vector. These differences are (local) martingales. Therefore, plots of

$$\int_0^{X_{(m)}^h} \sum_{h(i)=h} p_i(u,\hat{\beta})dN(u), m = 1,...,N_h(\tau), h = 1,...,k$$

versus m, where $X_{(m)}^h, m = 1,...,N_h(\tau)$ are the ordered jump times in stratum h, should be approximately straight lines with unit slope.

This plot is a Total Time on Test plot for the residuals $\hat{r}_i$. Tests for the proportional hazards model based on these residuals were briefly discussed by Arjas (1988) and Arjas and Haara (1988).

Figure 2.6.1 shows an example of an Arjas plot of estimated cumulative hazard versus number of failures in each stratum (here the strata are the two groups).

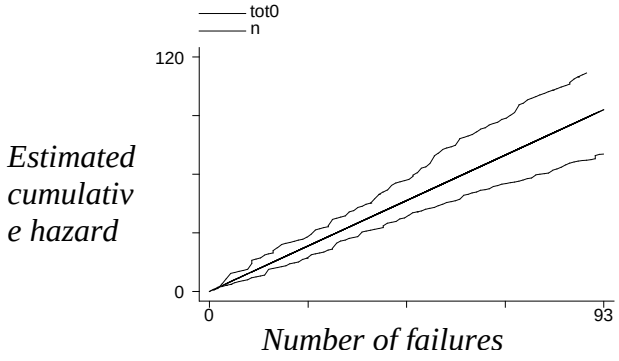


Figure 2.6.1 Arjas plot

If the hazards are proportional, then these curves should be approximately linear with slope close to one. However, even under proportional hazards the curves may differ from the 45 degree line, as seen in Figure 2.6.1, but they are still fairly linear. When the hazards are not proportional, then the curves are roughly as close to the 45 degree line as under proportionality, but the curves are not linear. To determine if these curves differ nonlinearly from the 45 degree line, one can estimate, for each stratum, a linear regression of $\sum_{j\in\ell} H_j(t_i)$ on $\sum_{j\in\ell} N_j(T_i)$, where $H_j(t_i)$ is the cumulative hazard for the j[th] individual in the sample at time t_i, i = 1, …, D. Let $\text{Diff}_{\text{max6}}$ denote the maximum absolute difference between $\sum_{j\in\ell} H_j(t_i)$ and the estimated (fitted) values from the regression. Figure 2.6.2 shows the distribution of 10,000 generated $\text{Diff}_{\text{max6}}$ values under proportional hazards.

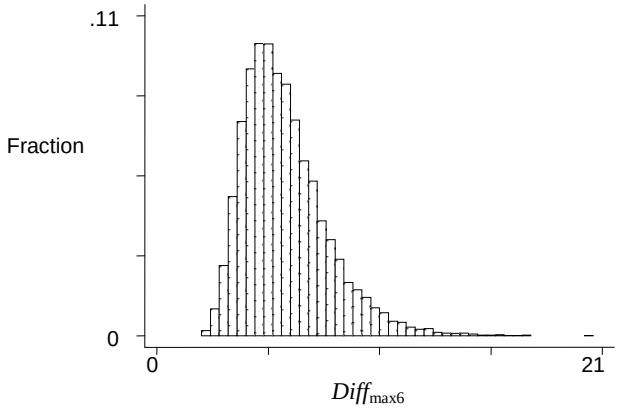


Figure 2.6.2 10,000 $Diff_{max6}$ values generated under proportional hazards

The hazards are concluded to be nonproportional if the calculated value of $\text{Diff}_{\text{max6}}$ exceeds the 95[th] percentile of this distribution.

3. SIMULATION STUDY

The six graphical methods described in subsections 2.1 – 2.6 are evaluated under proportional hazards and five different forms of nonproportional hazards: (1) increasing hazards, (2) decreasing hazards,
(3) crossing hazards, (4) diverging hazards, and (5) nonmonotonic hazards. The methods are compared in the two-sample case corresponding to two groups with different hazard functions. Equal sample sizes

of 30, 50, and 100 observations per group are used along with average censoring rates of 10, 25, 50, and 70 percent. Random (noninformative) censoring using an exponential censoring distribution is incorporated. The smallest sample size is not used at the highest censoring rate because of the small number (18) of events that would result. The number of repetitions used in each simulation is 10,000.

Random samples of survival times, T_s, are generated from the Weibull distribution in all cases except for the nonmonotonic hazards, where the lognormal distribution, having probability density function

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(t) - \mu}{\sigma} \right)^2 \right],$$

is used. The hazard of the Weibull distribution is defined as $h(t) = \alpha\gamma(\alpha t)^{\gamma-1}$, where α is the scale parameter and γ is the shape parameter. Details about parameter values are described in each case below. Censoring times, T_c, are generated from the exponential distribution with hazard function $h(t) = \lambda$, where the value of the parameter λ is adjusted to achieve the desired censoring rates. The time on study, T, is defined as $T = \min(T_s, T_c)$, where T_s and T_c are independent.

The criteria described in subsections 2.1 – 2.6 are used for rejection of the null hypothesis of proportional hazards for a test procedure that is adjusted for the appropriate sample size and censoring rate. A significance level of 5% is used. The results from the simulations are presented in the form of a plot and a numerical summary (table) presenting the proportion of times that the criterion, $Diff_{maxk}$, k = 1, 2, …, 6, exceeds the 95th percentile of the corresponding reference distribution, thus indicating "strong" evidence that the hazards are not proportional.

In subsections 3.1 – 3.6 below, the parameter settings for the survival distributions and the figures are given as follows.

Subsec.	Hazards	Survival Dist.	Group 1	Group 2	Fig
3.1	Proportional	Weibull(α,γ)	(α,γ)=(1,1)	(α,γ)=(2,1)	3.1.1
3.2	Increasing	Weibull(α,γ)	(α,γ)=(2,1.5)	(α,γ)=(2,2)	3.2.1
3.3	Decreasing	Weibull(α,γ)	(α,γ)=(2,.3)	(α,γ)=(2,.5)	3.3.1
3.4	Crossing	Weibull(α,γ)	(α,γ)=(2,1.5)	(α,γ)=(5,1)	3.4.1
3.5	Diverging	Weibull(α,γ)	(α,γ)= (1,.95)	(α,γ)= (1,1.5)	3.5.1
3.6	Nonmon-otonic	Log-normal(μ,σ)	(μ,σ)=(.3,1)	(μ,σ)=(1,1)	3.6.1

3.1 Proportional Hazards

The proportion of times that $Diff_{maxk}$ exceeds the 95th percentile of the reference distribution is given for each censoring rate, sample size, and k = 1, 2, …, 6 in Figure 3.1.2. The numerical values can be found in Appendix Table A1.

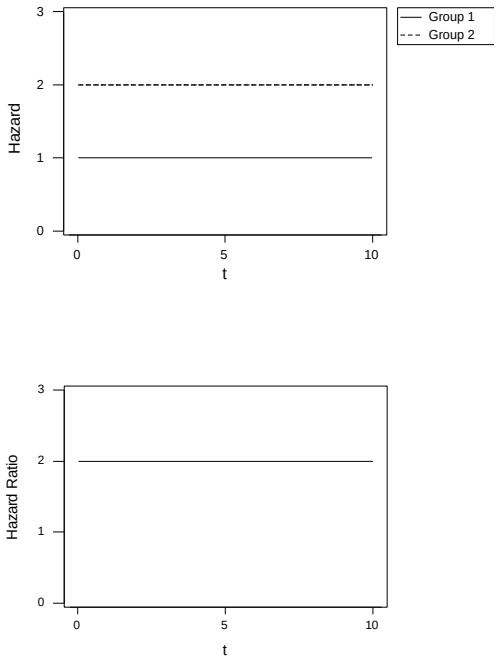


Figure 3.1.1 Constant hazards.

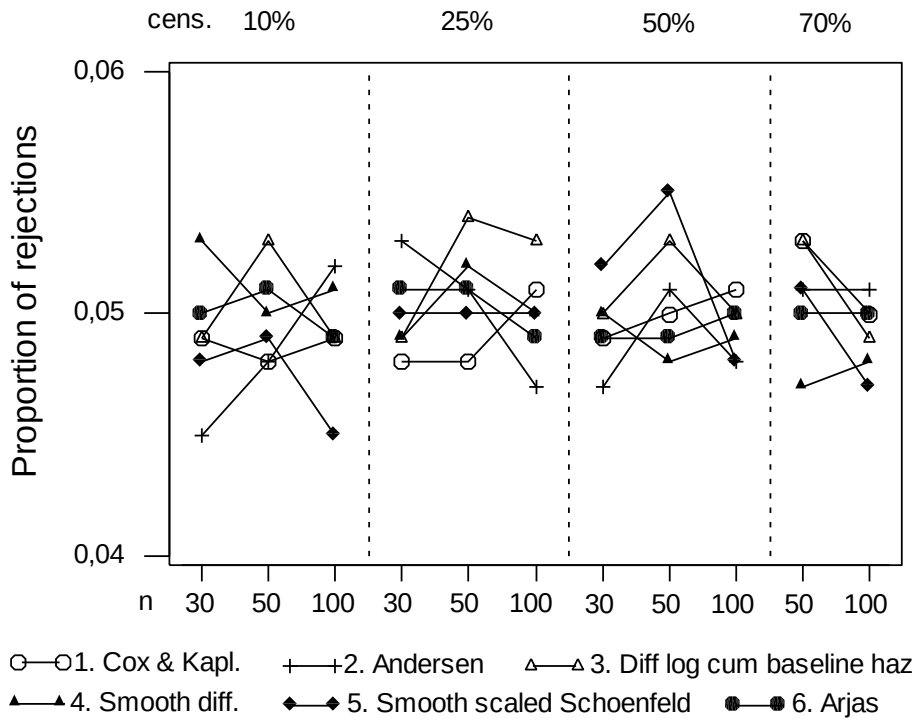


Figure 3.1.2 Proportional hazards

In order to compare the percentages, one can calculate the standard deviation of the proportion under the proportional hazards model:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n_{rep}}} = \sqrt{\frac{(0.05)(0.95)}{5000}} \approx 0.0031$$

. The standard deviation of the difference between two independent proportions is $\sigma_{diff} \approx 0.0044$, so $3\sigma_{diff} \approx 0.013$ can be used as an informal benchmark for a real difference between significance levels. However, due to the multiplicity of comparisons, this benchmark must be used cautiously.

All of the graphical procedures behave as expected, with the percentage of rejections of proportional hazards close to the significance level of 5%. In fact, all percentages fall between 0.045 and 0.055.

3.2 Increasing Hazards

The proportion of times that $\text{Diff}_{\text{maxk}}$ exceeds the 95$^{\text{th}}$ percentile of the reference distribution is given for each censoring rate, sample size, and k = 1, 2, …, 6 in Figure 3.2.2. The numerical values can be found in Appendix Table A2.

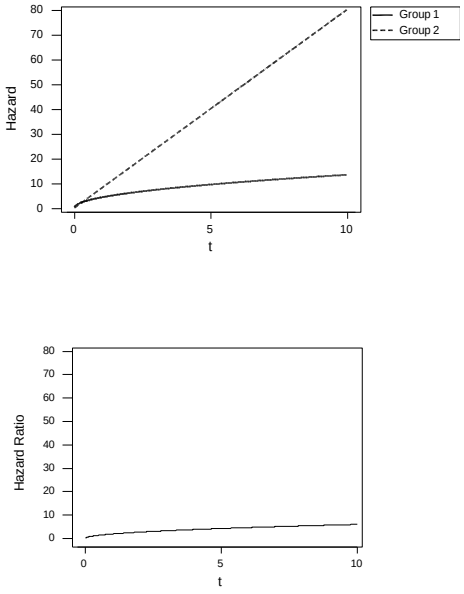


Figure 3.2.1 Increasing hazards.

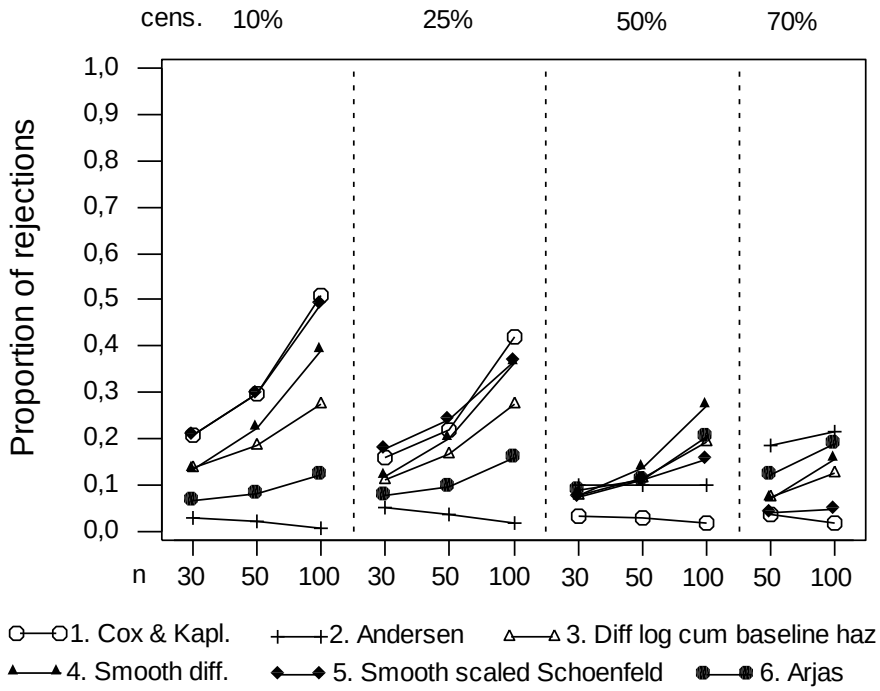


Figure 3.2.2 Increasing hazards

In order to compare the percentages, one can calculate the maximum standard deviation of the difference between two proportions:

$$\sigma_{diff}^{\max} = \sqrt{2 \times \frac{(0.5)^2}{5000}} \approx 0.01,$$ and

$3\sigma_{diff}^{\max} \approx 0.03$ serves as a conservative benchmark signifying a real difference. The same benchmark can be used in subsections 3.3 – 3.6. Again, because of the multiplicity of comparisons this benchmark must be used cautiously.

Methods 2 and 6 perform relatively poorly at the 10% and 25% censoring rates, while method 1 performs poorly at the 50% and 70% censoring rates.

3.3 Decreasing Hazards

The proportion of times that $\text{Diff}_{\text{maxk}}$ exceeds the 95$^{\text{th}}$ percentile of the reference distribution is given for each censoring rate, sample size, and k = 1, 2, …, 6 in Figure 3.3.2. The numerical values can be found in Appendix Table A3.

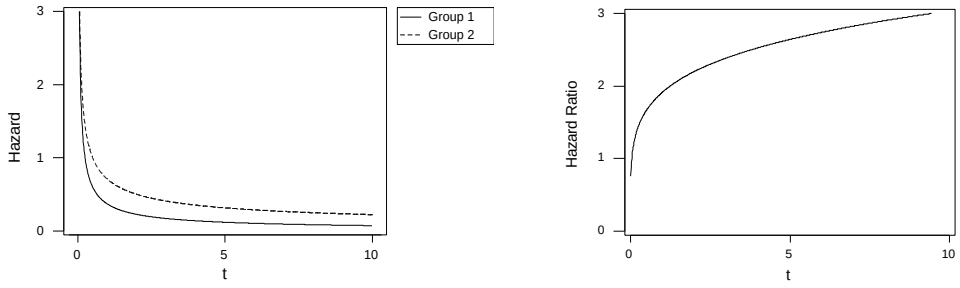


Figure 3.3.1 Decreasing hazards.

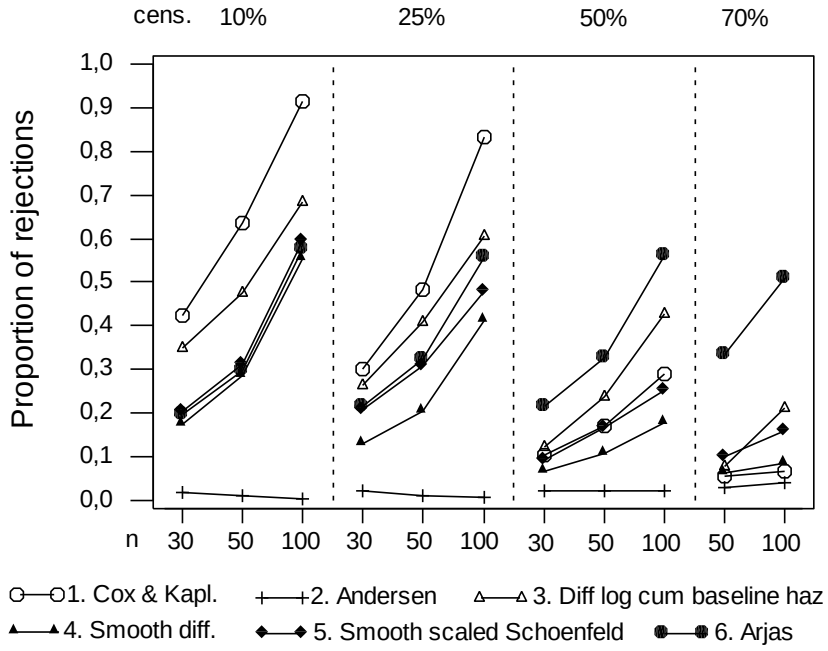


Figure 3.3.2 Decreasing hazards

Method 2 is consistently inferior. Methods 1 and 3 perform well at the 10% and 25% censoring rates. Method 6 performs well at the 50% and 70% censoring rates.

The proportion of times that $Diff_{maxk}$ exceeds the 95^{th} percentile of the reference distribution is given for each censoring rate, sample size, and k = 1, 2, …, 6 in Figure 3.4.2. The numerical values can be found in Appendix Table A4.

3.4 Crossing hazards

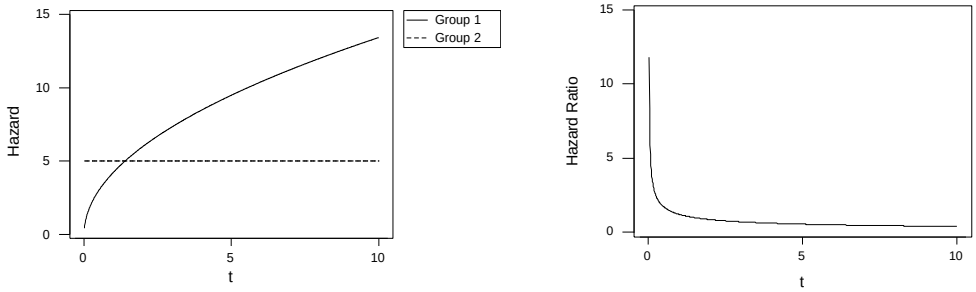


Figure 3.4.1 Crossing hazards.

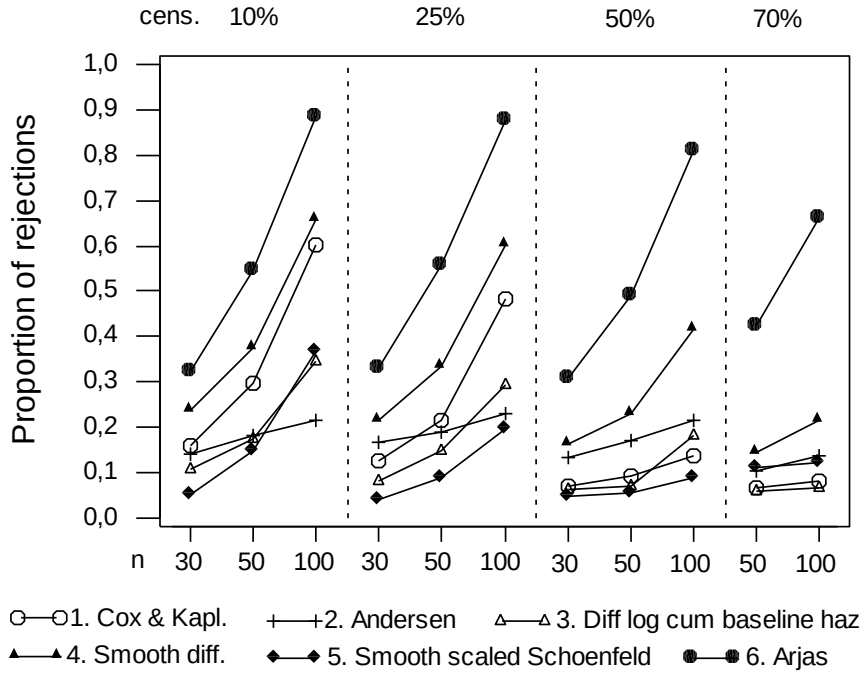


Figure 3.4.2 Crossing hazards

Method 6 is consistently superior. Method 4 also performs consistently well.

3.5 Diverging Hazards

The proportion of times that $Diff_{maxk}$ exceeds the 95^{th} percentile of the reference distribution is given for each censoring rate, sample size, and k = 1, 2, …, 6 in Figure 3.5.2. The numerical values can be found in Appendix Table A5.

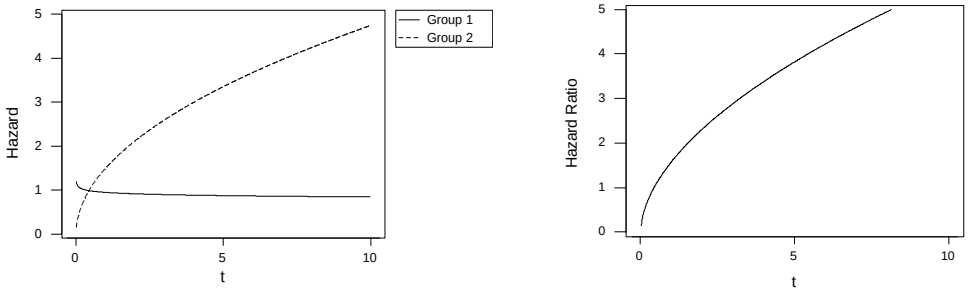


Figure 3.5.1 Diverging hazards.

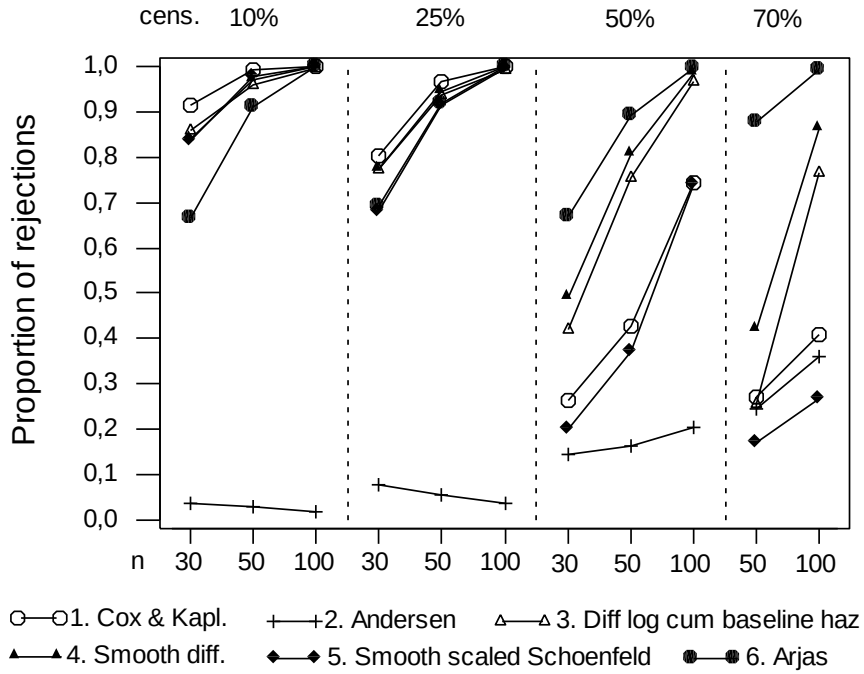


Figure 3.5.2 Diverging hazards

Method 2 performs consistently poorly, especially at the 10% and 25% censoring rates.

The proportion of rejections is shown for each censoring rate, sample size, and k = 1, 2, …, 6 in Figure 3.6.2. The numerical values can be found in Appendix Table A.6.

3.6 Nonmonotonic Hazards

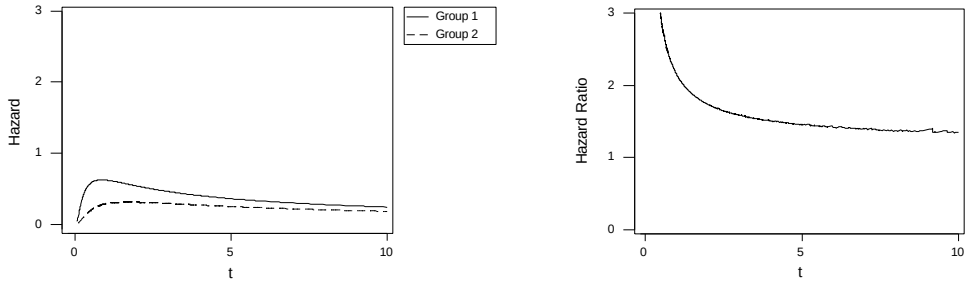


Figure 3.6.1 Nonmonotonic hazards.

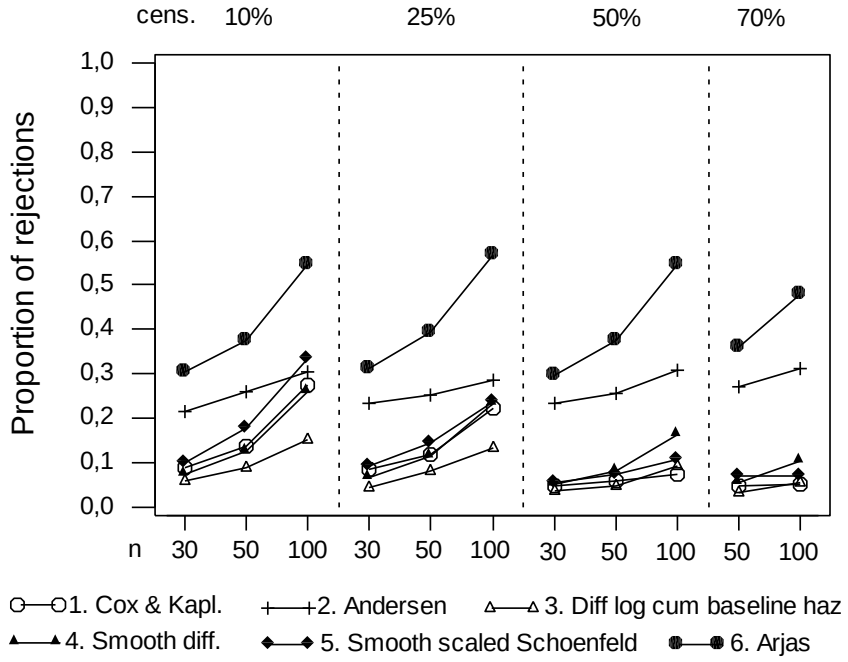


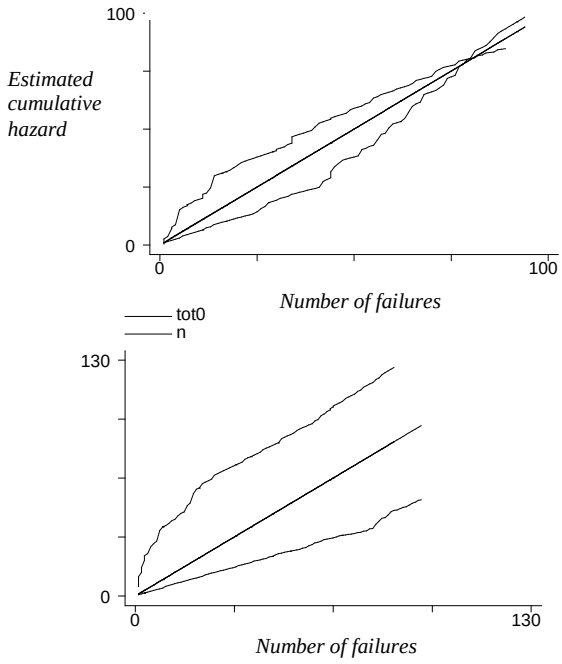
Figure 3.6.2 Nonmonotonic hazards

Method 6 is consistently superior. Method 2 also performs consistently well.

4. DISCUSSION

In each of the five forms of nonproportional hazards, the proportion of rejections generally increases with sample size and decreases with censoring rate. The conclusions from the simulations is that Method 6, the Arjas plot, finds nonproportionality more often than the other methods, especially
(1) for crossing and nonmonotonic hazards and (2) at the higher censoring rates for decreasing and diverging hazards. For decreasing hazards, Method 1, the Cox and Kaplan-Meier survival versus time plot, is superior at the low censoring rates. The Andersen plot, Method 2, performs poorly in all situations except for nonmonotonic hazards where it performs well.

These results are consistent with a viewing of the plots derived from data sets. Figure 4.1 shows examples of Method 6, which performed well for crossing and nonmonotonic hazards. The sample size is 100 and the censoring rate is 10% for both plots.



a) increasing hazards b) crossing hazards

Figure 4.1 Arjas plot

It is fairly easy to see that at least one of the curves differs nonlinearly from the 45 degree line. Under crossing hazards the distance between the curves and the 45 degree line is also larger than it is under proportional hazards (see Figure 2.6.1). Even though the Arjas plot did not perform as well under increasing hazards, it is still easy to see that the curves differ nonlinearly from the 45 degree line (Figure 4.1a; compare to Figure 2.6.1).

The other method that performed well in the simulations, especially at low censoring rates, except for crossing and nonmonotonic hazards, is Method 1. Figure 4.2 shows an example of that plot under decreasing hazards, sample size 100 and 10% censoring rate. A departure of the two estimates can be seen in the figure (compare to Figure 2.1.1).

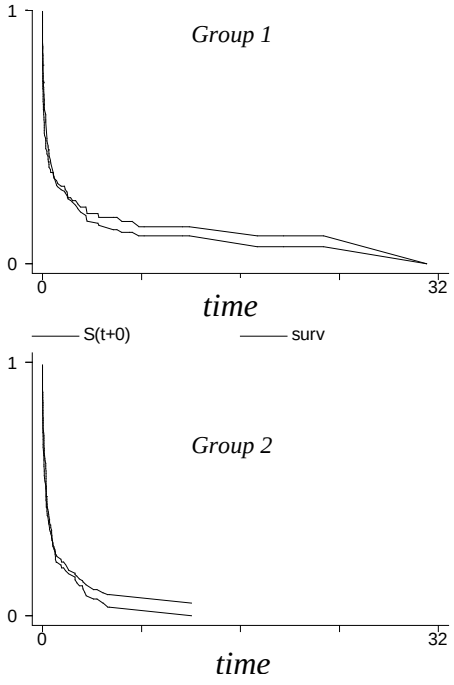


Figure 4.2 Survival curves based on the Cox model along with Kaplan-Meier estimates for each of two groups of patients

Figure 4.3 shows an example of the Andersen plot, the method that performed worse than the others in almost every situation, under diverging hazards, with sample size 100 and 10% censoring rate. It is difficult to conclude that the line does not follow a straight line through the origin (compare to Figure 2.2.1).

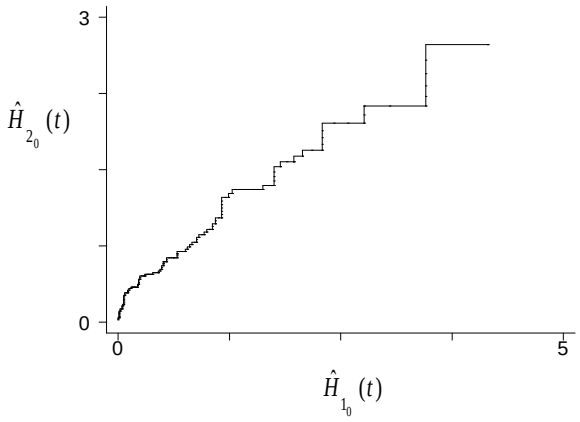


Figure 4.3 Estimated cumulative baseline hazard rate in group 2 versus group 1

The Andersen plots look similar to this under any of the nonproportional hazards cases; only in a few cases would it be possible to detect a deviation from a straight line with the naked eye.

5. APPLICATION

A multicenter study of the disease CML, chronic myeloid leukemia, was initiated in 1984 at the University Hospital in Uppsala, Sweden. CML is a cancer of the blo[illegible] where the patient has a high number of w[illegible] od cells, granulocytes, in bone marrow and blood. The treatment of this disease aims to reduce the number of white blood cells. "Cell-restraining drugs" which reduce the production of these blood cells are used in treatment. The two treatments, busulphane and hydroxyurea, were widely used all over the world at the time of this study. In previous studies, these treatments were found to be equally effective at prolonging the lifetimes of the patients (Hehlmann et al., 1993 and Alan et al., 1995).

Patients were recruited from all hospitals in Sweden. All patients older than five years and willing to participate, diagnosed with CML from January 1, 1984 until December 31, 1988, were included in the study. The patients were randomized to one of the two treatments at the date of diagnosis. All patients younger than approximately 45 years of age with a compatible donor (only brothers or sisters) were offered bone marrow transplantation. The last patient was included in the study in May 1988, and all patients were followed until February 1998. A total of 63 patients were included in the study, 26 of which received bone marrow transplantation. Figure 5.1 shows the Kaplan-Meier survival curves for patients who received a transplant (transpl 1) and those who did not receive a transplant (transpl 0).

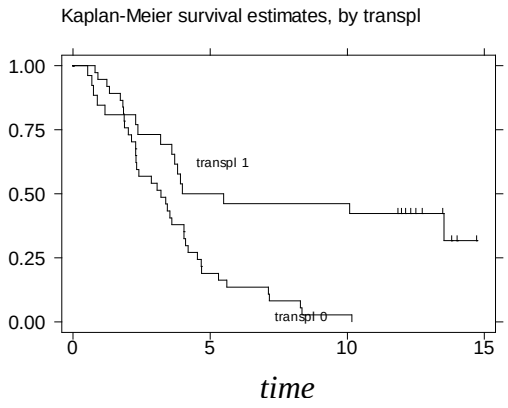


Figure 5.1 Kaplan-Meier survival for transplanted and not transplanted patients.

The censoring rate for these data is 16%. The transplantation covariate (1 = Yes, 0 = No) was believed to be time-dependent, so that the proportional hazards assumption for the Cox model was under question. Figure 5.2 shows the hazard rates for the two groups.

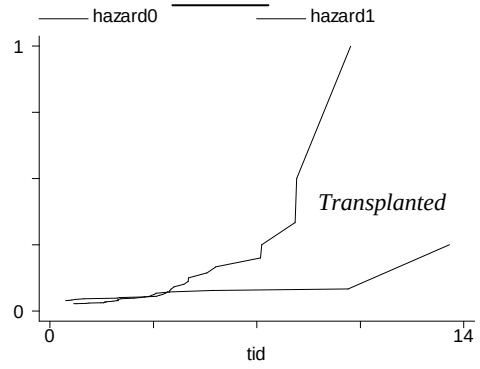


Figure 5.2 Hazard rates for transplanted and not transplanted patients.

The hazard rate *time* t early times and then diverge. From the results of section 3, the Arjas plot (Method 6) should be an effective graphical method to assess the proportional hazards assumption.

Figures 5.3 – 5.8 show the six different graphical methods described in sections 2.1 – 2.6 applied to the CML data with transplantation as a single binary covariate.

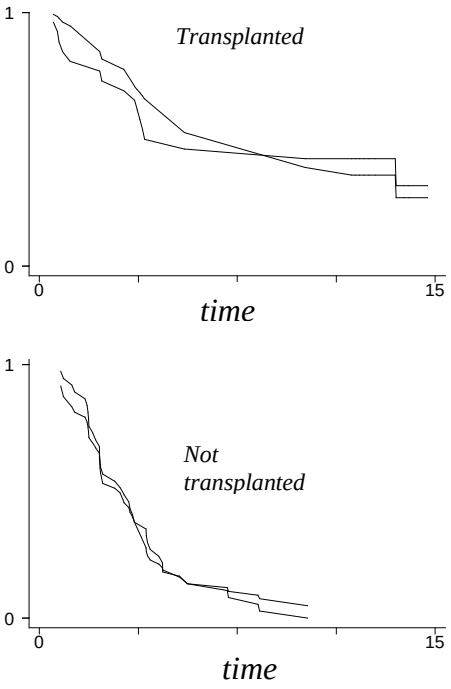


Figure 5.3 Method 1, Survival curves based on the Cox model along with Kaplan-Meier estimates for transplanted and not transplanted patients

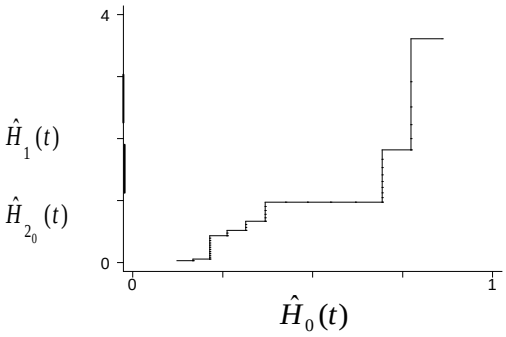


Figure 5.4 Method 2, Estimated cumulative baseline hazard rate for transplanted patients versus not transplanted patients (Andersen plot)

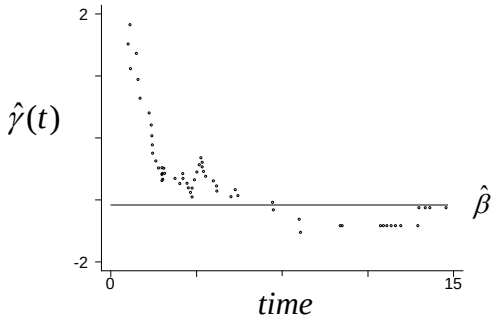


Figure 5.5 Method 3, Plot of $\hat{\gamma}(t)$ versus time

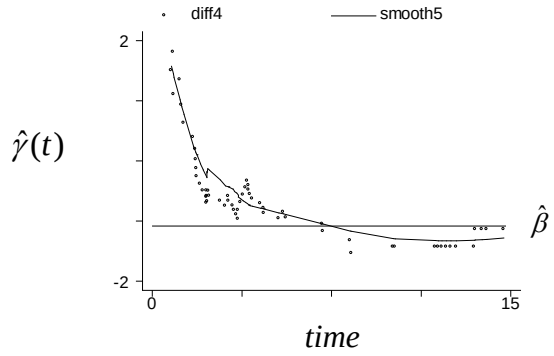


Figure 5.6 Method 4, Smoothed plot of $\hat{\gamma}(t)$ versus time

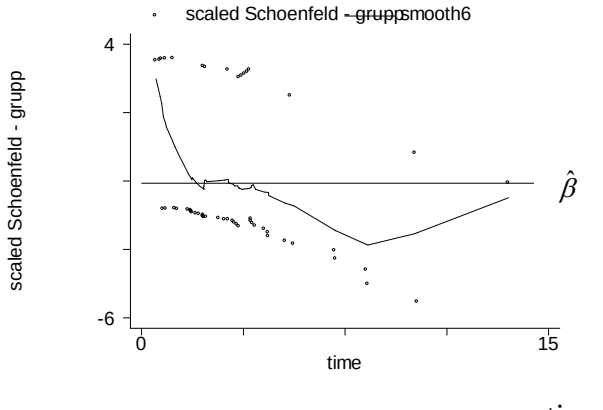


Figure 5.7 Method 5, Smoothed plot of scaled Schoenfeld residuals versus time

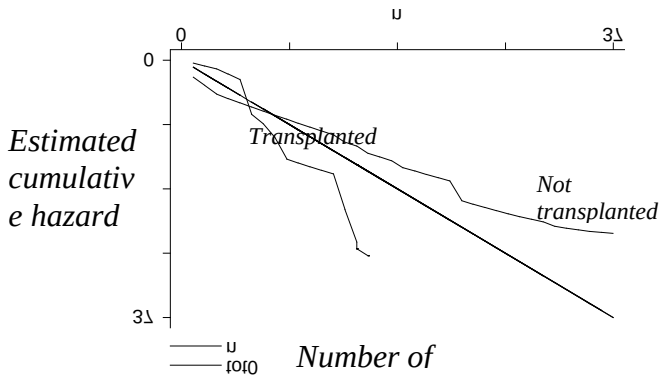


Figure 5.8 Method 6, the Arjas plot

General conclusions from these plots are given as follows.

Method 1.

There is a clear difference between the two estimates, especially for transplanted patients (compare to Figure 2.1.1), which signals a violation of the proportional hazards assumption.

Method 2.

A deviation from linearity can be detected in the Andersen plot (compare to Figure 2.2.1).

Method 3.

This plot is not constant over time, as would be the case if the hazards were proportional (compare to Figure 2.3.1).

Method 4.

The smoothing helps the analyst to determine that the plot is not constant over time (compare to Figure 2.4.1).

Method 5.

The residuals do not tend to center around $\hat{\beta}$ (compare to Figure 2.5.1).

Method 6.

Both curves cross the 45 degree line and differ nonlinearly from it, especially the curve for the transplanted patients (compare to Figure 2.6.1).

CONCLUSION

Assessing graphs for the purpose of determining the severity of model assumption violations can be difficult because of the lack of objectivity involved. To the untrained eye, several of the plots in Figures 2.1.1 – 2.6.1 may appear to signal a violation of the proportional hazards assumption (e.g., Figures 2.2.1 and 2.4.1) even though they were generated from models having proportional hazards.

By using a Kolmogorov-Smirnov like maximum deviation criterion upon which to base comparisons of six different graphical procedures, this simulation study shows that the Arjas plot is generally the most effective at identifying nonproportional hazards, especially for (1) crossing and nonmonotonic hazards and (2) decreasing and diverging hazards where the censoring rate is high. It is interesting to note that the effectiveness of the Arjas plot at identifying nonproportional hazards remains relatively constant across the censoring rates while for most all of the other methods the

proportion of rejections tends to decrease with censoring rate.

When the proportion of rejections is averaged over sample sizes and censoring rates, Method 2 performs the worst under increasing, decreasing, and diverging hazards while Method 6 performs the best under crossing and nonmonotonic hazards. The average rejection rates are given as follows:

	Method					
Hazards	1	2	3	4	5	6
Increasing	.18	.08	.16	.20	.20	.12
Decreasing	.39	.02	.35	.21	.26	.37
Crossing	.21	.17	.14	.33	.12	.56
Diverging	.71	.13	.79	.83	.65	.87
Nonmono-tonic	.11	.27	.08	.12	.13	.41

Method 6, the Arjas plot, has one of the top two average rejection rates for four of the five forms of nonproportional hazards. For increasing hazards, where Method 6 has the fifth highest average rejection rate, Methods 4 or 5 would be recommended.

The maximum absolute deviation criterion used in this study is consistent with the practical useage of plots to determine if model assumptions are plausible. That is, when one visually analyzes a plot, one is searching for the deviation between the observed plot and the plot one expects to see under the model assumption. This study merely formalizes this process.

It is recommended that in general the Arjas plot (Method 6) be used as the preferred graphical procedure for checking the proportional hazards assumption if the form of the nonproportional hazards is anything but increasing. For increasing hazards, Methods 4 and 5 are superior.

Generally, it is recommended that the proportional hazards assumption always be checked in the Cox model, and that while a plot such as the Arjas plot is a helpful tool, it should not be the only basis upon which to make a decision regarding the plausibility of the proportional hazards assumption.

REFERENCES

Alan, N.C., Richards, S.M. and Stepherd, P.C.A. (1995). UK Medical Research Council randomised, multicentre trial of interferon-an1 for chronic myeloid leukaemia: improved survival irrespective of cytogenetic response. *The Lancet*, 345, 1392-1397.

Andersen, P.K. (1982). Testing Goodness of Fit of Cox's Regression and Life Model. *Biometrics*, 38, 67-77. Correction (1984): 40, 1217.

Andersen, P.K., Borgan, Ø., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Arjas, E. (1988). A Graphical Method for Assessing Goodness of Fit in Cox's Proporional Hazards Model. *J. American Statistical Association*, 83, 204-212.

Arjas, E. and Haara, P. (1988). A note on the exponentiality of total hazards before failure. *J. Multivariate Analysis*, 26, 207-218.

Breslow, N.E. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, 30, 89-99.

Cox, D.R. (1972). Regression Models and Life-Tables (with discussion). *Journal of the Royal Statistical Society – Series B*, 34, 187-220.

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

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman and Hall, London, Ch. 3, 4

Hehlmann, R., Heimpel, H., Hasford, J., Kolb, HJ., Pralle, H., et al., The German CML Study Group (1993). Randomized Comparison of Busulfan and Hydroxyurea in Chronic Myelogenous Leukemia (CML): Prolongation of Survival by Hydroxyurea. *Blood*, 82, 398.

Hess, K.R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statistics in Medicine*, 14, 1707-1723.

Hosmer, D.W. and Lemeshow, S. (1999). *Regression Modeling of Time to Event Data*. Wiley, New York.

Kalbfleisch, J.D. and Prentice , R.L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York, pp. 86-89.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. American Statistical Association*, 53, 457-481.

Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. Wiley, New York.

Kleinbaum, D.G. (1996). *Survival Analysis, A Self-Learning Text*. Springer, New York.

Lin, D.Y. and Wei, L.J. (1991). Goodness-of-fit tests for the general Cox regression model. *Statistica Sinica*, 1, 1-17.

Lin, D.Y., Wei, L.J., and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80, 557-572.

Link, C.L. (1984). Confidence intervals for the survival function using Cox's proportional-hazard model with covariates. *Biometrics*, 40, 601-610.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1, 27-52.

Persson, I. (2002). Essays on the assumption of proportional hazards in Cox regression. Acta Universitatis Upsaliensis, Uppsala University, Uppsala, Sweden. Unpublished Ph.D. Dissertation, Chapter 2.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67, 145-153.

Schumacher, M. (1990). Evaluation of nonproportional treatment effects in cancer clinical trials. *Cancer Investigation*, 8, 91-98.

Sokal, R. R. and Rohlf, F.J. (1995). *Biometry, 3rd edition*. W.H. Freeman and Company, New York.

APPENDICES

Table A.1. Proportion of rejections for constant hazards.

	Censoring										
	10%			25%			50%			70%	
	Sample size in each group										
Test statistic	30	50	100	30	50	100	30	50	100	50	100
1.Cox and Kaplan-Meier Survival vs. time	.049	.048	.049	.048	.048	.051	.049	.050	.051	.053	.050
2. Cumulative Baseline Hazard (Andersen plot)	.045	.048	.052	.053	.051	.047	.047	.051	.048	.051	.051
3. Difference of log Cum. Baseline Hazard vs. Time	.049	.053	.049	.049	.054	.053	.050	.053	.050	.053	.049
4. Smoothed Diff. of log Cum. Base-line Haz. vs. Time	.053	.050	.051	.049	.052	.050	.050	.048	.049	.047	.048
5. Smoothed scaled Schoenfeld residuals vs. Time	.048	.049	.045	.050	.050	.050	.052	.055	.048	.051	.047
6. Arjas plot of Cum. Haz. vs. number of failures	.050	.051	.049	.051	.051	.049	.049	.049	.050	.050	.050

Table A.2. Proportion of rejections for increasing hazards

	Censoring										
	10%			25%			50%			70%	
	Sample size in each group										
Test statistic	30	50	100	30	50	100	30	50	100	50	100
1.Cox and Kaplan-Meier Survival vs. time	.207	.296	.511	.159	.221	.420	.032	.029	.018	.038	.019
2. Cumulative Baseline Hazard (Andersen plot)	.029	.021	.008	.052	.036	.018	.101	.100	.101	.184	.214
3. Difference of log Cum. Baseline Hazard vs. Time	.138	.185	.274	.110	.169	.274	.077	.116	.195	.073	.127
4. Smoothed Diff. of log Cum. Base-line Haz. Vs. Time	.133	.224	.390	.118	.201	.365	.079	.139	.270	.072	.157
5. Smoothed scaled Schoenfeld residuals vs. Time	.210	.296	.492	.180	.243	.369	.076	.113	.156	.042	.050
6. Arjas plot of Cum. Haz. vs. number of failures	.066	.081	.124	.079	.097	.158	.088	.112	.203	.124	.190

Table A.3. Proportion of rejections for decreasing hazards

	Censoring										
	10%			25%			50%			70%	
	Sample size in each group										
Test statistic	30	50	100	30	50	100	30	50	100	50	100
1.Cox and Kaplan-Meier Survival vs. time	.424	.634	.915	.302	.485	.832	.105	.171	.289	.054	.067
2. Cumulative Baseline Hazard (Andersen plot)	.020	.011	.004	.021	.011	.006	.024	.023	.024	.031	.040
3. Difference of log Cum. Baseline Hazard vs. Time	.348	.477	.683	.264	.409	.607	.122	.237	.428	.078	.212
4. Smoothed Diff. of log Cum. Base-line Haz. vs. Time	.175	.285	.553	.129	.205	.414	.066	.107	.179	.062	.087
5. Smoothed scaled Schoenfeld residuals vs. Time	.204	.313	.596	.210	.310	.481	.094	.168	.251	.099	.158
6. Arjas plot of Cum. Haz. vs. number of failures	.198	.298	.578	.216	.322	.559	.217	.326	.561	.334	.509

Table A.4. Proportion of rejections for crossing hazards

	Censoring										
	10%			25%			50%			70%	
	Sample size in each group										
Test statistic	30	50	100	30	50	100	30	50	100	50	100
1.Cox and Kaplan-Meier Survival vs. time	.159	.296	.601	.127	.216	.484	.069	.094	.136	.068	.083
2. Cumulative Baseline Hazard (Andersen plot)	.141	.183	.214	.166	.188	.232	.133	.170	.215	.104	.136
3. Difference of log Cum. Baseline Hazard vs. Time	.108	.176	.345	.080	.147	.294	.062	.071	.181	.061	.066
4. Smoothed Diff. of log Cum. Base-line Haz. vs. Time	.239	.374	.657	.217	.333	.602	.163	.231	.418	.146	.216
5. Smoothed scaled Schoenfeld residuals vs. Time	.052	.148	.369	.040	.090	.196	.049	.056	.089	.112	.121
6. Arjas plot of Cum. Haz. vs. number of failures	.322	.546	.884	.330	.557	.878	.309	.492	.809	.424	.662

Table A.5. Proportion of rejections for diverging hazards

	Censoring										
	10%			25%			50%			70%	
	Sample size in each group										
Test statistic	30	50	100	30	50	100	30	50	100	50	100
1.Cox and Kaplan-Meier Survival vs. time	.916	.993	1.00	.803	.967	1.00	.263	.427	.745	.270	.408
2. Cumulative Baseline Hazard (Andersen plot)	.038	.030	.017	.077	.055	.039	.145	.163	.206	.247	.360
3. Difference of log Cum. Baseline Hazard vs. Time	.858	.960	.998	.772	.936	.994	.419	.755	.967	.256	.764
4. Smoothed Diff. of log Cum. Base-line Haz. vs. Time	.839	.969	1.00	.774	.945	.999	.490	.806	.986	.420	.863
5. Smoothed scaled Schoenfeld residuals vs. Time	.916	.993	1.00	.803	.967	1.00	.263	.427	.745	.270	.408
6. Arjas plot of Cum. Haz. vs. number of failures	.038	.030	.017	.077	.055	.039	.145	.163	.206	.247	.360

Table A.6. Proportion of rejections for nonmonotonic hazards

	Censoring										
	10%			25%			50%			70%	
	Sample size in each group										
Test statistic	30	50	100	30	50	100	30	50	100	50	100
1.Cox and Kaplan-Meier Survival vs. time	.091	.138	.274	.086	.119	.223	.047	.061	.073	.048	.051
2. Cumulative Baseline Hazard (Andersen plot)	.217	.260	.306	.236	.252	.286	.234	.257	.308	.273	.313
3. Difference of log Cum. Baseline Hazard vs. Time	.061	.089	.151	.046	.080	.132	.037	.050	.094	.035	.057
4. Smoothed Diff. of log Cum. Base-line Haz. vs. Time	.076	.128	.259	.066	.117	.235	.053	.082	.165	.057	.103
5. Smoothed scaled Schoenfeld residuals vs. Time	.100	.177	.336	.092	.146	.239	.056	.074	.107	.069	.071
6. Arjas plot of Cum. Haz. vs. number of failures	.303	.375	.545	.311	.394	.567	.296	.375	.545	.360	.481