Lifetime Data Analysis, 10, 139–157, 2004

Tests for the Proportional Intensity Assumption Based on the Score Process

JAN TERJE KVALØY jan.t.kvaloy@tn.his.no
Department of Mathematics and Natural Science, Stavanger University College, P.O. Box 8002, N-4068 Stavanger, Norway

LINDA REIERSØLMOEN NEEF* linda.neef@nr.no
Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Received October 3, 2002; Revised August 14, 2003; Accepted October 2, 2003

Abstract. Proportional intensity models are widely used for describing the relationship between the intensity of a counting process and associated covariates. A basic assumption in this model is the proportionality, that each covariate has a multiplicative effect on the intensity. We present and study tests for this assumption based on a score process which is equivalent to cumulative sums of the Schoenfeld residuals. Tests within principle power against any type of departure from proportionality can be constructed based on this score process. Among the tests studied, in particular an Anderson–Darling type test turns out to be very useful by having good power properties against general alternatives. A simulation study comparing various tests for proportionality indicates that this test seems to be a good choice for an omnibus test for proportionality.

Keywords: tests for proportional hazards, multiplicative intensity, counting process, nonmonotonic effects, censored data

1. Introduction

A widely used model for the conditional intensity of a counting process is the proportional intensities model, which can be written

$$\lambda(t) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{X}). \tag{1}$$

Here $\lambda_0(t)$ is an unspecified baseline intensity function, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ a vector of unknown regression parameters and $\mathbf{X} = (X_1, \ldots, X_p)^T$ a vector of covariates. The covariate vector $\mathbf{X}$ can generally be time-dependent, but for the problem considered in this paper we will assume that the covariates are constant in time.

One of the main assumptions in the proportional intensities model (1) is of course the proportionality, that the ratio of two intensities is constant in time,

$$\frac{\lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{X}_i)}{\lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{X}_j)} = \exp(\boldsymbol{\beta}^T(\mathbf{X}_i - \mathbf{X}_j)),$$

*Present address: Norwegian Computing Center, P.O. Box 114 Blindern, N-0314 Oslo, Norway

or that the impact of a covariate vector on the intensity is the same multiplicative factor $\exp(\boldsymbol{\beta}^T\mathbf{X})$ at any time. Note that this proportionality property also holds separately for each single covariate in the covariate vector. The proportionality assumption is a strong assumption which is not always necessarily reasonable and thus needs to be checked. One example of a situation where the assumption might not hold is in medical data with an initial treatment effect that fades over time. See for instance Therneau and Grambsch (2000, ch. 6.6) for a further discussion of causes for nonproportionality.

Several approaches for checking the proportionality assumption exist. A commonly used simple graphical method is based on dividing the data into $S \geq 2$ strata according to levels of the value of the covariate which the proportionality assumption is being checked for. Letting $\Lambda(t) = \int_0^t \lambda(u)\,du$ we get from (1) that

$$\log(\Lambda(t)) = \log(\Lambda_0(t)) + \boldsymbol{\beta}\mathbf{X}.$$

This shows that for a stratified model where a separate $\Lambda_{0s}(t)$ is fitted in each strata $s = 1, \ldots, S$, the estimated $\log(\Lambda_{0s}(t))$ for each strata should be parallel if proportionality holds for the examined covariate. Thus plotting all estimated $\log(\Lambda_{0s}(t))$ against t or some other function of time, for instance $\log(t)$, should give approximately parallel lines if proportionality holds. This is a simple and straightforward approach for checking the proportionality assumption, but there are several drawbacks. It is, for instance, unclear how parallel the lines need to be, there are no corresponding formal tests, the plots become sparse in regions with few events, it is not necessarily clear how to interpret deviations from parallelity and there are no general rules on how the stratification of the data should be done. See for instance Andersen et al. (1993, ch. 7) for details on this and other graphical approaches.

Statistical tests for checking the proportionality assumption have been considered by a number of authors, see for instance Cox (1972), Wei (1984), Nagelkerke, Oosting and Hart (1984), Gill and Schumacher (1987), O'Quigley and Pessione (1989), Therneau, Grambsch and Fleming (1990), Pettitt and Bin Daud (1990), Lin (1991), Lin, Wei and Ying (1993), Murphy (1993), Grambsch and Therneau (1994), Martinussen, Scheike and Skovegaard (2002) and Scheike and Martinussen (2004). Many of the approaches in the papers listed above are based on Schoenfeld residuals (Schoenfeld, 1982), see Section 2.1, or cumulative sums of Schoenfeld residuals. Of particular interest among these is the approach of Grambsch and Therneau (1994), see also Therneau and Grambsch (2000, ch. 6), which contains many of the other tests as special cases, and which also includes a plotting method.

The approach of Grambsch and Therneau (1994) is based on checking proportional intensity against the alternative of time-dependent coefficients,

$$\lambda(t) = \lambda_0(t)\exp(\boldsymbol{\beta}(t)\mathbf{X}). \tag{2}$$

Grambsch and Therneau (1994) show that a plot of the Schoenfeld residuals, properly scaled, for covariate l gives a first order approximation to $\beta_l(t)$, $l = 1, \cdots, p$.

See also Winnett and Sasieni (2001). Further, writing $\beta_l(t)$ as a regression on some function of time $g_l(t)$

$$\beta_l(t) = \beta_l + \theta_l g_l(t), \quad l = 1, \ldots, p, \tag{3}$$

Grambsch and Therneau (1994) derive a corresponding test of proportionality, where $\theta_l = 0$ corresponds to proportionality. This test can roughly be thought of as a test of zero slope in a regression line fit to a plot of the scaled Schoenfeld residuals against $g_l(t)$. Both a simultaneous test of proportionality and separate tests for each covariate are provided. Different choices of $g_l(t)$ correspond to different tests of deviation from proportionality, and a number of earlier proposed tests of proportionality are special cases of this test corresponding to particular choices of the $g_l(t)$ function. A limitation with the Grambsch and Therneau (1994) approach is that only one specific alternative to proportional intensity, namely time-dependent coefficients, is checked. Other kinds of deviations can possibly be wrongly interpreted or not detected at all. Another limitation is the need to choose specific $g_l(t)$ functions. This may lead to low power against deviations of a nature not described by this function, for instance a nonmonotonic deviation when a monotonic $g_l(t)$ function has been chosen. See Winnett and Sasieni (2001) and Scheike and Martinussen (2004) for further comments on the Grambsch and Therneau (1994) approach.

More omnibus tests can be constructed based on the score process. This process is equivalent to cumulative sums of the Schoenfeld residuals. In this paper, we will look further at tests based on the score process. Under certain conditions the score process converges to a Brownian bridge, and measures of deviation from Brownian bridge can thus be used as tests for proportional intensity. Wei (1984), Therneau, Grambsch and Fleming (1990) and Lin, Wei and Ying (1993) use the maximum deviation of the score process from the zero line as a test of deviation from proportionality. We have looked at other measures of deviation from Brownian bridge which use more of the available information and which thus presumably should lead to more powerful tests.

The tests based on the score process are presented in Section 2. In Section 3 these and some of the tests in the Grambsch and Therneau (1994) class of tests are compared in a simulation study. A real data example is given in Section 4 and some concluding comments are given in Section 5.

2. Tests for Proportionality

We observe the occurrence of a, possible recurrent, event of interest for n different independent subjects with corresponding covariate vectors $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$, $i = 1, \ldots, n$, which are constant in time. We thus observe n independent counting processes $N_i(t)$, $i = 1, \ldots, n$, where each process is observed until time τ_i, $i = 1, \ldots, n$, and the censoring is assumed to be noninformative. Further, we introduce the predictable 0–1 process $Y_i(t)$ which indicates whether subject i is at

risk at time t ($Y_i(t) = 1$) or not ($Y_i(t) = 0$). Finally, we let t_{ij} denote the time when event j, $j = 1, \ldots, N_i(\tau_i)$ occurs in process i, $i = 1, \ldots, n$.

2.1. The Score Process

For the proportional intensity model (1) the partial likelihood can be written (Johansen, 1983)

$$\mathrm{PL} = \prod_{i=1}^{n} \prod_{j=1}^{N_i(\tau_i)} \frac{\exp(\beta^T \mathbf{X}_i)}{\sum_{k=1}^{n} Y_k(t_{ij}) \exp(\boldsymbol{\beta}^T \mathbf{X}_k)}.$$

The derivative of the logarithm of the partial likelihood with respect to β_l, $l = 1, \ldots, p$, becomes

$$\begin{aligned}\frac{\partial \log \mathrm{PL}}{\partial \beta_l} &= \sum_{i=1}^{n} \sum_{j=1}^{N_i(\tau_i)} \left\{ X_{il} - \frac{\sum_{k=1}^{n} Y_k(t_{ij}) X_{kl} \exp(\boldsymbol{\beta}^T \mathbf{X}_k)}{\sum_{k=1}^{n} Y_k(t_{ij}) \exp(\boldsymbol{\beta}^T \mathbf{X}_k)} \right\} \\ &= \sum_{i=1}^{n} \int_0^{\infty} \{ X_{il} - \bar{x}_l(\boldsymbol{\beta}, u) \}\, dN_i(u),\end{aligned}$$

where

$$\bar{x}_l(\boldsymbol{\beta}, u) = \frac{\sum_{k=1}^{n} Y_k(u) X_{kl} \exp(\boldsymbol{\beta}^T \mathbf{X}_k)}{\sum_{k=1}^{n} Y_k(u) \exp(\boldsymbol{\beta}^T \mathbf{X}_k)} \tag{4}$$

is a weighted mean of the covariates over the risk set at time u. We define the score process as

$$U_l(\hat{\boldsymbol{\beta}}, t) = \sum_{i=1}^{n} \int_0^{t} \left\{ X_{il} - \bar{x}_l(\hat{\boldsymbol{\beta}}, u) \right\} dN_i(u),$$

where $\hat{\boldsymbol{\beta}}$ is the maximum partial likelihood estimate found by solving $U_l(\hat{\boldsymbol{\beta}}, \infty) = 0$ simultaneously for all $l = 1, \ldots, p$. The Schoenfeld residuals (Schoenfeld, 1982) mentioned in Section 1 are simply the increments of this score process. Under an independent covariate condition to be explained in detail later and certain regularity conditions specified by Andersen and Gill (1982), it was shown by Therneau, Grambsch and Fleming (1990) that for each covariate $l = 1, \ldots, p$

$$\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_l)}\, U_l(\hat{\boldsymbol{\beta}}, \cdot) \xrightarrow{L} W^0\left(\frac{\sigma_{ll}(\cdot)}{\sigma_{ll}(\infty)} \right), \tag{5}$$

where $W^0(\cdot)$ is a Brownian bridge and the $\sigma_{ll}(t)$ function will be defined later. Thus tests for proportionality can be constructed by applying measures of deviation from the Brownian bridge to the score process scaled by $\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_l)}$. Plotting the scaled score process can also give useful information. Some typical examples of score process plots are given in Figure 1.

The upper plot illustrates a score process behaving similar to a Brownian bridge, while the middle and lower plots illustrate typical deviations from Brownian bridge

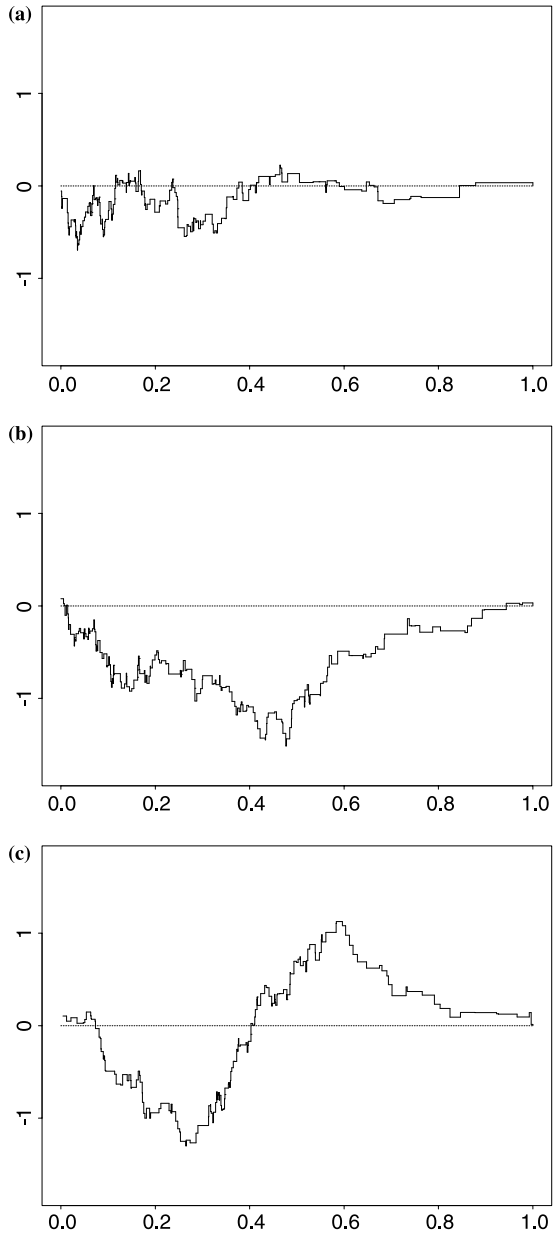

Figure 1. Examples of scaled score processes $\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_l)}U_l(\hat{\boldsymbol{\beta}}, \cdot)$ for data generated from $n = 200$ processes without recurrent events using the model $\lambda(t) = 2\exp(\beta(t)X)$ with respectively (a) $\beta(t) = 1$ (b) $\beta(t) = 2t$ and (c) $\beta(t) = -1 + 2 * I(0.3 \leq t \leq 0.6)$, constant censoring at $t = 1$ and X generated from the Uniform[0,1] distribution.

behavior in a case with monotonic and a case with nonmonotonic deviations from proportionality, respectively.

Therneau, Grambsch and Fleming (1990) and Lin, Wei and Ying (1993) propose to use the statistic

$$\mathrm{KS} = \sup_t \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_l)}|U_l(\hat{\boldsymbol{\beta}}, t)| \tag{6}$$

to test deviations from proportionality. This is a Kolmogorov–Smirnov type statistic with the well-known distribution of $\sup_{0 \le u \le 1} |W^0(u)|$ as asymptotic distribution (see for instance Billingsley, 1968). On a 5% level, the null hypothesis is rejected when $\mathrm{KS} \ge 1.36$. This is a handy test which is easy to calculate but which only uses a part of the available information. We want to look at alternative tests which use more of the available information and which thus might be more powerful. Before we present these tests we need to define the quantity $\sigma_{ll}(\cdot)$ and look at the condition needed to get (5).

Let $\mathbf{U}(\boldsymbol{\beta}, t)$ and $\bar{\mathbf{x}}(\boldsymbol{\beta}, t)$ be the column vectors with $U_l(\boldsymbol{\beta}, t)$ and $\bar{x}_l(\boldsymbol{\beta}, t)$, $l = 1, \ldots, p$ as components, respectively. The negative first derivative of $\mathbf{U}(\boldsymbol{\beta}, t)$ is the $p \times p$ matrix

$$\mathbf{I}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \mathbf{V}(\boldsymbol{\beta}, u)\, dN_i(u), \tag{7}$$

where $\mathbf{V}(\boldsymbol{\beta}, u)$ is the weighted covariance of $\mathbf{X}$ at time u,

$$\mathbf{V}(\boldsymbol{\beta}, u) = \frac{\sum_{i=1}^n Y_i(u) \exp(\boldsymbol{\beta}^T \mathbf{X}_i)[\mathbf{X}_i - \bar{\mathbf{x}}(\boldsymbol{\beta}, u)][\mathbf{X}_i - \bar{\mathbf{x}}(\boldsymbol{\beta}, u)]^T}{\sum_{i=1}^n Y_i(u) \exp(\boldsymbol{\beta}^T \mathbf{X}_i)}.$$

Note that $\mathbf{U}(\boldsymbol{\beta}, \infty)$ is the score vector and $\mathbf{I}(\boldsymbol{\beta}, \infty)$ the information matrix. It can be shown that

$$\frac{1}{n}\mathbf{I}(\hat{\boldsymbol{\beta}}, t) \xrightarrow{p} \mathbf{V}(t) \tag{8}$$

uniformly where $\mathbf{V}(t)$ is the asymptotic covariance matrix of $n^{-1/2}\mathbf{U}(\boldsymbol{\beta}_0, t)$ and $\boldsymbol{\beta}_0$ is the true parameter vector. See Andersen et al. (1993, ch. 7) for details. Now $\sigma_{ll}(t) = \mathbf{V}_{ll}(t)$, and the condition required for (5) to hold for each separate covariate is that $\mathbf{V}_{ll'}(t) = 0$ for all $l \ne l'$ and all t (Therneau, Grambsch and Fleming 1990), or in other words that $\mathbf{V}(t) = \mathrm{diag}(t)$. Thus the condition is that the off-diagonal elements of the asymptotic covariance matrix of $n^{-1/2}\mathbf{U}(\boldsymbol{\beta}_0, t)$ are zero, and we see from the above relations that this essentially requires the covariates to be independent. However, unless all processes are observed over the same time interval, selection effects may give nonzero off-diagonal elements in $\mathbf{V}(t)$ for $t > 0$ even for independent covariates, but these will in practice typically be small. We shall first assume that $\mathbf{V}(t) = \mathrm{diag}(t)$ holds and come back to what can be done if it does not hold later on. Note that this condition, is of course, always fulfilled in the special case with only a single covariate.

From (8) we see that a uniformly consistent estimator of $\sigma_{ll}(t)$ is given by $\hat{\sigma}_{ll}(t) = \left\{\frac{1}{n}\mathbf{I}(\hat{\boldsymbol{\beta}}, t)\right\}_{ll}$. Defining $q_l(t) = \sigma_{ll}(t)/\sigma_{ll}(\infty)$ we get from (5) that $\sqrt{\widehat{Var}(\hat{\beta}_l)}\, U_l(\hat{\boldsymbol{\beta}}, \cdot) \xrightarrow{L} W^0(q_l(\cdot))$.

2.2. *The Tests*

We here first present the tests under the assumption that $\mathbf{V}(t) = \mathrm{diag}(t)$, generalizations are presented in Section 2.3.

The first measure of deviation from proportionality we consider is simply the integral of the scaled score process. By the uniform consistency of $\hat{q}_l(t)$ we have that

$$G = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_l)} \int_0^\infty U_l(\hat{\boldsymbol{\beta}}, t)\, d\hat{q}_l(t) \xrightarrow{L} \int_0^\infty W^0(q_l(t))\, dq_l(t) = \int_0^1 W^0(u)\, du,$$

where $\int_0^1 W^0(u)\, du$ is normally distributed with expectation 0 and variance 1/12. For calculating the test statistic more explicitly we introduce $0 < t_1^* \le t_2^* \le \cdots \le t_N^* \le \tau$ as the event times from all processes, t_{ij} , $i = 1, \ldots, n$, $j = 1, \ldots, N_i(\tau_i)$, sorted in increasing order, and where $\tau = \max_i \tau_i$ and $N = \sum_{i=1}^n N_i(\tau_i)$. Letting $u = \hat{q}_l(t)$ we get that

$$G = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_l)} \int_0^1 U_l(\hat{\boldsymbol{\beta}}, \hat{q}_l^{-1}(u))\, du$$

and since $U_l(\hat{\boldsymbol{\beta}}, \hat{q}_l^{-1}(u))$ is a constant for $\hat{q}_l^{-1}(u) \in [t_{k-1}^*, t_k^*]$ or $u \in [\hat{q}_l(t_{k-1}^*), \hat{q}_l(t_k^*)]$ we get that

$$\begin{aligned} G &= \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_l)} \sum_{k=2}^N U_l(\hat{\boldsymbol{\beta}}, t_{k-1}^*)\left[\hat{q}_l(t_k^*) - \hat{q}_l(t_{k-1}^*)\right] \\ &= \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_l)} \sum_{k=2}^N U_l(\hat{\boldsymbol{\beta}}, t_{k-1}^*)\left[\hat{\sigma}_{ll}(t_k^*) - \hat{\sigma}_{ll}(t_{k-1}^*)\right]/\hat{\sigma}_{ll}(\infty), \end{aligned} \tag{9}$$

where the summation can start at $k = 2$ since $U_l(\hat{\boldsymbol{\beta}}, t) = 0$ for $t < t_1^*$, and where $\hat{\sigma}_{ll}(t_N^*) = \hat{\sigma}_{ll}(\infty)$. This test should have good power properties against monotonic deviations from proportionality, for instance a time-dependent coefficients model (2) with monotonic $\beta_l(t)$ which typically give a $\cup$ or $\cap$ shaped one-sided deviation from the Brownian bridge as seen on the plot in the middle in Figure 1. However, other kinds of nonproportionalities may lead to score processes with deviations to both sides of the zero line, as seen in the lower plot in Figure 1. In such cases the integrated process is not a good choice for test statistic since the deviations on the two sides of the zero line will tend to cancel out in the integration. A better choice, with in principle power against any kind of deviation, is the Cramér–von Mises type statistic

$$\mathrm{CV} = \int_0^\infty \widehat{\mathrm{Var}}(\hat{\beta}_l) U_l(\hat{\boldsymbol{\beta}}, t)^2\, d\hat{q}_l(t) \xrightarrow{L} \int_0^1 W^0(u)^2\, du,$$

where the right-hand side has the Cramér–von Mises distribution (Cramér, 1928; von Mises, 1931). Doing the same kind of calculations as above, this statistic can be written

$$\mathrm{CV} = \widehat{\mathrm{Var}}(\hat{\beta}_l) \sum_{k=2}^{N} U_l(\hat{\boldsymbol{\beta}}, t_{k-1}^*)^2 \left[\hat{\sigma}_{ll}(t_k^*) - \hat{\sigma}_{ll}(t_{k-1}^*)\right] / \hat{\sigma}_{ll}(\infty). \tag{10}$$

On a 5% level the null hypothesis is rejected when $\mathrm{CV} \geq 0.461$. A variant of this statistic which could possibly have even better properties is the Anderson–Darling type statistic

$$\mathrm{AD} = \widehat{\mathrm{Var}}(\hat{\beta}_l) \int_0^\infty \frac{U_l(\hat{\boldsymbol{\beta}}, t)^2}{\hat{q}_l(t)(1-\hat{q}_l(t))} d(\hat{q}_l(t)) \xrightarrow{L} \int_0^1 \frac{W^0(u)^2}{u(1-u)} du,$$

where the right-hand side has the Anderson–Darling distribution (Anderson and Darling, 1952). The factor $1/(\hat{q}_l(t)(1-\hat{q}_l(t)))$ puts more weight on the tied down ends of the score process than the Cramér–von Mises statistic. This can improve the power properties of the test, but also implies that much weight is placed in a region with possibly few observations. The Anderson–Darling statistic can be simplified to

$$\mathrm{AD} = \widehat{\mathrm{Var}}(\hat{\beta}_l) \sum_{k=2}^{N-1-I(t_N^*=\tau)} U_l(\hat{\boldsymbol{\beta}}, t_{k-1}^*)^2 \left[\log\left(\frac{\hat{\sigma}_{ll}(t_k^*)}{\hat{\sigma}_{ll}(t_{k-1}^*)}\right) + \log\left(\frac{\hat{\sigma}_{ll}(\infty) - \hat{\sigma}_{ll}(t_{k-1}^*)}{\hat{\sigma}_{ll}(\infty) - \hat{\sigma}_{ll}(t_k^*)}\right)\right]. \tag{11}$$

To avoid dividing by zero in the test statistic the sum can only go to $N-1$ when $t_N^* < \tau$ since $\hat{\sigma}_{ll}(t_N^*) = \hat{\sigma}_{ll}(\infty)$, and only to $N-2$ if $t_N^* = \tau$ since then $\hat{\sigma}_{ll}(t_{N-1}^*) = \hat{\sigma}_{ll}(\infty)$ caused by all entries in $\mathbf{V}(\hat{\boldsymbol{\beta}}, t)$ being zero at $t = t_N^* = \tau$ (unless a censoring also occurs exactly at $t = \tau$). On a 5% level the null hypothesis is rejected when $\mathrm{AD} \geq 2.492$.

A special case in which things simplify greatly is if all subjects have been observed over exactly the same interval, implying that all $Y_i(t)$, $i = 1, \ldots, n$, are identical. In that case $\mathbf{V}(\hat{\boldsymbol{\beta}}, t)$ will be constant, implying from (7) that the estimator $\hat{q}_l(t) = \hat{\sigma}_{ll}(t)/\hat{\sigma}_{ll}(\infty)$ can simply be written $\hat{q}_l(t) = \sum_{i=1}^n N_i(t)/N$.

2.3. *Generalization*

The test statistics presented above all require independent covariates in the sense that $\mathbf{V}(t) = \mathrm{diag}(t)$ for all t. If this assumption does not hold, the process $\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_l)} U_l(\hat{\boldsymbol{\beta}}, \cdot)$ will still be a tied down mean zero Gaussian process starting and ending at zero, but not exactly a Brownian bridge. See for example Andersen et al. (1993, p. 554). However, the process will be something similar to a Brownian bridge, and one possibility is simply to use the same tests and study what properties they will have. Simulations reported in Section 3 indicate that the violation of the independence assumption must be quite severe before it makes any

notable impact on the tests. Thus the assumption need not be strictly fulfilled for the tests to work.

If, however, there are strong dependencies in the covariates, an alternative approach is to adopt the simulation method presented by Lin, Wei and Ying (1993) to simulate the null distribution of the score process and thus of the test statistics. Lin, Wei and Ying (1993) study how to simulate various processes which are cumulative sums of martingale-based residuals and how to use this to test various aspects of the proportional hazards model. Applying this to our score process, the process to be simulated from can be written

$$\hat{\mathbf{U}}(\hat{\boldsymbol{\beta}}, t) = \sum_{i=1}^{n} \sum_{j=1}^{N_i(\tau_i)} I(t_{ij} \leq t)\left\{\mathbf{X}_i - \bar{\mathbf{x}}(\hat{\boldsymbol{\beta}}, t_{ij})\right\} G_{ij}$$
$$- \mathbf{I}(\hat{\boldsymbol{\beta}}, t)\mathbf{I}(\hat{\boldsymbol{\beta}}, \infty)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N_i(\tau_i)} \left\{\mathbf{X}_i - \bar{\mathbf{x}}(\hat{\boldsymbol{\beta}}, t_{ij})\right\} G_{ij},$$

where G_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, N_i(\tau_i)$ is a random sample from the standard normal distribution. For covariate l we consider the lth component of this process vector, $\hat{U}_l(\hat{\boldsymbol{\beta}}, t)$. Lin, Wei and Ying (1993) show that the conditional distribution of $n^{-1/2}\hat{\mathbf{U}}(\hat{\boldsymbol{\beta}}, t)$ given the observed data in the limit is the same as the unconditional distribution of $n^{-1/2}\mathbf{U}(\hat{\boldsymbol{\beta}}, t)$. Thus the asymptotic null distribution of a test statistic, say AD, for testing the proportionality of covariate l can be approximated by generating a large number of processes $\hat{U}_l(\hat{\boldsymbol{\beta}}, t)$ and calculating the AD statistic for all processes. If the value of the AD statistic calculated from the original data lies in the upper 5% of the statistics from the simulated processes, the null hypothesis is rejected on a 5% level.

2.4. Comments

Formally tests based on the score process are general tests for goodness of fit of the model. However, in practice these tests will primarily detect departures from proportionality, other kinds of departures are generally not picked up by tests based on the score process. More omnibus tests for goodness of fit are among others studied by Lin, Wei and Ying (1993), Grønnesby and Borgan (1996), Marzec and Marzec (1997) and Parzen and Lipsitz (1999).

When proportionality is tested for one covariate, formally this is done under the assumption that proportionality holds for the other covariates. If some covariates have a nonproportional effect this might thus potentially affect the testing of other covariates. However, intuitively, this should mainly be a problem in cases with strongly correlated covariates, as the effect of other covariates only influences the score process of a covariate through the weights in the calculation of the weighted mean (4). Thus as long as the variation of other covariates is fairly nonsystematic compared to the covariate under examination a possible nonproportionality in some of the other covariates should not play any important role. This is confirmed by

simulations in Section 3. A sequential test procedure where each covariate is tested without assuming proportionality for the other covariates is presented by Scheike and Martinussen (2004).

3. Simulation Study

In this section, the level and power properties of the tests presented in Section 2 are studied. For comparisons, two of the tests in the Grambsch and Therneau (1994) class of tests are included as well. Rejection probabilities are estimated by Monte Carlo simulations, in all examples studied 2000 repetitions are used. This implies that the standard deviation of the reported estimated rejection probabilities is at most $\sqrt{0.5 \cdot 0.5/2000} = 0.011$ and for level simulations typically around $\sqrt{0.05 \cdot 0.95/2000} = 0.005$. For the tests based on simulated null distributions 1000 realizations are generated each time as recommended by Lin, Wei and Ying (1993). The simulations are performed in S-plus. In the reported simulations the most common case of counting processes terminated at the first event or censoring are considered. Other simulations not reported have indicated that the same relationships among the tests as reported here seem to hold for recurrent event data as well.

The two tests in the Grambsch and Therneau (1994) class of tests included are the tests corresponding to setting $g_l(t) = \log(t)$ and $g_l(t) = N - \sum_{i=1}^{n} N_i(t)$ in (3), respectively. The first test is a version of the test originally proposed by Cox (1972). The second test, with $g_l(t)$ being the rank of the event times, is equivalent to a test proposed by Breslow, Edler and Berger (1984). This test does, in practice, typically give the same result as a version of a test proposed by Lin (1991) obtained by setting $g_l(t)$ equal to the Kaplan–Meier estimator. See Grambsch and Therneau (1994) and Therneau and Grambsch (2000, ch. 6) for details. Below these two tests are abbreviated GT-log and GT-rank, respectively.

We also use the abbreviations KS for the Kolmogorov–Smirnov type statistic (6), G for the asymptotically Gaussian distributed statistic (9), CV for the Cramér–von Mises type statistic (10) and AD for the Anderson–Darling type statistic (11). The subscript s is used to denote when simulated null distributions are used for these tests. The results for the CV test are not reported explicitly below as it turned out that this test behaves very similar to the AD test, the only difference is that the CV test generally is slightly less powerful than the AD test.

3.1. Case 1

We start the study of level and power properties of the various tests by generating data from the model $\lambda(t) = 2\exp(at^b X)$ with no censoring and Uniform[0,1] distributed covariates. Rejection probabilities for the tests for different values of a and b and different sample sizes are reported in Table 1.

Table 1. Estimated rejection probabilities for different values of a and b and different sample sizes n in the model $\lambda(t) = 2\exp(at^bX)$ with no censoring and X being Uniform[0,1] distributed.

			Test							
a	b	n	AD	AD_S	G	G_S	KS	KS_S	GT-log	GT-rank
1	0	50	0.052	0.055	0.051	0.056	0.031	0.055	0.039	0.034
1	0	100	0.055	0.051	0.052	0.054	0.034	0.056	0.042	0.037
1	0	200	0.050	0.050	0.052	0.051	0.034	0.053	0.039	0.040
1	0	500	0.051	0.048	0.049	0.044	0.044	0.051	0.046	0.044
2	1	100	0.273	0.221	0.277	0.234	0.195	0.201	0.185	0.228
4	1	100	0.484	0.449	0.478	0.462	0.335	0.395	0.338	0.403
6	1	100	0.654	0.571	0.645	0.591	0.478	0.512	0.479	0.552
8	1	100	0.739	0.684	0.731	0.695	0.564	0.587	0.553	0.626

Table 2. Estimated rejection probabilities for various sample sizes n for the model $\lambda(t) = 2\exp(\beta(t)X)$, where $\beta(t) = -\ln(4) + \ln(4) \cdot I(0.3 \leq t \leq 0.6))$, X is Uniform[0,2] distributed and the censoring is constant at time 1.2, giving around 33% censoring.

	Test							
n	AD	AD_S	G	G_S	KS	KS_S	GT-log	GT-rank
50	0.138	0.111	0.072	0.062	0.103	0.116	0.082	0.068
100	0.311	0.316	0.068	0.071	0.250	0.307	0.115	0.064
200	0.698	0.699	0.093	0.092	0.593	0.639	0.197	0.090
500	0.998	0.998	0.143	0.136	0.988	0.992	0.478	0.139

The cases with $b = 0$ in Table 1 illustrate level properties of the tests. We see that the AD and G tests based on asymptotic null distributions and all score process tests based on simulated null distributions have very good level properties even for fairly small sample sizes. The GT-log and GT-rank tests tend to be slightly conservative for small sample sizes, while the KS test based on asymptotic null distribution is a bit conservative even for larger sample sizes. Using the simulated null distribution improves the level properties of the KS test. Simulations from other models have shown the same tendencies regarding level properties as reported in Table 1.

The cases with $b = 1$ in Table 1 illustrate power properties. We see that the AD and G tests perform very similarly and are better than the other tests in these cases. Among the other tests the GT-rank test is slightly better than the KS and GT-log tests. Also notice that the AD and G tests are a bit less powerful when simulated null distributions are used instead of asymptotic null distributions, while the KS test is a bit more powerful when the simulated null distribution is used due to better level properties. We also note that the deviation from proportionality has to be quite strong before the probability of detecting it becomes large.

Similar relationships among the tests as reported here have been seen in other simulations of models with monotonic deviations from proportionality. The AD and G tests are always clearly better than the KS and GT-log tests, while the GT-rank

test in some cases, in particular for large sample sizes, can have power properties comparable to the AD and G tests. It also seems to be a general tendency that the AD and G tests are equally or a bit less powerful when the simulated null distributions are used, while the KS test typically is a bit more powerful. The examples below also illustrate this.

3.2. Case 2

An example of a nonmonotonic deviation from proportionality is given by the time-dependent coefficients model $\lambda(t) = 2\exp(\beta(t)X)$ where $\beta(t) = -\ln(4) + \ln(4)I(0.3 \leq t \leq 0.6)$. Samples of different sizes with constant censoring at $\tau = 1.2$ and X being Uniform[0,2] distributed are generated. This gives a censoring percentage of around 33%. The results are reported in Table 2.

We see in Table 2 that the AD test has the largest power among the tests for detecting this nonmonotonic deviation. The KS test also has good power properties, while the G test, as expected, does not have power to detect this kind of deviation from proportionality. The GT tests are also far less powerful than the AD and KS tests in this case, in particular the GT-rank test. Of course, in this case other GT tests with a nonmonotoic $g_l(t)$ function would presumably have better power properties for detecting this nonmonotonic kind of deviation from proportionality.

In other simulations of models with nonmonotonic deviations from proportionality, similar results as reported in Table 2 have been seen. The AD test is the most powerful test while the G and GT-rank tests in all such cases have been the least powerful tests. The KS and GT-log tests, however, vary a lot in performance, in some cases they are close to the AD test, as the KS test in Table 2, while they in other cases are far less powerful.

3.3. Case 3

We now proceed to consider situations with dependent covariates. We start by studying level properties by generating data from the model $\lambda(t) = \exp(X_1 + X_2 - 8)$ where the covariates X_1 and X_2 are jointly drawn from a binormal distribution with both covariates having expectation 4 and variance 1, and with the correlation between them being denoted ρ. Censoring variables are drawn from the Uniform[0,5] distribution, giving approximately 30% censoring. The simulated rejection probabilities for the tests, used for testing the proportionality assumption for one of the covariates, X_1, for different values of the correlation ρ and samples of size $n = 100$ and $n = 500$, are reported in Table 3.

From this table we see that the correlation between the covariates has to be quite strong before it seriously affects the level of the score process tests based on asymptotic null distributions, first at a correlation of 0.7 does the correlation start to affect the level properties of the tests a bit. The score process tests based on simulated null distributions do as expected have good level properties in all cases. We also note that the level properties are the same for the different sample sizes.

Table 3. Estimated rejection probabilities for testing the proportionality assumption for X_1 for different values of ρ, the correlation between the normally distributed covariates X_1 and X_2 in the model $\lambda(t) = \exp(X_1 + X_2 - 8)$, with $n = 100$ and $n = 500$ observations. The censoring variables are drawn from the Uniform[0,5] distribution, giving around 30% censoring. The covariates X_1 and X_2 both have expectation 4 and variance 1.

		Test							
ρ	n	AD	AD_S	G	G_S	KS	KS_S	GT-log	GT-rank
0.5	100	0.046	0.040	0.046	0.044	0.033	0.049	0.050	0.044
0.6	100	0.054	0.047	0.052	0.053	0.039	0.057	0.055	0.045
0.7	100	0.068	0.052	0.062	0.059	0.054	0.059	0.051	0.045
0.8	100	0.131	0.045	0.113	0.053	0.115	0.052	0.062	0.060
0.9	100	0.286	0.040	0.186	0.049	0.285	0.052	0.048	0.042
0.5	500	0.047	0.055	0.050	0.055	0.044	0.060	0.064	0.054
0.6	500	0.050	0.048	0.048	0.052	0.050	0.057	0.057	0.047
0.7	500	0.070	0.053	0.058	0.052	0.062	0.048	0.050	0.043
0.8	500	0.106	0.041	0.082	0.047	0.106	0.047	0.051	0.045
0.9	500	0.259	0.046	0.172	0.049	0.283	0.048	0.050	0.055

Table 4. Estimated rejection probabilities for both of the covariates for different values of ρ, the correlation between the normally distributed covariates X_1 and X_2 in the model $\lambda(t) = \exp(0.5tX_1 + X_2 - 8)$, with $n = 100$ and $n = 500$ observations. The censoring variables are drawn from the Uniform[0,5] distribution, giving around 45% censoring. The covariates X_1 and X_2 both have expectation 4 and variance 1.

			Test							
ρ	n	Cov.	AD	AD_S	G	G_S	KS	KS_S	GT-log	GT-rank
0.3	100	X_1	0.542	0.488	0.551	0.518	0.373	0.427	0.472	0.457
0.5	100	X_1	0.520	0.453	0.531	0.482	0.352	0.399	0.466	0.445
0.7	100	X_1	0.540	0.393	0.538	0.424	0.410	0.337	0.322	0.299
0.9	100	X_1	0.771	0.325	0.690	0.361	0.722	0.279	0.152	0.129
0.3	100	X_2	0.056	0.041	0.052	0.050	0.033	0.055	0.077	0.043
0.5	100	X_2	0.052	0.048	0.051	0.059	0.033	0.067	0.070	0.048
0.7	100	X_2	0.200	0.104	0.188	0.129	0.156	0.112	0.062	0.040
0.9	100	X_2	0.621	0.232	0.533	0.254	0.589	0.192	0.059	0.039
0.3	500	X_1	0.998	0.998	0.998	0.999	0.986	0.994	0.993	0.996
0.5	500	X_1	0.997	0.999	0.996	0.999	0.983	0.986	0.988	0.994
0.7	500	X_1	0.994	0.988	0.992	0.989	0.983	0.971	0.957	0.970
0.9	500	X_1	0.998	0.976	0.996	0.977	0.997	0.945	0.669	0.688
0.3	500	X_2	0.058	0.064	0.048	0.054	0.047	0.051	0.077	0.049
0.5	500	X_2	0.112	0.093	0.089	0.072	0.090	0.087	0.079	0.047
0.7	500	X_2	0.505	0.427	0.483	0.422	0.452	0.392	0.070	0.043
0.9	500	X_2	0.971	0.858	0.962	0.874	0.969	0.796	0.054	0.037

3.4. *Case 4*

Finally we study both level and power properties in a case with two correlated covariates where one of the covariates has a non-proportional effect. We do this by generating data from the model $\lambda(t) = \exp(0.5tX_1 + X_2 - 8)$ where the covariates X_1 and X_2 as in the previous case are binormal with expectation 4, variance 1 and correlation ρ. Censoring variables are drawn from the Uniform[0,5] distribution giving approximately 45% censoring. Both of the covariates are tested for proportionality, and the simulated rejection probabilities for different values of ρ and samples of size 100 and 500 are reported in Table 4.

Considering X_1, we see that the score process based tests have good power properties compared to the GT tests, in particular, when the correlation is strong. However, considering X_2, we also see that the score process tests in this case achieve far too high levels for strong correlations. As commented in Section 2.4 this is due to the fact that proportionality for each covariate is tested under the assumption that proportionality holds for the other covariates. In Section 2.4, it was claimed that this should mainly be a problem when we have quite strong correlations among the covariates. This is largely confirmed in Table 4, but for the $n = 500$ case level deviations are seen also for more moderate correlations. This is of course due to the fact that for very large sample sizes small influences on the score process is more easily detected. In the present case the GT tests show relatively reasonable level behavior, but in other cases similar level problems have also been reported for GT tests (Scheike and Martinussen, 2004).

3.5. *Comments*

The simulation study has shown that as long as there are only weak correlations in the covariates, the AD, CV, G and KS tests based on asymptotic null distributions can safely be applied. If, however, there are quite strong correlations in the covariates the simulated null distributions should be used for these tests. If we have correlations and nonproportionality has been detected for one of the covariates, care should be taken in the interpretation of the results for other covariates strongly correlated to this covariate. The AD test (closely followed by the CV test) seems to have the best overall power properties of the tests studied.

4. Example

For illustrating application of the tests on real data we use the primary biliary cirrhosis (PBC) data from the Mayo Clinic. PBC is a fatal chronic liver disease, and out of the 418 patients followed in the study, 161 died before study closure. A listing of the data can be found in Fleming and Harrington (1991). In a study of the data by Dickson et al. (1989) a proportional hazards model with the five covariates age, edema, log(bilirubin), log(protime) and log(albumin) was fitted. This model is

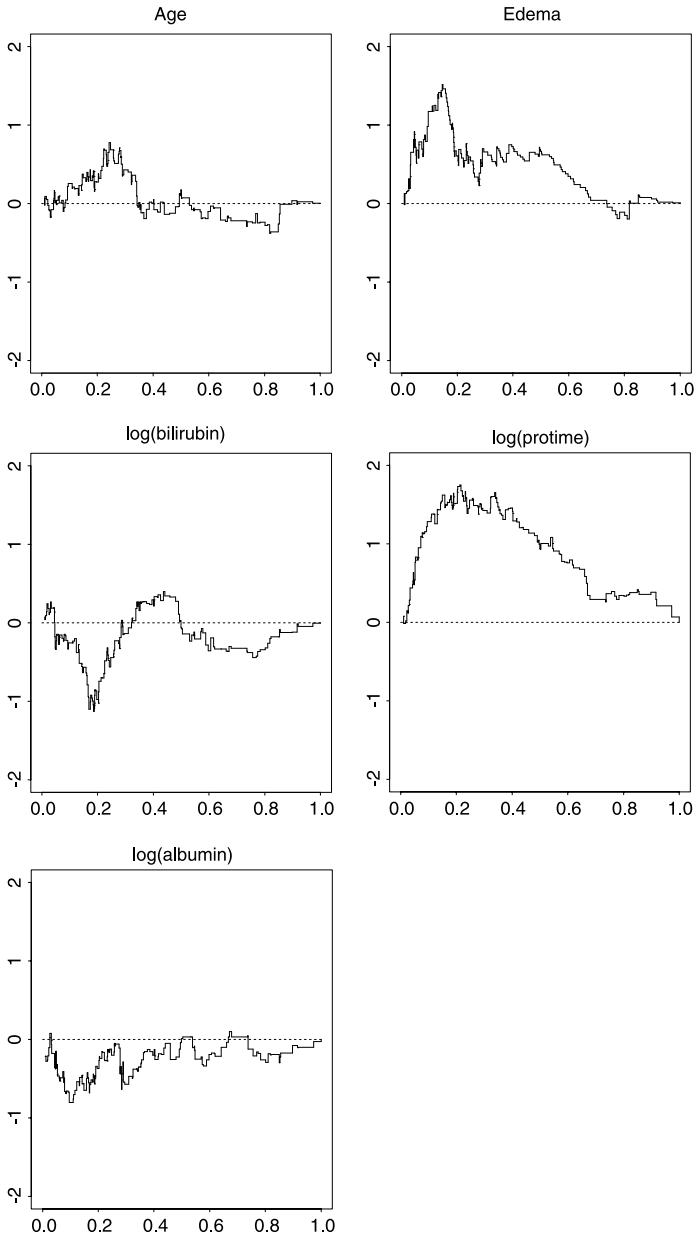

Figure 2. Plots of the scaled score processes for the five covariates in the proportional hazards model fitted for the PBC data.

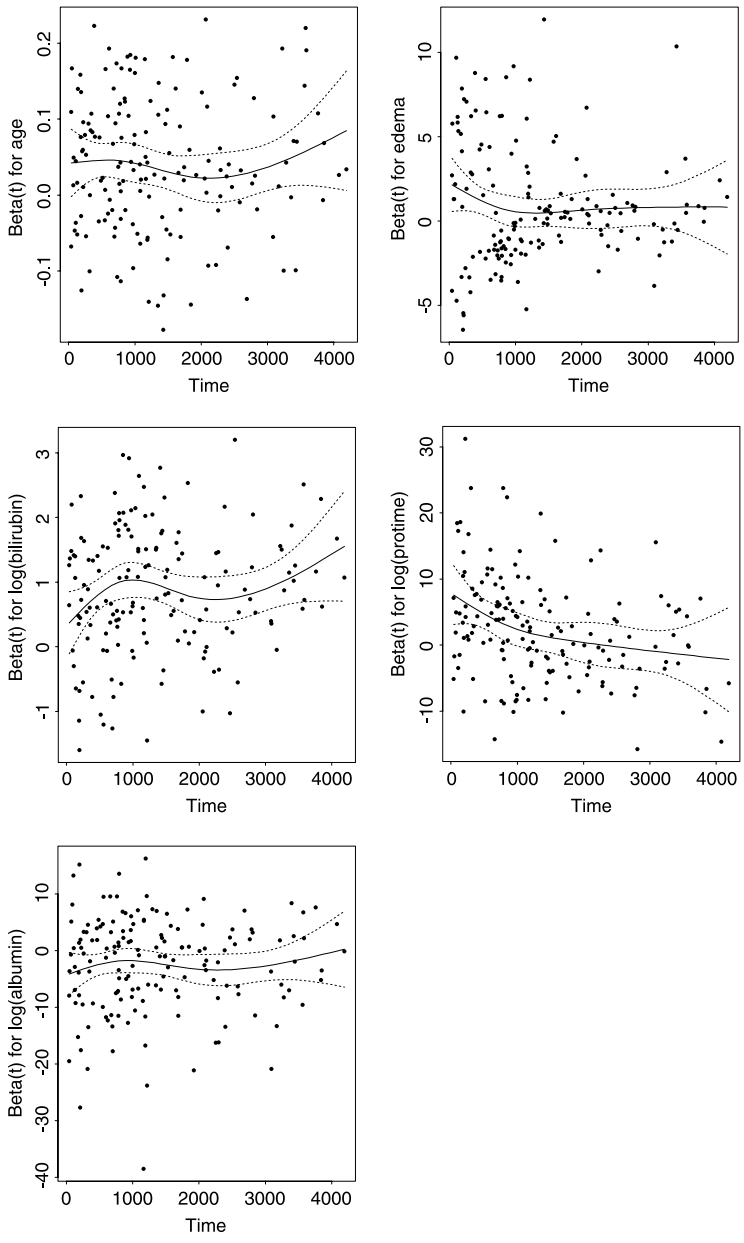

Figure 3. Plots of scaled Schoenfeld residuals (Grambsch and Therneau, 1994) for the five covariates in the proportional hazards model fitted for the PBC data. A spline smoother with 90% confidence interval is added to the plots (Therneau and Grambsch, 2000, ch. 6).

Table 5. The table reports *P*-values for each test applied to the five covariates in the proportional hazards model for the PBC-data. In the tests based on simulated null distributions 10,000 realizations are used.

	Test							
Covariate	AD	AD_S	G	G_S	KS	KS_S	GT-log	GT-rank
Age	0.729	0.652	0.633	0.604	0.584	0.417	0.825	0.740
Edema	0.042	0.055	0.033	0.041	0.020	0.020	0.154	0.268
log(bil)	0.238	0.230	0.353	0.360	0.155	0.098	0.187	0.132
log(pro)	0.001	0.001	0.001	0.001	0.004	0.001	0.002	0.001
log(alb)	0.437	0.538	0.290	0.330	0.539	0.513	0.567	0.715

further studied by Therneau, Grambsch and Fleming (1990), Lin, Wei and Ying (1993) and Therneau and Grambsch (2000).

Plots of the scaled score processes for each of the five covariates are displayed in Figure 2, while plots of the scaled Schoenfeld residuals of Grambsch and Therneau (1994) are displayed in Figure 3.

Both plots clearly indicate that proportionality does not hold for log(protime), while the score process plots also indicate that proportionality might not hold for edema either. In both cases a possible explanation according to the plots could be a covariate effect which fades over time. The five covariates are only slightly correlated, the maximum correlation is 0.34. Thus the score process tests can safely be applied without using simulated null distributions. However, for comparison and verification of this claim, the results obtained with both the asymptotic and the simulated null distributions are reported. The results of applying the tests to the five covariates are reported in Table 5.

We see that all the tests detect the nonproportionality in log(protime), but only the score process based tests detect nonproportionality in edema. GT tests with other time transformations than log and rank have also been used but none of these detected the nonproportionality in edema either. One explanation for this could be that the deviation from proportionality in edema is of a different nature than explained by a time-dependent coefficients model.

We also see in Table 5 that the score process tests based on asymptotic and simulated null distributions give fairly similar results as expected. The minor differences seen could be explained by the observation from Section 3 that the AD and G tests are often slightly less powerful when simulated null distributions are used, while the KS test typically is slightly more powerful due to better level properties.

5. Concluding Comments

Some new tests for the proportional intensity assumption based on the score process are proposed and studied. The simulation study show that in particular the AD test is a useful test for general use by having very good power properties against various

alternatives. The usefulness of applying flexible score process tests is also demonstrated in the PBC-data example where the score process tests are able to detect a nonproportional effect not detected by various commonly used tests in the GT class of tests.

The score process based tests do in principle have power against any type of departure from proportionality. The main limitation is that care should be taken in interpreting the result for a covariate strongly correlated to a covariate with a detected nonproportionality. In cases with no or weak correlations among the covariates the tests based on asymptotic null distributions reported in Section 2.2 could be used directly. In cases with stronger correlations the simulation approach of Lin, Wei and Ying (1993) presented in Section 2.3 should be used to approximate the null distribution of the tests. Due to its wider applicability this could of course be the default procedure, but applying the simulation method is more computationally demanding and the simpler asymptotic null distribution approach also often tends to give a bit more powerful tests in cases where both approaches are applicable.

Plotting the score process, as demonstrated in Figures 1 and 2, can also give useful insight in possible deviations from proportionality.

Acknowledgments

We would like to thank the associate editor and the referees for useful comments and suggestions that improved the paper.

References

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

P. K. Andersen and R. D. Gill, "Cox's regression model for counting processes: A large sample study," *Ann. Stat.* vol. 10 pp. 1100–1120, 1982.

T. W. Anderson and D. A. Darling, "Asymptotic theory of certain goodness of fit criteria based on stochastic processes," *Ann. Math. Stat.* vol. 23 pp. 193–212, 1952.

P. Billingsley, *Convergence of Probability Measures*, Wiley: New York, 1968.

N. E. Breslow, L. Edler, and J. Berger, "A two-sample censored-data rank test for acceleration," *Biometrics* vol. 40 pp. 1049–1062, 1984.

D. R. Cox, "Regression models and life-tables," *J. Roy. Stat. Soc., Ser. B* vol. 34 pp. 187–220, 1972.

H. C. Cramér, "On the composition of elementary errors," *Skandinavisk Aktuarietidsskrift* vol. 11 pp. 13–74, 141–180, 1928.

E. R. Dickson, P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy, "Prognosis in primary biliary cirrhosis: Model for decision making," *Hepatology* vol. 10 pp. 1–7, 1989.

T. R. Fleming and D. P. Harrington, *Counting Processes and Survival Analysis,* Wiley: New York, 1991.

R. Gill and M. Schumacher, "A simple test for the proportional hazards assumption," *Biometrika* vol. 74 pp. 289–300, 1987.

P. M. Grambsch and T. M. Therneau, "Proportional hazards tests and diagnostics based on weighted residuals," *Biometrika* vol. 81 pp. 515–526, 1994.

J. K. Grønnesby and Ø. Borgan, "A method for checking regression models in survival analysis based on the risk score," *Lifetime Data Anal.* vol. 2 pp. 315–328, 1996.

S. Johansen, "An extension of Cox's regression model," *Int. Stat. Rev.* vol. 51 pp. 258–262, 1983.

D. Y. Lin, "Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators," *J. Am. Stat. Assoc.* vol. 86 pp. 725–728, 1991.

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

T. Martinussen, T. Scheike, and I. M. Skovgaard, "Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models," *Scand. J. Stat.* vol. 29 pp. 57–74, 2002.

L. Marzec and P. Marzec, "On fitting Cox's regression model with time-dependent coefficients," *Biometrika* vol. 84 pp. 901–908, 1997.

S. A. Murphy, "Testing for a time dependent coefficient in Cox's regression model," *Scand. J. Stat.* vol. 20 pp. 35–50, 1993.

N. J. D. Nagelkerke, J. Oosting, and A. A. M. Hart, "A simple test for goodness of fit of Cox's proportional hazards model," *Biometrics* vol. 40 pp. 483–486, 1984.

J. O'Quigley and F. Pessione, "Score tests for homogeneity of regression effects in the proportional hazards model," *Biometrics* vol. 45 pp. 135–144, 1989.

M. Parzen and S. R. Lipsitz, "A global goodness-of-fit statistic for Cox regression models," *Biometrics* vol. 55 pp. 580–584, 1999.

A. N. Pettitt and I. Bin Daud, "Investigating time dependence in Cox's proportional hazards model," *Appl. Stat.* vol. 39 pp. 313–329, 1990.

T. Scheike and T. Martinussen, "On estimation and tests of time-varying effects in the proportional hazards model," Scand. J. Stat. Vol. 31 pp. 51–62, 2004.

D. Schoenfeld, "Partial residuals for the proportional hazards regression model," *Biometrika* vol. 69 pp. 239–241, 1982.

T. M. Therneau and P. M. Grambsch, *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag: New York, 2000.

T. M. Therneau, P. M. Grambsch, and T. R. Fleming, "Martingale-based residuals for survival models," *Biometrika* vol. 77 pp. 147–160, 1990.

N. von Mises, *Wahrscheinlichkeitsrechnung*, Deuticke, Leipzig, 1931.

L. J. Wei, "Testing goodness of fit for the proportional hazards model with censored observations," *J. Am. Stat. Assoc.* vol. 79 pp. 649–652, 1984.

A. Winnett and P. Sasieni, "A note on scaled Schoenfeld residuals for the proportional hazards model," *Biometrika* vol. 88 pp. 565–571, 2001.