Validation of the Smooth Test of Goodness-of-Fit for Proportional Hazards in Cancer Survival Studies

Collins Odhiambo[1], John Odhiambo[1] and Bernard Omolo[2,*]

[1]*Strathmore Institute of Mathematical Sciences, Strathmore University, Ole Sangale Road, Nairobi, Kenya*

[2]*Division of Mathematics & Computer Science, University of South Carolina-Upstate, 800 University Way, Spartanburg, South Carolina, USA*

Abstract: In this study, we validate the smooth test of goodness-of-fit for the proportionality of the hazard function in the two-sample problem in cancer survival studies. The smooth test considered here is an extension of Neyman's smooth test for proportional hazard functions. Simulations are conducted to compare the performance of the smooth test, the data-driven smooth test, the Kolmogorov-Smirnov proportional hazards test and the global test, in terms of power. Eight real cancer datasets from different settings are assessed for the proportional hazard assumption in the Cox proportional hazard models, for validation. The smooth test performed best and is independent of the number of covariates in the Cox proportional hazard models.

Keywords: Cancer, Cox proportional hazards model, Global test, Neyman's smooth test, Two-sample problem.

1. BACKGROUND

Most cancer studies that focus on the identification of survival risk factors use models that assume proportional hazards. With over 100 different types of cancer known today, targeted research in this area is particularly important for the ten most prevalent cancers (i.e. lung and bronchus, prostate, breast, colon and rectum, pancreas, liver and intrahepatic bile duct, leukemia, urinary bladder, non-hodgkin lymphoma, brain and other nervous system [1]) in the fight against cancer. Oncologic research models failure-time data in order to identify the risk factors. The time-to-failure data analyses in this aspect have consistently used the Cox proportional hazard (CPH) models. An important aspect of analysis using the CPH model is the verification of the proportional hazards assumption. The CPH model assumes that the effects of covariates do not change over time. Verification of the proportionality assumption is critical because the model will be rendered invalid if it is violated [2]. Since the development of the CPH model [3], different authors have proposed several methods of ascertaining the proportionality assumption, ranging from graphical methods (e.g. [4-7]) to non-graphical methods (e.g. [3, 8-14]).

To the best of our knowledge, the validation of existing methods for assessing the proportional hazards assumption using real datasets from different bio-medical settings has not been performed. Here, we assess the performance of the smooth test against the Kolmogorov-Smirnov and the global test for the proportional hazards assumption. The null hypothesis is that the coefficient of the jth covariate in the CPH model is independent of time (i.e. $\beta_j(t) = \beta_j$).

The smooth test considered here is an extension of Neyman's smooth tests to proportional hazard functions. Neyman's classical approach [15] has been extended to hazard functions (see [16-22]). In these studies, one or two example datasets have been used to illustrate the validity of the test. Our approach employs datasets from eight different cancer survival studies.

The test is essentially a score test and is derived by nesting the null hypothesis in a larger class of hazard functions. Directional tests are designed to have high power against a specific departure from the null hypothesis and omnibus tests are constructed without any specific alternative. Neyman's smooth test is considered to be a compromise between the two [19] and is capable of detecting a wider range of alternatives.

We focused on smooth tests for the two-sample problem in the presence of censoring, a situation that is common in failure-time data. We also compared data-driven versions of smooth test as proposed by Kraus [18] and Ledwina [15]. The eight cancer datasets analyzed are from the most prevalent and deadly cancers for men and women in United States [1].

The Cox PH model is defined as

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta^T X_i(t)\}, \tag{1}$$

*Address correspondence to this author at the Division of Mathematics & Computer Science, University of South Carolina-Upstate, 800 University Way, Spartanburg, South Carolina, USA; Tel: +1 864-503-5362; Fax: +1 864-503-5930; E-mail: bomolo@uscupstate.edu

where $\lambda_i(t)$ is the intensity process of the i-th component of an n-variate counting process $N(t) = \{(N_1(t), N_2(t), \cdots, N_n(t)) : t \in T\}$; $Y_i(t)$ is the risk indicator process; $X_i(t)$ is a p-dimensional covariate (predictable process); $\lambda_0(t)$ is the unknown baseline hazard; and β is a vector of unknown regression coefficients. Under the proportional hazards assumption, the covariate process is time-dependent but the baseline hazard function is not, so that

$$\frac{\lambda(t \mid X^{(1)}(t))}{\lambda(t \mid X^{(2)}(t))} = \frac{\lambda_0(t)\exp\{\beta^T X^{(1)}(t)\}}{\lambda_0(t)\exp\{\beta^T X^{(2)}(t)\}} \qquad (2)$$
$$= \exp\{\beta^T(X^{(1)}(t) - X^{(2)}(t))\}.$$

Here $X^{(1)}$ is the covariate vector from sample 1 and $X^{(2)}$ is the covariate vector from sample 2. Testing the proportional hazard assumption is also equivalent to testing $H_0 : \lambda_0(t) = \lambda_0$ (constant) versus $H_A : \lambda_0(t) \neq \lambda_0$ (time-dependent).

This paper is organized as follows. We present the general overview of the smooth tests, the Kolmogorov-Smirnov test and the global test in the presence of censoring in section 2. In section 3, we give a motivation for the study and describe the cancer datasets used for validation. We also provide results for tests of proportionality. Section 4 provides a discussion of main results from the analysis in section 3. Finally, we provide concluding remarks and limitations of our study.

2. TESTS OF GOODNESS-OF-FIT

2.1. Two-Sample Smooth Test

By nesting the Cox PH model defined in (1), we obtain the model of analysis (3) below [18].

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta^T X_i(t) + \theta^T \Psi_i(\cdot, \beta)X_{iq}(t)\}, \qquad (3)$$

where $\Psi_i(\cdot, \beta) = \{(\psi_1(\cdot, \beta), \psi_2(\cdot, \beta), \cdots, \psi_k(\cdot, \beta))\}$ is a vector of some basis functions in standardized time, i.e.

$$\psi_i(\cdot, \beta) = \varpi_i(1 - \exp\{-\hat{\gamma}_0(\cdot, \beta)\}). \qquad (4)$$

Here, $\hat{\gamma}_0(\cdot, \beta)$ is the Breslow estimator of $\gamma_0(\cdot, \beta) = \int_0^t \lambda_0(s)ds$ and $\varpi_i(\cdot)$ is some bounded function on $[0,1]$ [18]. Details of the derivation of the test statistic, its asymptotic properties and how to choose k, including the data-driven version of the test, have been discussed by [18, 19, 21].

Note that testing for the hypothesized proportional hazard function in (3) above is equivalent to testing for the constant baseline hazard function ($\lambda_0(t) = \lambda_0$) and is obtained by setting $\theta = 0$ (i.e. within this nesting, $H_0 : \lambda(\cdot) = \lambda_0(\cdot)$ is equivalent to $H_0 : \theta = 0$). The score test statistic for testing $H_0 : \theta = 0$ against $H_A : \theta \neq 0$ is defined by

$$S_k = U(\cdot; \hat{\beta})^T \Sigma(\cdot; \hat{\beta})U(\cdot; \hat{\beta}), \qquad (5)$$

where $U(\cdot; \hat{\beta})$ is a score process for θ evaluated at $(\beta^T, \theta^T) = (\beta^T, (0, 0, \cdots, 0))$ and has been shown by [18] to be

$$\sum_{i=1}^{n} \int_0^t \psi(s, \beta)X_\varpi(s)dN_i(s) - \int_0^t$$
$$\frac{\sum_{i=1}^{n} Y_i(t)\Psi(s, \beta)X_\varpi(s)\exp\{\beta^T X_i(t)\}}{\sum_{i=1}^{n} Y_i(t)\exp\{\beta^T X_i(t)\}} d\bar{N}(s). \qquad (6)$$

The estimated variance of the score is given as

$$\Sigma(\cdot; \hat{\beta}) = \Sigma_{22}(\cdot; \hat{\beta}) - \Sigma_{21}(\cdot; \hat{\beta})\Sigma_{11}(\cdot; \hat{\beta})^{-1}\Sigma_{12}(\cdot; \hat{\beta}). \qquad (7)$$

Here $\Sigma_{11}(\cdot; \beta) = [U_1(\cdot; \beta)](t)$, $\Sigma_{22}(\cdot; \beta) = [U_2(\cdot; \beta)](t)$ and

$\Sigma_{21}(\cdot; \beta) = [U_2(\cdot; \beta), U_1(\cdot; \beta)](t)$. $S_k \to \chi_k^2$ for $n \to \infty$ under H_0 and H_0 is rejected for large values of S_k.

2.2. Two-Sample Kolmogorov-Smirnov (D_n) Test

The Kolmogorov-Smirnov test for the two-sample proportional hazards with right censoring has been discussed by [4, 23], and its asymptotic properties, by [24]. The test is based on the simplified partial likelihood score process and it tests the hypothesis that transformation hazards are proportional in the two samples with right censored data. The test uses the Kolmogorov-Smirnov supremum statistic. Martingale simulations are used to compute the *p*-values. Let

$$U(\hat{\beta}, t) = \{U_1(\hat{\beta}, t), U_2(\hat{\beta}, t), \cdots, U_p(\hat{\beta}, t)\} \qquad (8)$$

be the empirical score process. Then the standardized score process, $\mathfrak{F}^{-1/2}U(\hat{\beta}, t)$, is asymptotically equivalent to the Brownian bridge (B^0), where $\hat{\beta}$ is the nonparametric maximum likelihood estimator (NPMLE) of β. For the supremum test, if $p \geq 1$, each of the proportional hazards test statistics [25, 26],

$$\sup_t \{ \mathfrak{F}^{-1}(\hat{\beta})_{jj} \}^{\frac{1}{2}} \, | \, U_j(\hat{\beta}, t) \, | \, (j = 1, 2, \cdots, p), \tag{9}$$

has the asymptotic distribution of $\sup_{0 \le u \le 1} | B^0(u) |$ if $\{ V(t) \}_{jk} = 0$ for $(j \neq k)$ for all t, where $V(\cdot)$ is the limiting covariance matrix for $n^{-\frac{1}{2}} U(\beta_0, \cdot)$ [4, 26]. Therefore, testing for the overall proportionality can be done using the test statistic

$$\sup_t \| U(\hat{\beta}, t) \| \ \text{or} \ \sup_t \sum_{j=1}^{p} \{ \mathfrak{F}^{-1}(\hat{\beta})_{jj} \}^{\frac{1}{2}} \, | \, U_j(\hat{\beta}, t) \, | . \tag{10}$$

Details about the consistency of the test against nonproportional hazards alternative have been discussed by [4, 9, 12, 21, 26].

2.3. Global Test

The global test is widely used to test the proportional hazards assumption for the CPH model. The test was first proposed by [8] and is based on a semi-parametric generalization of the proportional hazards regression model. The hazard function corresponding to a covariate vector X has the time-function defined as

$$\Phi(t) = 1 + \log[\Lambda_0(t)], \tag{11}$$

where $\Lambda_0(t)$ is the cumulative baseline hazard function, which is essentially Breslow's maximum likelihood estimator under H_0 [27]. The hypothesis of proportional hazards, $H_0 : \theta = 0$, is tested using a score statistic derived from the partial likelihood. The Breslow estimator for $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ is

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(u)}{\sum\limits_{j=1}^{n} Y_j(u) \exp(\hat{\beta}^T X_j(u))}, \tag{12}$$

where $\hat{\beta}$ maximizes the partial log-likelihood of β. As in the special case of the proportional hazards model, $\hat{\beta}$ and $\hat{\Lambda}_0(t)$ are the NPMLEs. For an exhaustive coverage for $\Lambda_0(t)$, see [8, 23, 27, 28].

3. SIMULATIONS

We conducted a simulation study to ascertain proportionality under right censoring in the CPH model. Independent samples of size 10, 50, 100, 200, 500 and 1,000 were simulated and adjusted to give a chosen percentage of censored observations before the end of

follow-up (i.e. 25% to 35% censoring, 45% to 55% censoring and 65% to 75% censoring). Each simulated dataset had a treatment covariate stratified by group (i.e. 1 or 2) and one other covariate arranged to contain equal numbers of observations. The power of the test was calculated as the percentage of rejection at the 5% level of significance. All simulations and comparative analyses were performed using the R packages *survival, eha, prodlim* and *surv2sample*. For each sample size (i.e $n \in (10, 50, 100, 200, 500, 1000))$, 1,000 samples were generated and percentage rejection was computed as the number of cases rejected (with $p < 0.05$).

Simulation studies show that as the censoring percentage increases, the percentage rejection of the global and Kolmogorov-Smirnov tests increases. Also the Kolmogorov-Smirnov test is strongly affected by sample size, such that as the sample size increases, percentage rejection increases as well. It fails to detect proportionality. The smooth tests are not affected by the percentage of censoring and sample size (see Figures **16** to **21**).

4. DATA SETTING AND ANALYSIS

4.1. Motivation for Analysis of Different Cancer Datasets

Despite decades of research in cancer, the overall prognosis, recurrences and survival rates are still attracting huge research interest. Much of the research is specific to cancer-type and is beneficial to patients through advanced technologies and cancer treatment protocols. Cancer is a major public health problem and is the second leading cause of death in the United States [1]. Prostate, lung and bronchus, and colorectal cancers account for 44% of all cases in men, with prostate cancer alone accounting for 20% of new diagnoses. For women, the three most commonly diagnosed cancers are breast, lung and bronchus, and colorectum, representing 50% of all cases. Breast cancer alone accounts for 29% of all new cancer diagnoses in women. The National Center for Health Statistics (NCHS) estimated 1,600 deaths per day due to cancer in 2016. The most common causes of cancer death are cancers of the lung and bronchus, prostate, and colorectum in men and lung and bronchus, breast, and colorectum in women. These four cancer-types account for 46% of all cancer deaths, with more than one-quarter (27%) due to lung cancer. The largest geographic variation in cancer occurrence by far is for lung cancer. Cancers in adolescents (aged 15 to 19

years) differ from those in children in terms of type and distribution.

With these variations in mind, this study does not aim to provide an exhaustive performance of smooth tests for proportionality for all types of cancer, but instead it aims to statistically compare its performance in selected eight different practical settings. Our goal is to provide an overview of the performance of smooth test and to help validate the test in different cancer settings. Ultimately, we hope that our findings will lead to higher overall standards and quality of oncological research by the survival analysis community, and limit the risk of using invalid models.

4.2. Dataset 1: Survival with Malignant Melanoma

This dataset consists of measurements made on patients with malignant melanoma. Each patient had their tumour removed by surgery at the Department of Plastic Surgery, University Hospital of Odense, Denmark during the period 1962 to 1977. The surgery consisted of a complete removal of the tumour together with about 2.5cm of the surrounding skin. Measurements taken included the thickness of the tumour and whether it was ulcerated or not. Patients were followed until the end of 1977. Time was defined as survival time in days since the operation, possibly censored. The patients' status at the end of the study were death from melanoma, alive and death from causes unrelated to their melanoma. The sex of the patients was also recorded. Age was recorded in years at the time of the operation. Other variables measured included tumor thickness and an indicator of ulceration [29].

We fitted a Cox PH model with sex, tumor thickness, ulceration indicator and age as covariates.

Table 1:   Fitting of the CPH Model: Malignant Melanoma Data

Covariate	β	chisq	*p*-value
sex	0.151	1.35	0.2456
tumour thickness	-0.249	3.02	0.0823
ulceration	0.163	1.52	0.2182
age	0.207	3.08	0.0791

From Table **1** above, all the covariates are insignificant at $\alpha = 0.05$. Our focus, however, was on testing the proportionality assumption and so we created the plots of the Schoenfeld residuals against

time for the overall fit. Testing the time-dependent covariates is equivalent to testing for a non-zero slope. A non-zero slope indicates a violation of the proportional hazards assumption. We started by looking at the graphs of the Cox regression models before performing the tests of non-zero slopes.
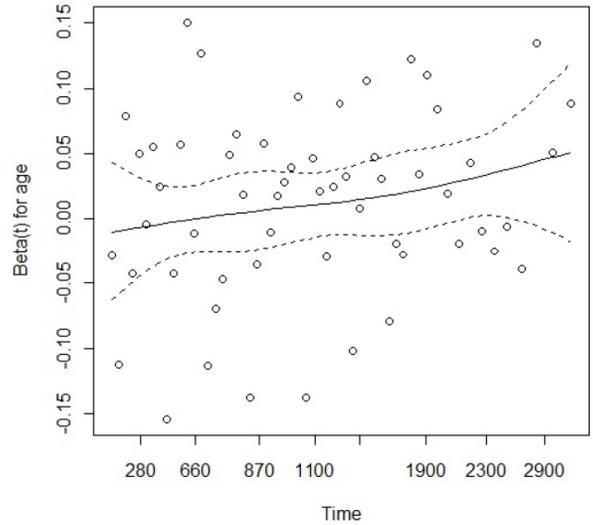

Figure 1: Schoenfeld residual plot for the overall fit: Malignant Melanoma Data.

The overall fit of the CPH model shows residuals scattered all over with a general zero slope (Figure **1**). Hence proportionality exists despite the fact that the covariates are insignificant. The next step was to create Schoenfeld residual plots for each of the four covariates, including a lowess smoothing curve. The graphs for the residuals were still scattered for the four covariates (Figure **2**).

Like in the plots, we expect all tests to fail to reject the null hypothesis, indicating that the proportionality assumption holds. We then compared the power of rejection between Kolmogorov-Smirnov test for proportional hazard, the smooth test (Legendre polynomials with $d = 3$ with 3 degrees of freedom), data-driven smooth test (Legendre polynomials as the basis functions, nested with 5 dimensions) and the global test for all the interactions tested at once. Note that a *p*-value less than 0.05 indicates a violation of the proportionality assumption.

Both the smooth test of order 3 and the data-driven version fail to reject the null hypothesis, with *p*-values of 0.12 and 0.15, respectively, whereas the global test rejects the null hypothesis at $\alpha < 0.05$. On the other hand, the Kolmogorov-Smirnov test also fails to reject the null at $\alpha < 0.05$ but does not do well at $\alpha < 0.1$ (null hypothesis is rejected).
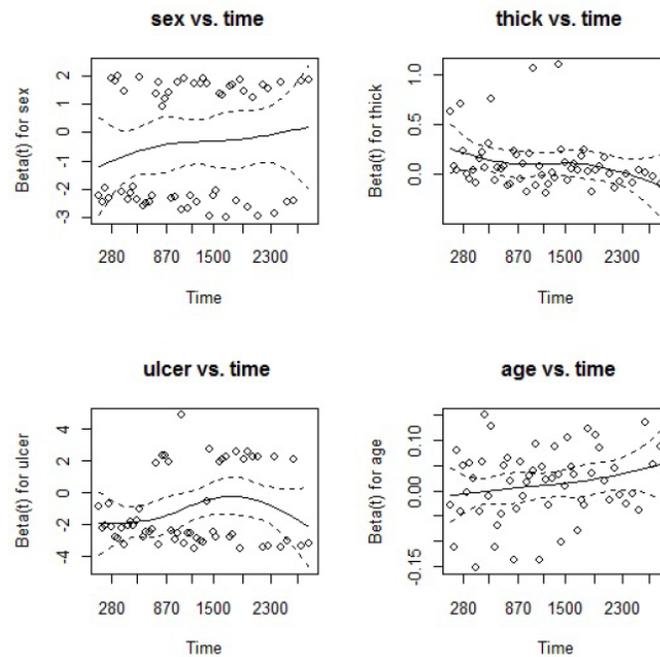
Figure 2: Schoenfeld residual plots for the covariates: Malignant Melanoma Data.

Table 2:   Tests of Proportionality: Malignant Melanoma Data

Test	Statisic	*p*-value
Global test	10.03	0.04
Two-sample Kolmogorov-Smirnov test	3.89	0.09
Smooth test of order 3	5.84	0.12
Data-driven Smooth test	2.28	0.15

4.3. Dataset 2: Cohort Study On Breast Cancer Patients From Netherlands

This dataset contains follow-up data on 2,982 women with breast cancer who went through breast surgery. The women were followed from the time of surgery until death, relapse or censoring. Only female patients diagnosed with primary epithelial breast cancer between 1 January 1990 and 31 December 2010 were selected from the Netherlands Cancer Registry (NCR). The register is a population-based independent cancer registry containing clinical administrative data of every newly diagnosed cancer patient in the Netherlands. Topography and morphology is coded according to the International Classification of Diseases for Oncology and staging according to the TNM-classification. Patients were included from hospitals in the Northern Netherlands and the Rotterdam region. Patients from hospitals from other regions that never participated before 2009 were included in the control group. Patients that were diagnosed with neuroendocrine tumors, synchronous tumors, diagnosed at autopsy and that had any type of previous malignancy were excluded. Hospitals from the intervention group were categorized by the implementation proportion (IP) of recommendations that were given in the final reports of each peer review. Rating the implementation was performed by studying final reports from subsequent reviews, follow-up correspondence, hospital documents and interviews with shareholders when necessary. Implementation of a recommendation was ranked on a scale from 0 to 4. The IP per hospital was expressed as a percentage of the total possible score. When implementation of a recommendation could not be determined (lost to follow-up), this recommendation was subtracted from the total possible score. The average IP of all peer reviews per hospital was used because it is not known what the time period is in which changes based on organizational change can occur and quality improvement is a continuous process. Ranking the implementation of recommendations was performed by the principal investigator and is described in [30].

We fitted a Cox PH model with age, menopausal status (meno), oestrogen receptors (er), differentiation grade (grade), number of nodes (nodes) and progesterone receptors (pr) as covariates.

Table 3:   Fitting of the CPH Model: Breast Cancer Data

Covariate	β	chisq	p-value
age	0.0519	4.28	3.85e-02
meno	0.0293	1.10	2.94e-01
er	0.1375	26.10	3.24e-07
grade	-0.0304	1.20	2.74e-01
nodes	-0.0569	2.54	1.11e-01
pr	0.1077	18.42	1.77e-05

From Table **3** above, all the covariates are significant at $\alpha = 0.5$. Figure **3** below shows the Schoenfeld residual plot for the overall fit. The solid line is a smoothing spline fit to the plot, with the broken lines representing a ± 2-standard-error band around the fit.
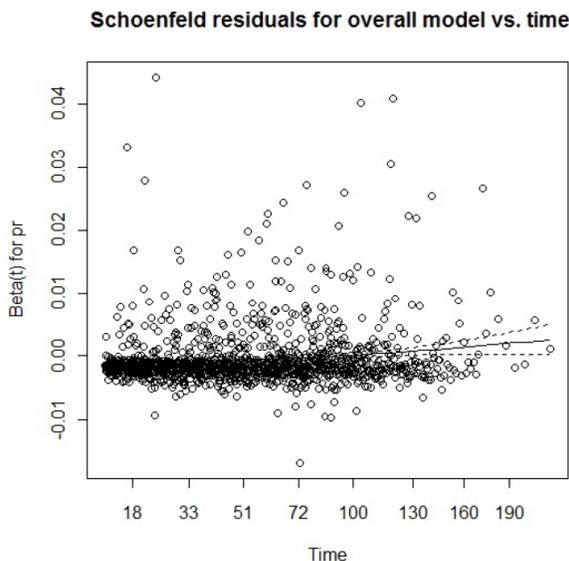

Figure 3: Schoenfeld residual plot for overall fit: Breast Cancer Patient Data.

The Schoenfeld residual plots show scatter plots with general non-zero slopes, indicating time-dependence (Figures **3** and **4**). The proportionality assumption does not hold in this dataset. Table **4** shows the proportionality tests for this dataset. We also compared the power of rejection for the Kolmogorov-Smirnov test for proportional hazard, the smooth test (Legendre $d = 3$ with 3 degrees of freedom), the data-driven smooth test (Legendre functions as basis, nested with 5 dimensions) and the global test for all the interactions.
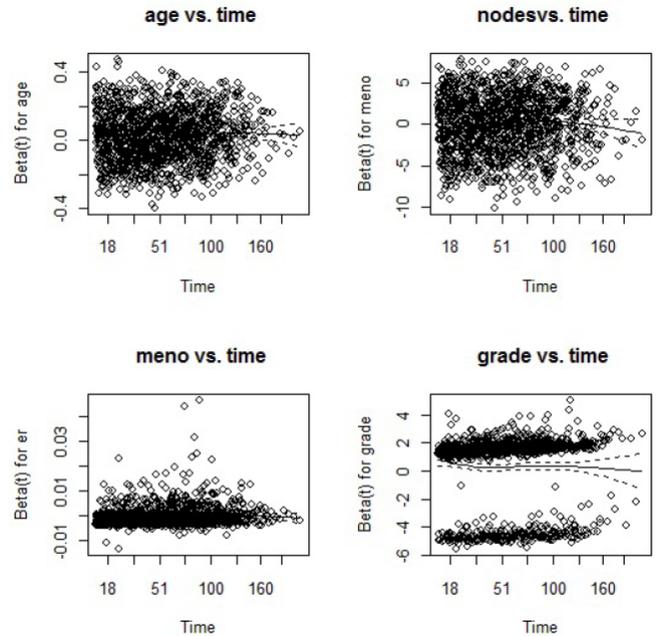

Figure 4: Schoenfeld residual plots for the covariates: Breast Cancer Patient Data.

Table 4:   Tests of Proportionality: Breast Cancer Patient Data

Test	Statisic	*p*-value
Global test	83.48	6.66e-16
Two-sample Kolmogorov-Smirnov test	29.63	0.01
Smooth test of order 3	17.73	0.00
Data-driven Smooth test	8.48	0.01

All the four tests are consistent in the rejection of the null hypothesis, which is supported by the Schoenfeld residual plots as well.

4.4. Dataset 3: Ovarian Cancer Survival Data

Between mid-1974 to mid-1977, 82 patients with advanced ovarian carcinoma and 29 patients with minimal residual disease were followed. Patients included in the minimal disease group had surgical excision of all turmor > 2 cm in diameter at the time of total abdominal hysterectomy, bilateral salpingo-oophorectomy and omentectomy within one month before enrolment. Following surgery, they were classified according to the distribution of residual diseases in arbitrary defined stages II to IIIA. All patients in each of the groups had histologically proved epithelial type ovarian carcinoma and all had adequate renal hepatic and marrow functions. The dataset is described in [31]. We fitted a Cox PH model for

censoring time (futime) and censoring status (fustat) with covariates age, ECOG performance status (ecog.ps), treatment group (rx) and residual disease present (resid.ds).

Table 5: Fitting the CPH Model: Ovarian Cancer Data

Covariates	β	chisq	*p*-value
age	-0.0399	0.0262	0.871
ecog.ps	0.4845	1.8819	0.170
rx	0.1325	0.2001	0.655
resid.ds	-0.1417	0.2463	0.620

From Table **5** above, all covariates are insignificant at $\alpha = 0.05$.

For testing the proportionality assumption, we plotted the Schoenfeld residual plot for the overall fit and each of the four covariates (Figures **5** and **6**).

Figure 5: Schoenfeld residual plot for overall fit: Ovarian Cancer Survival Data.

The Schoenfeld residual plots show non-zero slopes, suggesting time-dependence. The proportionality assumption holds in this dataset. Table **6** shows the proportionality tests for this dataset. We compared the power of rejection between Kolmogorov-Smirnov test for proportional hazard, the smooth test (Legendre $d = 3$ with 3 degrees of freedom), data-driven smooth test (Legendre functions as basis, nested with 5 dimensions) and the global test for all the interactions.

The global test and the smooth tests (fixed dimension and data-driven) fail to reject the null hypothesis. This is in agreement with the Schoenfeld residual plots for the general zero slope. However, the two-sample Kolmogorov-Smirnov test rejects the null hypothesis at $\alpha < 10\%$. This is misleading and inconsistent with the Schoenfeld residual plots.
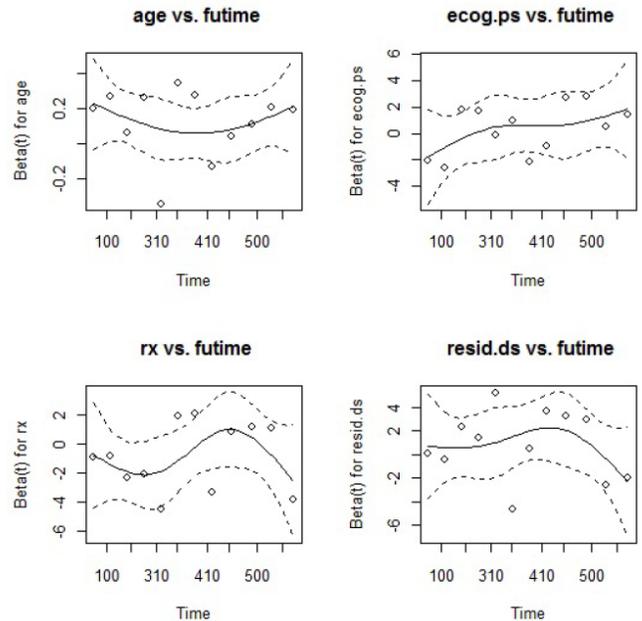

Figure 6: Schoenfeld residual plots for the covariates: Ovarian Cancer Survival Data.

Table 6: Tests of Proportionality: Ovarian Cancer Data

Test	Statisic	*p*-value
Global test	3.36	0.50
Two-sample Kolmogorov-Smirnov test	1.89	0.07
Smooth test of order 3	4.84	0.18
Data-driven Smooth test	2.23	0.20

4.5. Dataset 4: Remission Times for Acute Myelogenous Leukaemia

Acute myeloid leukemia (AML) represents a group of clonal hematopoietic stem cell disorders in which both a block in differentiation and unchecked proliferation result in the accumulation of myeloblasts at the expense of normal hematopoietic precursors. The patients in the study of maintenance therapy included 22 adults with AML, two with promyelocytic leukemia and two who had subacute myelogenous leukemia before conversion to classical AML. Patients had received no previous therapy for AML and there had been complete remission with standardized induction regimens supervised by the Stanford University Hematology Division. The median age of patients entered on the study was 45 years, with a

range of 18 to 72 years. The induction program was modified from the programs of Clarkson, Gee and colleagues by the addition of daunarubicin. With minor modifications, therapy was administered as follows: daunarubicin, 60 mg per sq meter by rapid intravenous infusion, was given on the first day. This was followed in 12 hours by cytarabine, 3 mg per kg of body weight by rapid intravenous infusion, and 6-thioguanine, 2.5 mg per kg of body weight given orally. Administration of the last two agents was continued every 12 hours until biopsy-proven marrow hypoplasia was achieved. A second dose of daunarubicin between days 7 and 10 was nearly always given, the dose varying, depending on the cellularity of a marrow biopsy specimen. Changes in therapy from the original program were undertaken so as to shorten the treatment program and decrease the time at risk from severe neutropenia and thrombopenia. That this was achieved is reflected in the shorter treatment period required to reach hypoplasia with the current drug program compared with earlier regimen employing only a single daily dose of cytarabine and 6-thioguanine. The question at the time was whether the standard course of chemotherapy should be extended ('maintenance') for additional cycles. The dataset is described in [32].

We fitted a Cox PH model for remission time and status with covariate X, representing 'maintenance' or non-maintenance' of patients in chemotherapy.

Table 7: Fitting the CPH Model: Acute Myelogenous Leukaemia Data

Covariates	β	chisq	p-value
x('non-maintenance)'	0.0198	0.00691	0.934

The covariate in this case ('non-maintenance') is not significant.

The Schoenfeld residual plots show a general zero-slope indicating proportionality. The proportionality tests are indicated in Table **8** below. We compared the power of rejection between Kolmogorov-Smirnov test for proportional hazard, the smooth test (Legendre $d = 3$ with 3 degrees of freedom), data-driven smooth test (Legendre functions as basis, nested with 5 dimensions). The global tests did not yield any result.

The global test did not give any result but the other 3 tests (i.e. the two-sample Kolmogorov-Smirnov test, smooth test and data-driven smooth test) failed rejects the null hypothesis. That is, they detected proportionality.
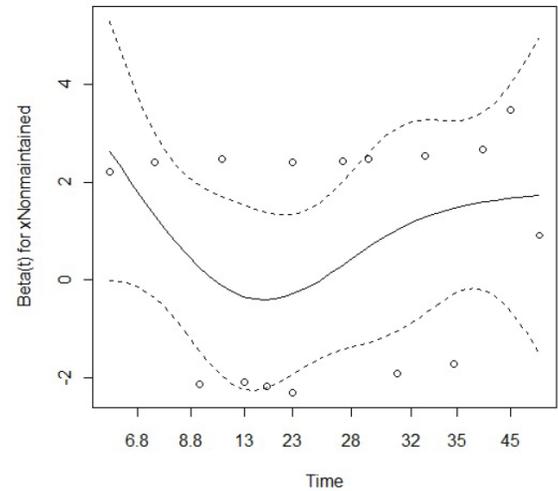

Figure 7: Schoenfeld residuals versus log(time) for overall fit: Acute Myelogenous Leukaemia Data.

Table 8: Tests of Proportionality: Acute Myelogenous Leukaemia Data

Test	Statisic	*p*-value
Global test	NA	NA
Two-sample Kolmogorov-Smirnov test	1.131	0.63
Smooth test of order 3	3.07	0.38
Data-driven Smooth test	0.15	0.75

4.6. Dataset 5: North Central Cancer Treatment Group Lung Cancer Data

This data shows survival of patients with advanced lung cancer from the North Central Cancer Treatment Group(NCCTG). The study looked at how performance scores can rate how well a patient performs usual daily activities. An initial detailed questionnaire was administered to approximately 150 patients with advanced cancer. This questionnaire was subsequently revised and given to a total of 1,115 patients with advanced colorectal or lung cancer. Thirty six variables showed significant prognostic information for survival in univariate analyses, even though many of these variables were associated with only a minimal increase in risk. A multivariate analysis demonstrated that there was a high correlation between many variables. Three major groups of variables became apparent as providing strong prognostic information (i.e. physician's assessment, patient's assessment and nutritional factor such as appetite). Data contained 228 patients with advanced lung cancer and includes measurements of the survival time in days, as well as other demographic and biological information for each patient. Variables such as weight loss were categorized by quartiles, and

ECOG scores were grouped into categories with subjects rated as either 0/1 or 2/3, with 0/1 representing the best and 2/3 representing a poor score. The data set was 28% censored, with a median observed failure time of 256 days. The baseline group (n = 16) were males with ECOG performance scores equal to 1 and a weight loss measure in the first quartile [33]. We fitted a Cox PH model with age, sex, ECOG performance score (ph.ecog), Karnofsky performance score rated by physician (ph.karno), Karnofsky performance score rated by patient (pat.karno), calories consumed at meals (meal.cal) and weight loss in last six months (wt.loss) as covariates.

Table 9:   Fitting Cox PH Model: NCCTG Lung Cancer Data

Covariates	β	chisq	*p*-value
age	0.0710	0.6553	0.4182
sex	0.1773	3.7609	0.0525
ph.ecog	-0.0189	0.0491	0.8247
ph.karno	0.1718	2.5791	0.1083
pat.karno	0.0298	0.1403	0.7080
meal.cal	0.1793	4.1493	0.0417
wt.loss	0.0764	1.0117	0.3145

From Table **9** above, meal.cal and sex are the only significant covariates. The other covariates are insignificant at $\alpha = 0.05$ significance level.
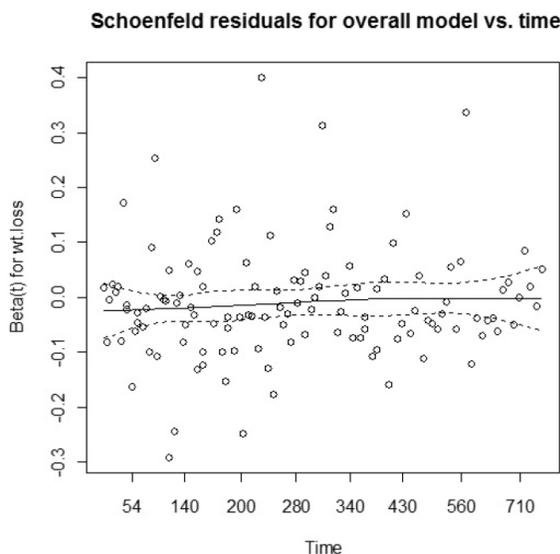

Figure 8: Schoenfeld residual plot for the overall fit: NCCTG Lung Cancer Data.

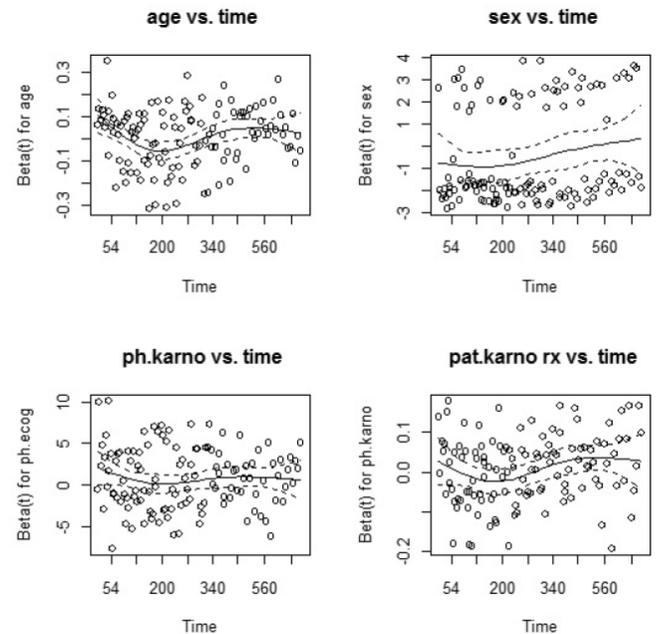The Schoenfeld residual plots for four covariates in the model including a lowess smoothing curve yields

Figure 9: Schoenfeld residual plots for the covariates: NCCTG Lung Cancer Data.

The overall Schoenfeld residual plot together with the plots for the four covariates show a zero slope, indicating proportionality. The hazard proportionality tests for the two-sample results are provided in Table **10** below. We compare the power of rejection between Kolmogorov-Smirnov test for proportional hazard, the smooth test (Legendre $d = 3$ with 3 degrees of freedom), data-driven smooth test (Legendre functions as basis, nested with 5 dimensions) and the global test.

Table 10: Tests of Proportionality in CPH: Lung Cancer Data

Test	Statisic	*p*-value
Global test	13.8	0.06
Two-sample Kolmogorov-Smirnov test	1.97	0.99
Smooth test of order 3	0.81	0.85
Data-driven Smooth test	0.01	0.96

Results show that the global test rejects the null hypothesis at $\alpha < 0.1$ but does well for $\alpha < 0.05$. The other three tests fail to reject the null hypothesis. This is coherent with the Schoenfeld residual plots.

4.7. Dataset 6: Stage C Prostate Cancer

Data contained 146 patients with stage C prostate cancer, from a study exploring the prognostic value of flow cytometry. Patients were followed and variables for time to progression or last follow-up (years) recorded. Other measurements were status (1=

progression observed, 0 = censored), age in years, status for endocrine captured (i.e. early endocrine therapy, 1 = no, 2 = yes), percent of cells in G2 phase, as found by flow cytometry, grade of the tumor, grade of the tumor, the ploidy status of the tumor, from flow cytometry, values for diploid, tetraploid, and aneuploid. A tumor was determined to be diploid (normal complement of dividing cells) if the fraction of cells in G2 phase was determined to be 13% or less. Aneuploid cells were given a measurable fraction with a chromosome count that is neither 24 nor 48. For these, the G2 percent is difficult or impossible to measure [34].

We fitted a Cox PH model for time and status with covariates age, early endocrine therapy (eet), percentage of cells in G2 phase (g2), grade of tumor by the Farrow system (grade), grade of tumor by the Gleason system (gleason), and the ploidy status of the tumor (diploid and tetraploid).

Table 11: Fitting Cox PH Model: Stage C Prostate Cancer Data

Covariates	β	chisq	p-value
age	0.0529	0.2361	0.6270
eet	-0.0406	0.1106	0.7395
g2	0.0134	0.0109	0.9170
grade	-0.0716	0.2733	0.6012
gleason	-0.1328	0.9858	0.3208
diploid	0.1311	1.2196	0.2694
tetraploid	0.1945	2.9799	0.0843

From Table **11** above all covariates are also insignificant at $\alpha = 0.05$. The Schoenfeld residual plot for the overall fit is shown in Figure **10** below.
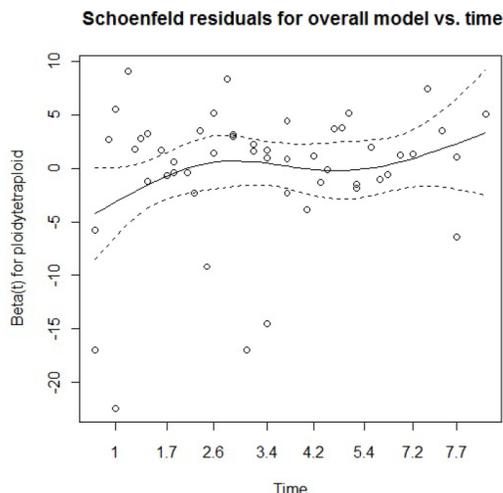

Figure 10: Schoenfeld residual plot for the overall fit: Stage C Prostate Cancer data.

We also created the Schoenfeld residual plots for the four covariates and fitted a lowess smoothing curve.
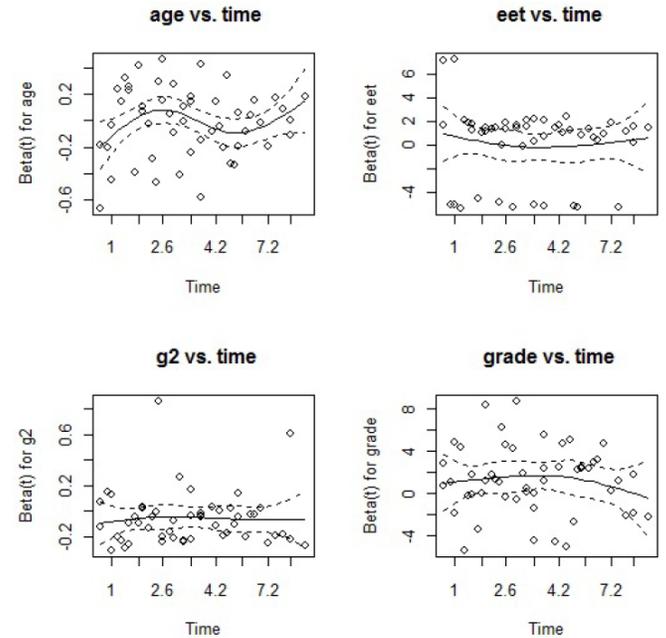

Figure 11: Schoenfeld residual plots for the covariates: Stage C Prostate Cancer data.

The Schoenfeld residual plots show a general slope of zero, indicating proportionality. The proportionality tests for the two-sample results are provided in Table **12** below. We compared the power of rejection between Kolmogorov-Smirnov test for proportional hazard, the smooth test (Legendre $d = 3$ with 3 degrees of freedom), the data-driven smooth test (Legendre functions as basis, nested with 5 dimensions) and the global test.

Table 12: Tests of Proportionality in CPH: Stage C Prostate Cancer Data

Test	Statisic	p-value
Global test	7.30	0.40
Two-sample Kolmogorov-Smirnov test	2.35	0.50
Smooth test of order 3	1.31	0.73
Data-driven Smooth test	1.12	0.31

Results show that all the tests are consistent and fail to rejects the null hypothesis at $\alpha < 0.05$.

4.8. Dataset 7: Chemotherapy for Stage B/C Colon Cancer Data

This was a national intergroup trial that was sponsored by the National Cancer Institute and

involved the Eastern Cooperative Oncology Group, the NCCTG, the Southwest Oncology Group, and the Mayo Clinic. Enrollment of patients started in March 1984, when a preliminary analysis of the NCCTG study indicated the likelihood of a treatment advantage for levamisole plus fluorouracil and for levamisole alone, with regard to time-to-recurrence. Enrollment was completed in October 1987. All patients were required to have undergone a potentially curative adenocarcinoma of the colon without gross or microscopic evidence of residual disease. Patients with rectal carcinoma were excluded from the study. The resected specimen in eligible patients showed one of two indicators of poor prognosis - invasion extending at least to the serosa or pericolonic fat (Stage B2) or metastasis to regional lymph nodes (Stage C). It was further required that the patient be able to swallow oral medication and have a leukocyte count of at least 4000 per microliter and a platelet count of at least 130,000 per microliter. Eligibility was determined by careful review of study forms, operative reports, and pathology reports. Entry into the study was allowed no earlier than one week and no later than five weeks after surgery. These are data from one of the first successful trials of adjuvant chemotherapy for colon cancer. Levamisole is a low-toxicity compound previously used to treat worm infestations in animals. There are two records per person, one for recurrence and one for death [35].

We fitted the Cox PH model for time and status with covariates age, sex, Levamisole (rxLev), Levamisole+5-FU (rxLev+5FU), obstruction of colon by tumor (obstruct), adherence to nearby organs (adhere),

Table 13: Fitting Cox PH Model: Stage B/C Colon Cancer Data

Covariates	β	chisq	p-value
age	-0.0107	0.111	7.39e-01
sex	0.0563	2.869	9.03e-02
rxLev	-0.0475	2.068	1.50e-01
rxLev+5FU	-0.0198	0.355	5.51e-01
obstruct	-0.1093	11.113	8.57e-04
adhere	0.0470	2.032	1.54e-01
differ	-0.1480	22.515	2.08e-06
extent	-0.0364	1.229	2.68e-01
surg	0.0135	0.166	6.84e-01
node4	-0.1121	10.788	1.02e-03
etype	0.3561	112.926	0.00e+00

differentiation of tumor (differ), extent of local spread(extent), time from surgery to registration (surg), more than 4 positive lymph nodes (node4) and event type (etype).

Results from Table **13** show that all the selected covariates are also significant at $\alpha = 0.05$. For testing proportionality the Schoenfeld residual plot for the overall fit is shown in Figure **12** below.
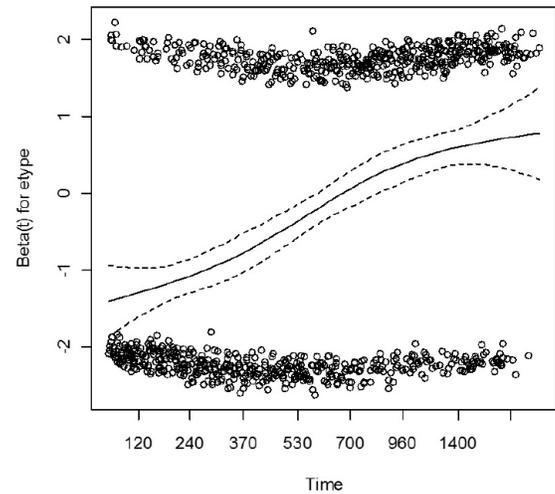

Figure 12: Schoenfeld residual plot for the overall fit: Stage B/C Colon Cancer Data.

The Schoenfeld residual plots for the four selected covariates in the model are shown in Figure **13**.
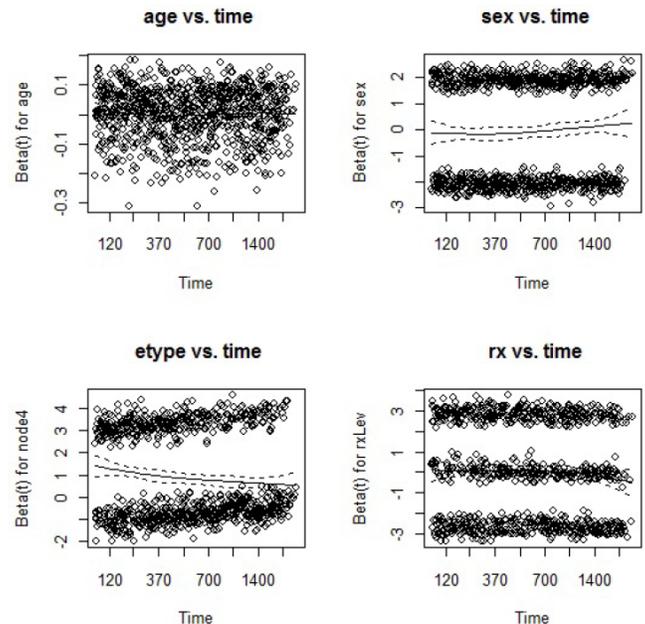

Figure 13: Schoenfeld residual plots for four selected covariates: Stage B/C Colon Cancer Data.

The Schoenfeld residual plots show a general non-zero slope, indicative of non-proportionality. The

proportionality tests for the two-sample results in are indicated in Table **14** below. We compared the power of rejection between Kolmogorov-Smirnov test for proportional hazard, the smooth test (Legendre $d = 3$ with 3 degrees of freedom), the data-driven smooth test (Legendre functions as basis, nested with 5 dimensions) and the global test.

Table 14: Tests of Hazard Proportionality in CPH: Colon Cancer Data

Test	Statisic	p-value
Global test	161.05	0.00
Two-sample Kolmogorov-Smirnov test	73.17	< 2.22e-16
Smooth test of order 3	97.76	< 2.22e-16
Data-driven Smooth test	97.21	< 2.22e-16

Results show that all the tests are consistent in rejecting the null hypothesis. Despite the fact that all covariates incorporated in this model are significant, proportionality does not hold. The covariates are therefore time dependent $\alpha < 0.05$.

4.9. Dataset 8: Veteran Administration Lung Cancer Study

The study population consisted of 109 patients with newly diagnosed Small Cell Lung Cancer (SCLC) investigated at the Pulmonary Division of Mainz University Hospital between 1989 and 1999. Clinical data were collected from chart review. The staging procedure for the majority of patients was standardized including a fiberoptic bronchoscopy, routine laboratory parameters, chest CT, abdomen CT and bone scan. In 89% of the patients chemotherapy was performed as first-line treatment. Three different standard combinations were applied with a median of four cycles. Response was first evaluated after two cycles of chemotherapy and every second cycles thereafter or if new clinical symptoms occurred. Response to chemotherapy was classified according to the WHO criteria in complete response, partial response, stable disease or progressive disease. Complete response was achieved in 24% of patients, partial response in 29%, and stable disease in 5% of patients. 42% of patients progressed during therapy. In 35% of all patients chemotherapy was followed by radiotherapy of the primary tumor. From all subjects four patients with complete response underwent surgical resection of the primary tumor side. The majority of patients were followed-up regularly in a time frame of 2 to 3 months.

The survival time was calculated from the date of histological diagnosis [36].

We fitted the Cox PH Model for time and status with covariates age, Karnofsky performance score (karno), treatment (trt), months from diagnosis to randomization (diagtime), prior therapy (prior) and celltype (small-cell, adeno, and large).

Table 15: Fitting Cox PH Model: Veteran Administration Lung Cancer Study

Covariates	β	chisq	p-value
age	0.1890	5.3476	0.020750
karno	0.3073	13.0449	0.000304
trt	-0.0273	0.1227	0.726104
diagtime	0.1491	2.9436	0.086217
prior	-0.1767	4.4714	0.034467
smallcell	0.0128	0.0261	0.871621
adeno	0.1424	2.9794	0.084329
large	0.1712	4.1093	0.042649

From Table **15** above some covariates (i.e. large, prior, karno and age) were significance at $\alpha < 0.05$. The other covariates were not significant at $\alpha = 0.05$. The Schoenfeld residual plot for the overall fit is shown in Figure **14** below.
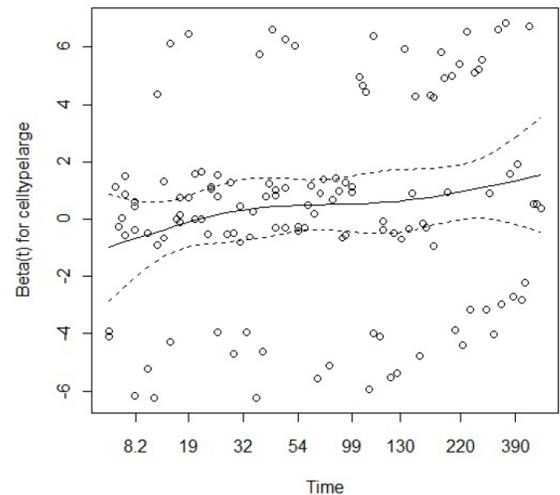

Figure 14: Schoenfeld residual plot for the overall fit: Veteran Administration Lung Cancer study.

The Schoenfeld residual plots for four selected covariates, including lowess smoothing curves, are shown in Figure **15** below.

The Schoenfeld residual plots show a non-zero slope, except prior, thereby indicating proportionality.

The proportionality tests for the two-sample results are indicated in Table **16** below. We compared the power of rejection between Kolmogorov-Smirnov test for proportional hazard, the smooth test (Legendre $d = 3$ with 3 degrees of freedom), the data-driven smooth test (Legendre functions as basis, nested with 5 dimensions) and the global test.
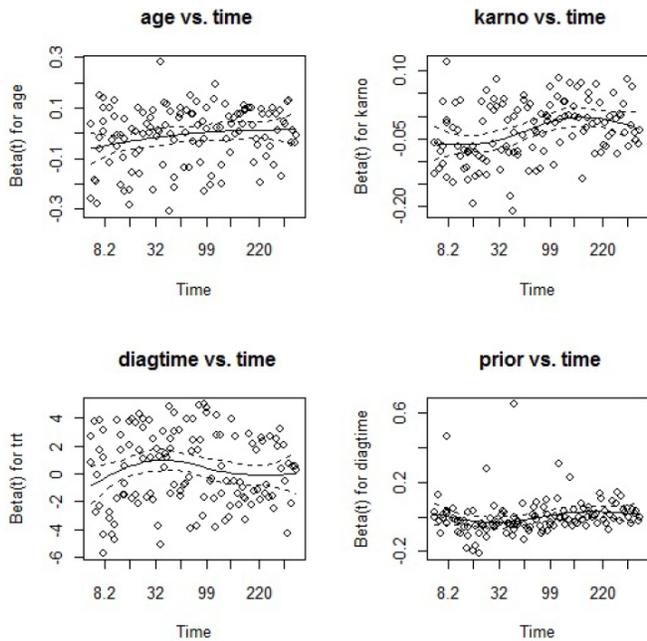

Figure 15: Schoenfeld residual plots for four selected covariates: Veteran Administration Lung Cancer study.

Table 16: Tests of Proportionality in CPH Model: Lung Cancer Data

Test	Statisic	*p*-value
Global test	28.00	0.00
Two-sample Kolmogorov-Smirnov test	9.09	0.06
Smooth test of order 3	6.84	0.08
Data-driven Smooth test	2.03	0.18

In this case the global test strongly rejects proportionality, whereas the two-sample Kolmogorov-Smirnov test and smooth test of order 3 rejects proportionality at $\alpha < 0.10$. The data-driven version of the smooth test however remains stable and fails to reject the null, an indication of proportionality. Table **17** below provides a concise summary of the results from the analysis of the eight datasets.

5. DISCUSSIONS

The CPH model is commonly used to determine risk factors. The assumption of proportional hazards is therefore important whenever the model is applied. Numerous methods for assessing the assumption of hazard proportions have been proposed. These methods (e.g. global test, G-test, Kolmogorov-Smirnov test, smooth test etc.) together with their asymptotic properties, have been studied theoretically by several authors. However, validation of these tests in light of real settings have generally utilized either none or at most two real datasets. Furthermore, the combined use of graphical and non-graphical analysis, which is one of the contributions of this manuscript, have been studied comparatively by few authors. Also, in practice there exist variations in real data settings, particularly in cancer studies and validations of these tests in multiple settings have not been done. Moreover, most researchers, usually fit CPH models using several explanatory variables in order to identify risk factors. However, in the fitted CPH model, the covariates included in the model should satisfy the assumption that the relative risk is proportional over the time for different levels.

This study sought to validate the performance of the smooth test in different cancer settings and compare with that of the global goodness-of-fit test and Kolmogorov-Smirnov proportional hazard test. In particular, we assessed the performance of these tests under different cancer study settings when testing for the PH assumption. For each of the eight datasets, we have displayed the projected hazard plots together with their $\log(\beta(t))$ projection. We chose graphs that are based on the Schoenfeld residuals because they are more robust compared to Kaplan-Meier (K-M) survival curves. Furthermore, the K-M curves with fewer time points are usually not straight-forward when detecting proportionality. In these cases, the resulting power of rejection is compared with the graphical presentation. Whereas there are certain types of non-proportionality that cannot be detected by the tests of non-zero slopes alone, it becomes obvious when looking at the graphs of the residuals to see a nonlinear relationship between the residuals and the function of time. In this regard, the behavior of smooth tests is similar to the other tests if we have a "sizeable" sample size.

The two versions of smooth tests provide a procedure with power that is more stable than the other methods. The smooth test is analyzed with a fixed dimension of order 3 with 3 degrees of freedom. For the data-driven version of the smooth test, we nested subsets in order to avoid the use of many components. The nested subsets selection procedure is not sensitive with respect to the choice of the maximum dimension

Table 17: Summary of Analyses of the Datasets when Testing Proportionality (H_0)

Dataset	Article	Schoenfeld residual plots	Global test	2-sample K-S test	Smooth test (order 3)	Data-driven smooth test
Dataset 1	Andersen *et al*. (2012)	zero line slope	rejects H_0	fails to reject H_0	fails to reject H_0	fails to reject H_0
Dataset 2	Royston *et al*. (2011)	non-zero slope	rejects H_0	rejects H_0	rejects H_0	rejects H_0
Dataset 3	Edmonson *et al*. (1979)	zero slope	fails to reject H_0	fails to reject H_0	fails to reject H_0	fails to reject H_0
Dataset 4	Embury *et al*. (1977)	zero slope	NA	fails to reject H_0	fails to reject H_0	fails to reject H_0
Dataset 5	Loprinzi *et al*. (1994)	zero slope	marginal	fails to reject H_0	fails to reject H_0	fails to reject H_0
Dataset 6	Zagars *et al*. (1993)	non-zero slope	fails to reject H_0	fails to reject H_0	fails to reject H_0	fails to reject H_0
Dataset 7	Moertel *et al*. (1990)	non-zero slope	rejects H_0	rejects H_0	rejects H_0	rejects H_0
Dataset 8	Micke *et al*. (2002)	zero slope	rejects H_0	marginal	fails to reject H_0	fails to reject H_0

(*d*) if *d* is large enough to cover realistic departures from the hypothesis (see [21]). Similarly, in all the Figures **1-15**, all the tests detect the non-proportionality in log (protime), but only the tests based on the score process detect non-proportionality in most covariates. Since we have analyzed these datasets in order to validate the smooth test in different settings, we have utilized Schoenfeld residuals plots to see if either proportionality does or does not hold for log (time). We have consequently applied the four tests to determine consistency between the plots and the level of rejection of null hypothesis (proportionality). We have done this simultaneously for comparison and verification of results obtained with both the graphical and data analysis. The results of applying the tests to the eight cancer datasets are reported in each subsection under the Methods section. The eight datasets were obtained from already published articles and are readily accessible in R. For the graphical representation, testing the time-dependent covariates is equivalent to testing for a non-zero gradient. Therefore if the proportional hazards assumption is true, beta (t) will be a horizontal line along $\beta(t) = 0$.

All analyses were performed using the $coxph$, $cox.zph$, $survival$ packages in R. In each situation we described the general setting of the study and the power of each test is computed. A CPH model was then fitted to the data, using forward selection procedure that ended up including as many covariates as possible into the model. It is important to note that our interest was not how good covariates fit in CPH model, but how accurate the hazard proportionality assumption is determined. Then the Schoenfeld residuals for overall and consequently four covariates were studied.

Dataset 1

The setting here is malignant melanoma with 205 patients. The overall Schoenfeld residual plot shows zero line slope. Further, the Schoenfeld residual plots for all the four covariates, except sex, show zero-slope. Analytical results on the other hand show smooth tests fail to reject the null hypothesis($p = 0.12$ and $p = 0.15$, respectively), whereas the global test rejects the null hypothesis at $\alpha < 0.05$. On the other hand, the Kolmogorov-Smirnov test also fails to reject the null at $\alpha < 0.05$ but does not do well at $\alpha < 0.1$ (null hypothesis is rejected). The smooth test in this setting does better than the other two tests in determining hazard proportionality. The smooth test is generally, coherent with the Schoenfeld residuals plots.

Dataset 2

For the cohort study on breast cancer patients analyzed here, the overall Schoenfeld residual plot and for three of the selected four covariates depict non-zero slope. This is an indication of time-dependent covariates. In this setting, the sample size is also significantly large ($n = 2,982$). Results show all the four tests are consistent in rejecting the null hypothesis, which is consistent with Schoenfeld residual plots.

Dataset 3

The setting here is ovarian cancer with 82 patients being observed. Results show that the global test and the smooth test fail to rejects the null hypothesis. This is in agreement with the Schoenfeld residual plots which show a general zero slope. However, the two-sample Kolmogorov-Smirnov test rejects the null hypothesis at $\alpha < 0.10$. The Kolmogorov-Smirnov test will still be consistent with the other two test at $\alpha < 0.05$ but may be misleading at $\alpha < 0.1$.

Dataset 4

The setting here involves patients with clonal hematopoietic stem cell disorder (acute myeloid leukemia). With a sample size of 23 and one covariate, the global test did not give any result but the other 3 tests (i.e. two-sample Kolmogorov-Smirnov test, smooth test and data-driven smooth test) failed to reject the null hypothesis. Despite the fact that the covariate was insignificant ($\chi = 0.00691, p = 0.934$), our interest was to check proportionality and not the best fit and it is after ascertaining proportionality assumption that we can objectively say that the covariate is insignificant. This is an indication that researchers can utilize the smooth test and two-sample Kolmogorov-Smirnov test whenever other variables (e.g. sample size, number of covariate etc.) are not appropriate in the global test.

Dataset 5

Survival data in this setting involves 228 patients with advanced lung cancer. The overall Schoenfeld residual plot shows a zero slope. Three of the four selected covariates (age, ph.karno and pat.karno rx) also show zero slope. Results show the global test rejects the null hypothesis at $\alpha < 0.1$ but not at $\alpha < 0.05$. The other three tests fail to reject the null hypothesis. This is consistent with the Schoenfeld residual plots. It is also an indication that the global test may not be accurate.

Dataset 6

Data analyzed in this setting involved 146 patients with stage C prostate cancer. All the seven covariates were statistically insignificant at $\alpha < 0.05$. The overall Schoenfeld residual plot depicts a zero slope. However, only one (eet) of the four selected covariates showed a non-zero slope. Analytical result showed that all the tests are consistent and fail to rejects the null hypothesis at $\alpha < 0.05$.

Dataset 7

This is the setting of national intergroup trial and involved 1,858 patients with stage B and stage C colon cancer. Results show that all the tests are consistent in rejecting the null hypothesis, which is consistent with both the overall Schoenfeld residual plot and the four selected covariates Schoenfeld residual plots. Despite the fact that all covariates incorporated in this model are significant, proportionality does not hold. The covariates are therefore time-dependent $\alpha < 0.05$.

Dataset 8

The study setting here involves a population consisting of 109 patients with small cell lung cancer. The overall Schoenfeld residual plot depicts a zero slope. Only one covariate (prior) of the four selected covariates shows non-zero slope. In this case the global test strongly rejects proportionality, whereas the two-sample Kolmogorov-Smirnov test and the smooth test of order 3 rejects the null (proportionality) at $\alpha < 0.10$. The data-driven version of the smooth test however remains stable and fails to reject the null hypothesis, an indication of proportionality. This is a situation where data-driven smooth test performs better than the other tests.

Analysis of cancer data showed that the smooth test and its data-driven version are stable compared to the global and the Kolmogorov-Smirnov tests when assessing the proportional hazards assumption in variety of practical settings. Furthermore, although the smooth test does not universally dominate the other two tests in different cancer study settings, it remains relatively stable irrespective of the sample size and the number of covariates. The application of the smooth test and its data-driven version to assess proportionality illustrates how the global test and Kolmogorov-Smirnov test inadequacies can result in invalid models. We therefore implore researchers to use smooth tests of goodness-of-fit whenever Schoenfeld residual plots conflicts with the global test and Kolmogorov-Smirnov test.

6. CONCLUSION

The smooth test for proportional hazard assumption in two-sample problem is revisited in light of real cancer datasets. We have shown that the smooth test is the "gold standard" for testing proportionality. The smooth test is robust and has better power against the other two tests when detecting departure from proportionality
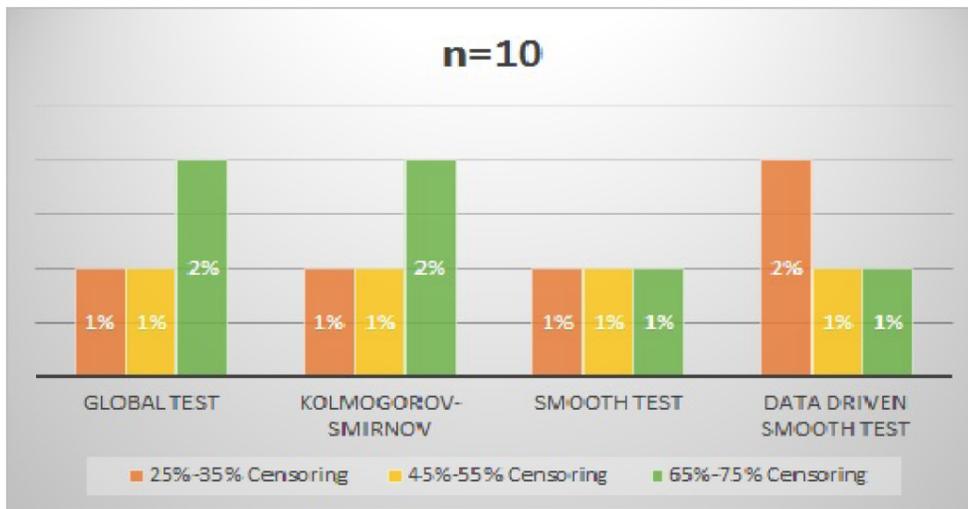
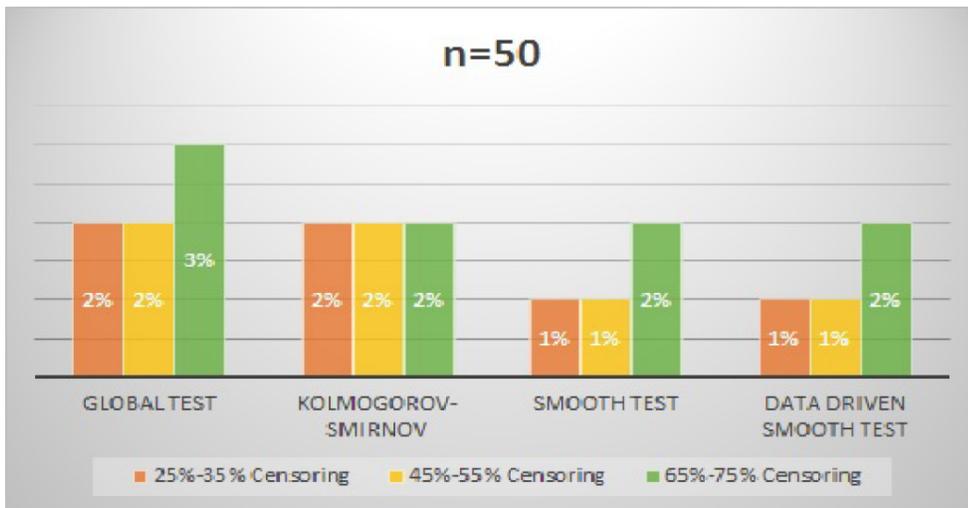**Figure 16:** Percentage rejection when n=10 at varying censoring percentages.

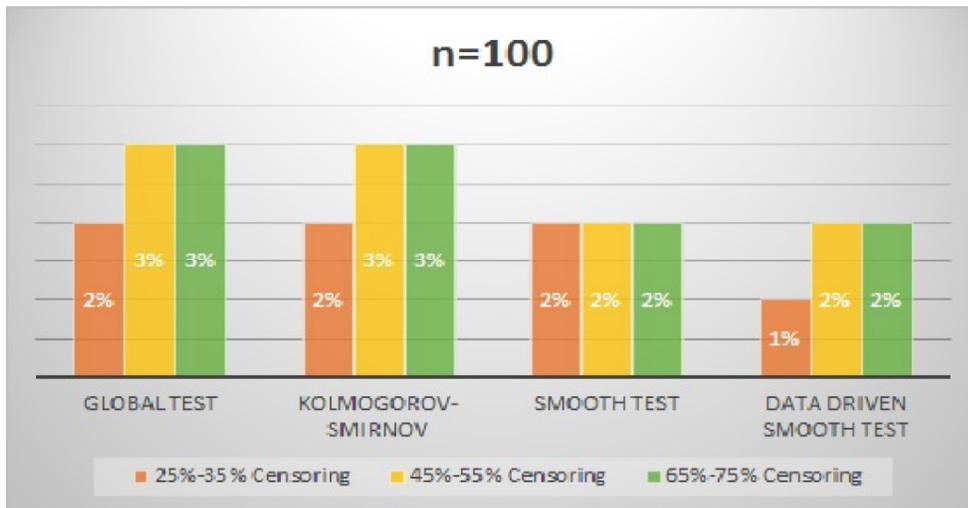**Figure 17:** Percentage rejection when n=50 at varying censoring percentages.

Figure 18: Percentage rejection when n=100 at varying censoring percentages.
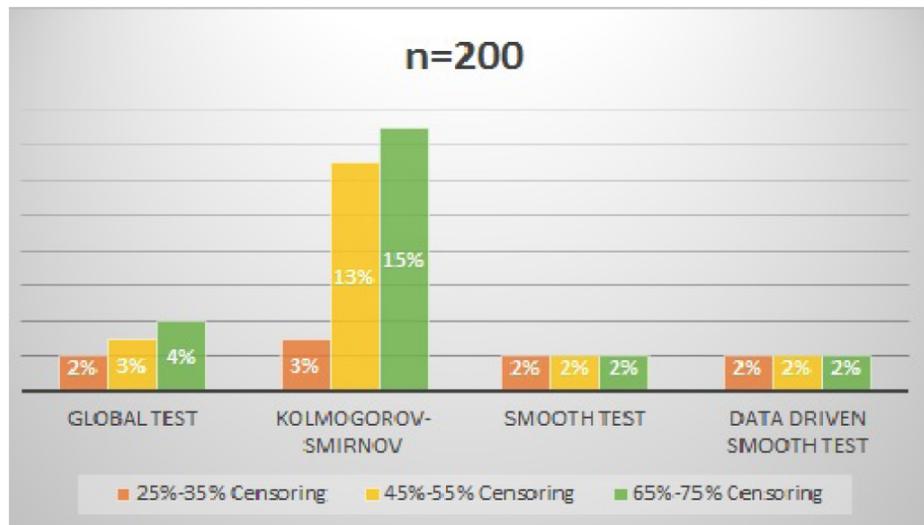
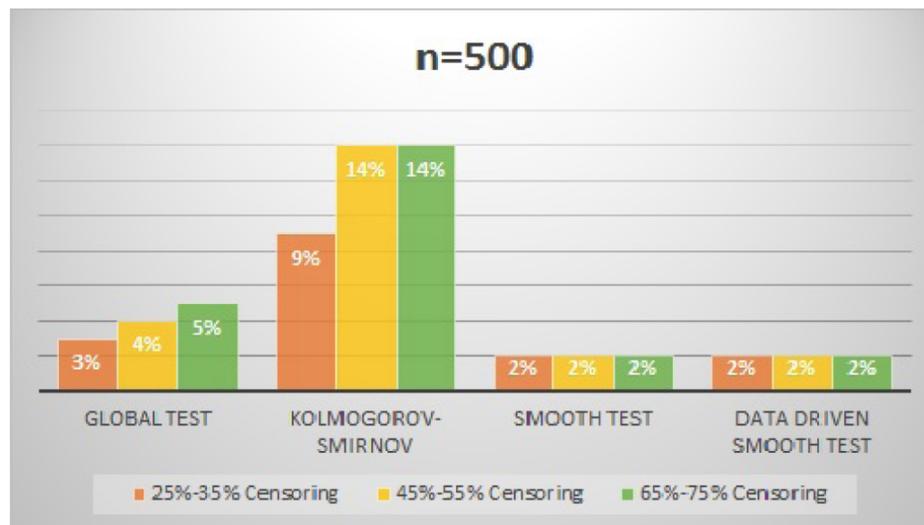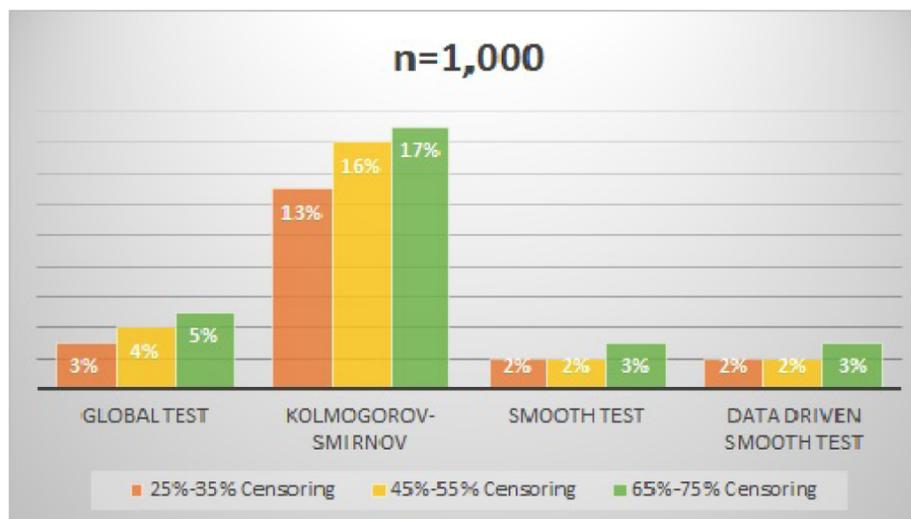**Figure 19:** Percentage rejection when n=200 at varying censoring percentages.

Figure 20: Percentage rejection when n=500 at varying censoring percentages.

Figure 21: Percentage rejection when n=1000 at varying censoring percentages.

under different practical settings. The limitation for this study is that the smooth test is not capable to distinguish which covariates are proportional and which are not [19].

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR'S CONTRIBUTIONS

BO conceived and supervised the study. CO performed the data analysis. BO and CO wrote the manuscript. BO and JO critically reviewed and discussed the manuscript. All authors read and approved the final manuscript.

BO was partially supported by the Simons Foundation (# 282174).

REFERENCES

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA: a Cancer Journal for Clinicians 2016; 66(1): 7-30.
https://doi.org/10.3322/caac.21332

[2] Persson I. Essays on the assumption of proportional hazards in Cox regression. Acta Universitatis Upsaliensis 2002.

[3] David CR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society 1972; 34: 187-220.

[4] Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival models. Biometrika 1990; 147-160.
https://doi.org/10.1093/biomet/77.1.147

[5] Barlow WE, Prentice RL. Residuals for relative risk regression. Biometrika 1988; 65-74.
https://doi.org/10.1093/biomet/75.1.65

[6] Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika 1982; 239-241.
https://doi.org/10.1093/biomet/69.1.239

[7] Kay R. Proportional hazard regression models and the analysis of censored survival data. Applied Statistics 1977; 227-237.
https://doi.org/10.2307/2346962

[8] Quantin C, Moreau T, Asselain B, Maccario J, Lellouch J. A regression survival model for testing the proportional hazards hypothesis. Biometrics 1996; 874-885.
https://doi.org/10.2307/2533049

[9] Lin D. Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. Journal of the American Statistical Association 1991; 86(415): 725-728.
https://doi.org/10.1080/01621459.1991.10475101

[10] O'Quigley J, Pessione F. Score tests for homogeneity of regression effect in the proportional hazards model. Biometrics 1989; 135-144.
https://doi.org/10.2307/2532040

[11] Nagelkerke N, Oosting J, Hart A. A simple test for goodness of fit of Cox's proportional hazards model. Biometrics 1984; 483-486.
https://doi.org/10.2307/2531400

[12] Wei L. Testing goodness of fit for proportional hazards model with censored observations. Journal of the American Statistical Association 1984; 79(387): 649-652.
https://doi.org/10.1080/01621459.1984.10478092

[13] Andersen PK. Testing goodness of fit of Cox's regression and life model. Biometrics 1982; 67-77.
https://doi.org/10.2307/2530289

[14] Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. Biometrika 1980; 67(1): 145-153.
https://doi.org/10.1093/biomet/67.1.145

[15] Neyman J. << Smooth test >> for goodness of fit. Scandinavian Actuarial Journal 1937; 1937(3-4): 149-199.
https://doi.org/10.1080/03461238.1937.10404821

[16] Pena EA. Smooth goodness-of-fit tests for the baseline hazard in Cox's pro-portional hazards model. Journal of the American Statistical Association 1998; 93(442): 673-692.

[17] Pena EA, *et al*. Smooth goodness-of-fit tests for composite hypothesis in hazard based models. The Annals of Statistics 1998; 26(5): 1935-1971.
https://doi.org/10.1214/aos/1024691364

[18] Kraus D. Data-driven smooth tests of the proportional hazards assumption. Lifetime Data Analysis 2007; 13(1): 1-16.
https://doi.org/10.1007/s10985-006-9027-8

[19] Kraus D. Neymans smooth tests in survival analysis. PhD thesis. Charles University in Prague, Dept. of Statistics 2007.

[20] Escanciano JC, Mayoral S. Data-driven smooth tests for the martingale difference hypothesis. Computational Statistics & Data Analysis 2010; 54(8): 1983-1998.
https://doi.org/10.1016/j.csda.2010.02.023

[21] Kraus D. Adaptive Neyman's smooth tests of homogeneity of two samples of survival data. Journal of Statistical Planning and Inference 2009; 139(10): 3559-3569.
https://doi.org/10.1016/j.jspi.2009.04.009

[22] Kang SI. Performance of generalized Neyman smooth goodness of fit tests 1978.

[23] Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika 1993; 80(3): 557-572.
https://doi.org/10.1093/biomet/80.3.557

[24] Bagdonavičius V, Nikulin M. On goodness-of-fit for the linear transformation and frailty models. Statistics & Probability Letters 2000; 47(2): 177-188.
https://doi.org/10.1016/S0167-7152(99)00155-8

[25] Ledwina T. Data-driven version of Neyman's smooth test of fit. Journal of the American Statistical Association 1994; 89(427): 1000-1005.
https://doi.org/10.1080/01621459.1994.10476834

[26] Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika 1993; 80(3): 557-572.
https://doi.org/10.1093/biomet/80.3.557

[27] Lin D. On the Breslow estimator. Lifetime data analysis 2007; 13(4): 471-480.
https://doi.org/10.1007/s10985-007-9048-y

[28] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994; 515-526.
https://doi.org/10.1093/biomet/81.3.515

[29] Andersen PK, Borgan O, Gill RD, Keiding N. Statistical models based on counting processes. Springer Science & Business Media 2012.

[30] Royston P, Lambert PC, *et al.* Flexible parametric survival analysis using Stata: beyond the Cox model 2011.

[31] Royston P, Lambert PC, *et al.* Flexible parametric survival analysis using Stata: beyond the Cox model 2011.

[32] Edmonson JH, Fleming TR, Decker D, Malkasian G, Jorgensen E, Jefferies J, *et al*. Different chemotherapeutic sensitivities and host factors affecting prog-nosis in advanced ovarian carcinoma versus minimal residual disease. Cancer Treatment Reports 1979; 63(2): 241-247.

[33]   Loprinzi CL, Laurie JA, Wieand HS, Krook JE, Novotny PJ, Kugler JW, *et al*. Prospective evaluation of prognostic variables from patient-completed question- naires. North Central Cancer Treatment Group. Journal of Clinical Oncology 1994; 12(3): 601-607.
https://doi.org/10.1200/JCO.1994.12.3.601

[34]   Zagars GK, Von Eschenbach AC, Ayala AG. Prognostic factors in prostate cancer: analysis of 874 patients treated with radiation therapy. Cancer 1993; 72(5): 1709-1725.
https://doi.org/10.1002/1097-0142(19930901)72:5<1709::AID-CNCR2820720535>3.0.CO;2-O

[35]   Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Goodman PJ, *et al*. Levamisole and uorouracil for adjuvant therapy of resected colon carcinoma. New England Journal of Medicine 1990; 322(6): 352-358.
https://doi.org/10.1056/NEJM199002083220602

[36]   Micke P, Faldum A, Metz T, Beeh KM, Bittinger F, Hengstler JG, *et al.* Staging small cell lung cancer: Veterans Administration Lung Study Group versus International Association for the Study of Lung Cancer - what limits limited disease? Lung Cancer 2002; 37(3): 271-276.
https://doi.org/10.1016/S0169-5002(02)00072-7