Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Adaptive Neyman's smooth tests of homogeneity of two samples of survival data

David Kraus[a,b,*]

[a] Institute of Information Theory and Automation, Prague, Czech Republic
[b] Department of Statistics, Charles University in Prague, Czech Republic

ARTICLE INFO

Article history:
Received 19 June 2007
Received in revised form
6 April 2009
Accepted 7 April 2009
Available online 18 April 2009

Keywords:
Censoring
Neyman's smooth test
Schwarz's selection rule
Survival analysis
Two sample test

ABSTRACT

The problem of testing whether two samples of possibly right-censored survival data come from the same distribution is considered. The aim is to develop a test which is capable of detection of a wide spectrum of alternatives. A new class of tests based on Neyman's embedding idea is proposed. The null hypothesis is tested against a model where the hazard ratio of the two survival distributions is expressed by several smooth functions. A data-driven approach to the selection of these functions is studied. Asymptotic properties of the proposed procedures are investigated under fixed and local alternatives. Small-sample performance is explored via simulations which show that the power of the proposed tests appears to be more robust than the power of some versatile tests previously proposed in the literature (such as combinations of weighted logrank tests, or Kolmogorov–Smirnov tests).

1. Introduction

This paper presents a new approach to testing homogeneity of two samples of right-censored survival data. The goal is to provide an 'omnibus' test procedure sensitive against a range of alternatives. Such a test is useful, for instance, in situations when the Kaplan–Meier curves for the two samples do not suggest an alternative against which one should test (e.g., when the curves cross, as in the example in Section 7), or in situations when the visual inspection of the Kaplan–Meier plots is impossible (e.g., when data must be analysed automatically).

Omnibus tests cannot have power superior to all tests in all situations. Therefore, the objective is different: it is desirable to have a test which should not fail against a rather broad spectrum of alternatives. The method proposed in this paper achieves this goal.

Consider two samples of survival data. The jth sample consists of observations $(T_{j,i},\delta_{j,i})$, $i=1,\ldots,n_j$, $j=1,2$, where $T_{j,i}=R_{j,i}\overset{1}{\wedge}C_{j,i}$ is the possibly censored survival time, $\delta_{j,i}=1_{[R_{j,i}\leqslant C_{j,i}]}$ is the failure indicator, $R_{j,i}$ is the unobserved survival time and $C_{j,i}$ is the unobserved censoring time. The survival time and censoring time are assumed to be independent. All $n=n_1+n_2$ observations $(T_{j,i},\delta_{j,i})$ are mutually independent. The times $R_{j,i}$ come from a distribution with hazard function $\alpha_j(t)$. The aim is to test the hypothesis $\mathrm{H}_0:\alpha_1=\alpha_2$ without any specific alternative in mind.

[*] Corresponding author at: Institute of Information Theory and Automation, Pod Vodárenskou věží 4, CZ-18208 Praha 8, Prague, Czech Republic.
E-mail address: david.kraus@matfyz.cz
URL: http://www.davidkraus.net/.

0378-3758/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.04.009

The traditional approach is to use a weighted logrank test statistic $\int_0^\tau L(t)\,dU_0(t)$, where the logrank process equals

$$U_0(t) = \int_0^t \frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}(s)} \left[\frac{d\bar{N}_2(s)}{\bar{Y}_2(s)} - \frac{d\bar{N}_1(s)}{\bar{Y}_1(s)} \right].$$

Harrington and Fleming (1982) proposed to use weight functions from the $G^{\rho,\gamma}$ class of the form $L(t)=K(\hat{S}(t-))$ with $K(u)=u^\rho(1-u)^\gamma$, $\rho,\gamma \geqslant 0$ and $\hat{S}$ being an estimator of the survival function computed from the pooled sample (e.g., the Kaplan–Meier estimator or the exponential of minus the Nelson–Aalen estimator). Various members of this class are suitable for discovering various departures from the null hypothesis. Obviously, tests with $\rho > 0$ and $\gamma = 0$ are sensitive against early differences in hazard functions, tests with $\rho = 0$ and $\gamma > 0$ are powerful against late differences, a choice with $\rho > 0$ and $\gamma > 0$ yields a test good at detecting middle differences and the logrank test $G^{0,0}$ does well under proportional hazards. More precise results on performance of $G^{\rho,\gamma}$ tests are can be found in Fleming and Harrington (1991, Chapter 7) or Andersen et al. (1993, Section V.2). For instance, the logrank test $G^{0,0}$ is optimal (locally efficient) against the proportional hazards alternative (as it is the partial likelihood score test in a Cox model with a group indicator covariate) and the Prentice–Wilcoxon statistic $G^{1,0}$ is optimal against shift alternatives in the logistic distribution.

The $G^{\rho,\gamma}$ tests are directed against specific alternatives. While such a test is highly sensitive (often optimal) against the particular direction in the space of alternatives it may fail to detect different kinds of alternatives. One often does not have a clear advance idea of the nature of heterogeneity of the samples. Therefore, more omnibus tests were developed. Fleming and Harrington (1991, Section 7.5) describe two classes of such tests: tests using the whole path of the logrank process U_0 and tests combining several statistics of the $G^{\rho,\gamma}$ type. The former include supremum (Kolmogorov–Smirnov, KS) tests and integral tests (of the Cramér–von Mises (CM) and Anderson–Darling (AD) type). See also Gill (1980, Section 5.4) and Schumacher (1984). The latter class uses the maximum or sum of a finite cluster of weighted logrank statistics. Yet another procedure has been proposed by Pecková and Fleming (2003) who select a statistic from this cluster on the basis of estimated asymptotic relative efficiencies (within the cluster) against location shift alternatives.

Here I make a further step towards versatile tests with robust power, that is towards tests which on one hand do not collapse against a wide range of alternatives and on the other hand do not lose much compared to optimal directional tests.

My approach is based on Neyman's embedding idea combined with Schwarz's selection rule. The null nonparametric model of homogeneous samples is viewed as a submodel of a larger semiparametric model in which the hazard ratio of the two samples is expressed in terms of several smooth functions. A score test is applied to testing the null model versus the smooth model. Furthermore, selection criteria are used for choosing the smooth model. This data-driven strategy is inspired by the approach of Ledwina (1994) and Inglot et al. (1997). Smooth tests in the context of event history analysis were previously considered by Peña (1998a, b) and adaptive smooth tests by Kraus (2007).

In Section 2 the smooth test is constructed. Section 3 provides its data-driven version. Consistency of the proposed procedures is investigated in Section 4 while in Section 5, I study their behaviour under sequences of local alternatives. The Monte Carlo study of Section 6 explores level properties and power. The method is illustrated on a real data set in Section 7.

2. Construction of Neyman's test

Neyman's goodness-of-fit idea is used here as follows. The null model with $\alpha_1 = \alpha_2$ is embedded in a d-dimensional model

$$\alpha_2(t) = \alpha_1(t)\exp\{\theta^\mathsf{T}\psi(t)\}, \tag{1}$$

where $\theta = (\theta_1, \ldots, \theta_d)^\mathsf{T}$ is a parameter and $\psi(t) = (\psi_1(t), \ldots, \psi_d(t))^\mathsf{T}$ are some bounded functions modelling possible difference of α_2 from α_1. The functions $\psi_k(t)$ are taken in the form $\psi_k(t) = \varphi_k(g(t))$ where $\{\varphi_1, \ldots, \varphi_d\}$ forms a set of linearly independent continuous functions on $[0,1]$ and g is an increasing transformation that maps the time period $[0,\tau]$ to $[0,1]$.

The task of testing $\alpha_1 = \alpha_2$ versus (1) is equivalent to testing $\mathrm{H}_0 : \theta = 0$ versus $\mathrm{H}_d : \theta \neq 0$. It is advantageous to introduce the group indicator variable $Z_{j,i} = 1_{[j=2]}$. With this notation the intensities admit the form

$$\lambda_{j,i}(t) = Y_{j,i}(t)\alpha(t)\exp\{\theta^\mathsf{T}\psi(t)Z_{j,i}\}. \tag{2}$$

Hence we arrive at a Cox proportional hazards model with d artificial time-dependent covariates $\psi_1(t)Z_{j,i}, \ldots, \psi_d(t)Z_{j,i}$ whose significance is to be tested. To this end we may use well-known partial likelihood tools, of which the score test is particularly appealing as it does not involve estimation of θ.

Before proceeding to asymptotic considerations let us make some assumptions. Throughout the paper we assume the following standard regularity conditions.

Assumptions. (a) $a_j = \lim_{n\to\infty} n_j/n$ exists and $a_j \in (0,1), j = 1,2$.

(b) The survival functions S_j of failure times satisfy $S_j(\tau) > 0, j = 1,2$.

(c) The distribution functions G_j of censoring times satisfy $1 - G_j(\tau) > 0, j = 1,2$.

Denote by $\bar{y}_1, \bar{y}_2$ the uniform limits in probability of $n^{-1}\bar{Y}_1, n^{-1}\bar{Y}_2$, respectively; let $\bar{y}$ stand for $\bar{y}_1+\bar{y}_2$. By the Glivenko–Cantelli theorem these functions are $\bar{y}_j(t)=a_jS_j(t)(1-G_j(t))$. The above conditions guarantee that the limit functions are bounded away from zero on $[0,\tau]$.

The score process for the Cox model (2) under $\theta=0$ takes the form

$$U(t)=\int_0^t \psi(s)\frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}(s)}\left[\frac{d\bar{N}_2(s)}{\bar{Y}_2(s)}-\frac{d\bar{N}_1(s)}{\bar{Y}_1(s)}\right]=\int_0^t \psi(s)\,dU_0(s).$$

Then it is known (Fleming and Harrington, 1991, Corollary 7.2.1; Andersen et al., 1993, Theorem V.2.1) that under the hypothesis $\alpha_1=\alpha_2$ the logrank process $n^{-1/2}U_0$ converges weakly in $D[0,\tau]$ with Skorohod topology to a zero mean Gaussian martingale V_0 whose variance function and its uniformly consistent estimator are

$$\sigma_0(t)=\int_0^t \frac{\bar{y}_1(s)\bar{y}_2(s)}{\bar{y}(s)}\,dA(s),\quad n^{-1}\hat{\sigma}_0(t)=n^{-1}\int_0^t \frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}(s)}\frac{d\bar{N}(s)}{\bar{Y}(s)},$$

where $A(t)=\int_0^t \alpha(s)\,ds$ is the cumulative hazard function of the survival distribution. Consequently, under the null (i.e., $\theta=0$) the process $n^{-1/2}U$ is asymptotically distributed as a d-variate zero mean Gaussian martingale V with covariance matrix function and its estimator

$$\sigma(t)=\int_0^t \psi(s)^{\otimes 2}\,d\sigma_0(s),\quad n^{-1}\hat{\sigma}(t)=n^{-1}\int_0^t \psi(s)^{\otimes 2}\,d\hat{\sigma}_0(s)$$

(that is, $cov(V_k(s),V_l(t))=\sigma_{k,l}(s\wedge t)$).

The partial likelihood score statistic

$$T_d=U(\tau)^{\mathsf{T}}\hat{\sigma}(\tau)^{-}U(\tau)$$

used for testing $\theta=0$ versus $\theta\neq 0$ is asymptotically chi-squared distributed with d degrees of freedom. The hypothesis is rejected for large values of T_d.

As mentioned before, the basis functions $\varphi_1,\ldots,\varphi_d$ are linearly independent. We can take several functions from a well-known orthonormal basis of $L^2[0,1]$. For instance, we can use the cosine basis $\varphi_k(u)=\sqrt{2}\cos(k\pi u)$, $k=1,\ldots,d$, or orthonormal Legendre polynomials on $[0,1]$. It is natural (but not always necessary) to have the unity in the linear span of these functions in order to capture possible proportional hazards alternatives. We can set $\varphi_1\equiv 1$ (note that a model of the form (2) containing an intercept is identifiable) and the other functions may be cosines or Legendre polynomials.

Modelling the logarithm of the hazard ratio by linear combinations of smooth functions is a flexible approach. For instance, consider d=3 polynomials (of order 0, 1, 2). Their linear span contains the weight functions $G^{0,0}, G^{1,0}, G^{0,1}$ and $G^{1,1}$. Hence $\varphi_1,\varphi_2,\varphi_3$ can capture the same alternatives as the four logrank weights (proportional hazards, early, middle and late differences). Moreover, also nonlocation alternatives (crossing hazards) can be expressed by combinations of $\varphi_1,\varphi_2,\varphi_3$.

The time-transformation $g:[0,\tau]\to[0,1]$ may be simply $g(t)=t/\tau$. However, the purpose of the transformation is to standardise the speed of time. Following Kraus (2007) we can use $g(t)=F(t)/F(\tau)$ (with $F(t)=1-S(t)$ being the common distribution function of survival times) or $g(t)=A(t)/A(\tau)$. Alternatively, one may consider $g(t)=\sigma_0(t)/\sigma_0(\tau)$. If the functions $\varphi_1,\ldots,\varphi_d$ are orthonormal, this transformation yields a diagonal asymptotic covariance matrix (that is, the components of the score vector are asymptotically independent). For survival data Kraus (2007) advocates the transformation based on the distribution function.

In practice, a transformation depending on unknown quantities is replaced by a uniformly consistent estimator $\hat{g}$ computed from the pooled sample, for instance $\hat{g}(t)=\hat{F}(t)/\hat{F}(\tau)$. Denote g^* the limit of $\hat{g}$ ($\hat{g}$ converges to g under the null hypothesis or local alternatives, and to a mixture counterpart of g under the fixed alternative considered later). The difference between the score with $\hat{g}$ and g^* converges in probability to zero (by the Cauchy–Schwarz inequality and the fact that $\sup_{t\in[0,\tau]}|\varphi_k(\hat{g}(t))-\varphi_k(g^*(t))|\to 0$ in probability). Therefore, it is enough to obtain asymptotic results for theoretic deterministic weights. They are predictable and thus standard martingale tools apply. Then the results carry over to the practical case of nonpredictable estimated weights involving $\hat{g}$. In the following sections we do not explicitly repeat this argument in each proof.

3. Selection rules and adaptive tests

The difference between α_1 and α_2 is often well described by less than all d smooth functions. However, one does not know which functions should be included in the model and which not. Omitting a function that highly contributes to the description of the data or including an improper function may result in bad performance of the test. Therefore, it is reasonable to let the test adapt to the data, make the test data-driven.

This is accomplished by means of Schwarz's selection rule (or Bayesian information criterion, BIC). The adaptive test consists of two steps. First, a subset of $\{\varphi_1,\ldots,\varphi_d\}$ is selected on the basis of Schwarz's rule. Once a subset is selected, the score test against this likely alternative is performed.

We must specify a class $\mathscr{S}$ of nonempty index subsets out of which the selection rule will pick the most suitable one. I consider two classes previously proposed in the literature. Ledwina (1994) used d nested subsets of the form $\mathscr{S}^{\text{nested}}=\{\{1\},$

$\{1,2\},\dots,\{1,\dots,d\}\}$. This choice is reasonable when the basis functions are naturally ordered, e.g., according to increasing complexity (which is the case, for instance, for the cosine basis with increasing frequencies but hardly for the indicator basis). Claeskens and Hjort (2004) proposed to use all nonempty subsets of $\{1,\dots,d\}$, that is $\mathscr{S}^{\text{all}}=2^{\{1,\dots,d\}}\setminus\{\emptyset\}$. I also consider the strategy proposed by Janssen (2003). He suggested to prescribe a set of basis functions of primary interest which are always included. Without loss of generality, let these functions be several first basis functions, i.e., let their indices be $C_0=\{1,\dots,d_0\}$ for some d_0 (with $d_0=0$ meaning $C_0=\emptyset$). Then the class of subsets is $\{C\cup C_0 : C\in\mathscr{S}'\}$ (where $\mathscr{S}'$ may be $\mathscr{S}^{\text{nested}}$, $\mathscr{S}^{\text{all}}$, or some other class of nonempty sets).

Schwarz's criterion (a modification proposed by Ledwina, 1994) selects the set S maximising the penalised score statistic, i.e.,

$$S=\arg\max_{C\in\mathscr{S}}\{T_C-|C|\log n\},$$

where $|C|$ denotes the number of elements of C and T_C stands for the score statistic computed in the model with basis functions φ_k, $k\in C$. The adaptive test is based on T_S.

The asymptotic behaviour of the statistic T_S is given by the following theorem.

Theorem 1. *Denote $d^*=\min\{|C| : C\in\mathscr{S}\}$ (that is $d^*=\max(d_0,1)$). Then, under the null hypothesis, the selection criterion asymptotically concentrates in sets of dimension d^*, i.e., $\Pr[|S|=d^*]\to 1$ as $n\to\infty$. Consequently, T_S is asymptotically distributed as*

$$\max\{V_C(\tau)^{\mathsf{T}}\sigma_{CC}(\tau)^{-1}V_C(\tau) : C\in\mathscr{S}, |C|=d^*\},$$

where $V_C(\tau)$ and $\sigma_{CC}(\tau)$ are, respectively, the subvector and submatrix of $V(\tau)$ and $\sigma(\tau)$ corresponding to the subset C.

Proof. Any d^*-dimensional set C asymptotically wins against any set $\tilde{C}$ of dimension $k>d^*$ because $\Pr[T_{\tilde{C}}-k\log n<T_C-d^*\log n]=\Pr[T_{\tilde{C}}/\log n-T_C/\log n<k-d^*]\to 1$. Among d^*-dimensional sets the one whose score statistic is maximal is selected. Hence T_S has the same asymptotic distribution as $\max\{T_C : C\in\mathscr{S}, |C|=d^*\}$ which converges to the variable in the assertion of the theorem by weak convergence of the score vector and the continuous mapping theorem. □

When no high priority directions are specified ($d_0=0$) the nested subsets test statistic is approximately χ^2_1-distributed. Although asymptotically valid the χ^2_1 approximation is known to be inaccurate for small samples. A two-term approximation taking into account the possibility of selection not only of the set $\{1\}$ but also $\{1,2\}$ was applied in Kraus (2007, Eq. (12)), see further references therein.

For the all subsets rule with $d_0=0$, T_S converges to $\max\{V_1(\tau)^2/\sigma_{11}(\tau),\dots,V_d(\tau)^2/\sigma_{dd}(\tau)\}$, the maximum of generally dependent χ^2_1 variables. It may be easily approximated by simulation from the distribution of $V(\tau)$ (zero-mean normal with variance matrix estimated by $n^{-1}\hat{\sigma}(\tau)$).

With $d_0>0$ both nested and all subsets criteria give a statistic with asymptotic χ^2 distribution with d_0 degrees of freedom.

Small-sample accuracy of asymptotic approximations is investigated via simulations in Section 6.

Finally note that although the proposed tests aim at detecting a wide range of alternatives without a specific direction in mind, they can provide an idea of the type of departure from the null when the null is rejected. For instance, when the test using nested subsets and Legendre polynomials rejects and the selected dimension is 1, it suggests proportional hazards. Similarly, the hazard ratio is likely to be monotonic for $S=2$ and convex/concave for $S=3$. In this regard, the proposed tests differ from the Kolmogorov–Smirnov and related tests which give no such idea in case of rejection.

4. Consistency

Let us investigate when the smooth tests and their data-driven versions are consistent. Consider a fixed general alternative of different hazards in the two samples, i.e., $\alpha_1(t)\neq\alpha_2(t)$ on a nonnull set. Recall that $g^*(t)$ denotes the function to which $\hat{g}$ converges in probability. Under the fixed alternative the pooled sample Nelson–Aalen estimator $\hat{A}$ consistently estimates

$$A^*(t)=\int_0^t\left[\frac{\bar{y}_1(s)}{\bar{y}(s)}\alpha_1(s)\,ds+\frac{\bar{y}_2(s)}{\bar{y}(s)}\alpha_2(s)\,ds\right].$$

Therefore, for instance, for $\hat{g}(t)=\hat{A}(t)/\hat{A}(\tau)$ we have $g^*(t)=A^*(t)/A^*(\tau)$ and for $\hat{g}(t)=\hat{F}(t)/\hat{F}(\tau)$ we have $g^*(t)=F^*(t)/F^*(\tau)$ where $F^*(t)=1-\exp\{-A^*(t)\}$. Denote $\psi^*(t)=\varphi(g^*(t))$.

Theorem 2. *Smooth tests (both fixed-dimensional and data-driven) are consistent against any alternative satisfying*

$$\int_0^\tau\psi^*(t)\frac{\bar{y}_1(t)\bar{y}_2(t)}{\bar{y}(t)}(\alpha_2(t)-\alpha_1(t))\,dt\neq 0 \tag{3}$$

(i.e., at least one component is nonzero).

Proof. The left-hand side in (3) is the limit in probability of $n^{-1}U(\tau)$ under the alternative by consistency of the Nelson–Aalen estimators. The variance estimator $n^{-1}\hat{\sigma}(\tau)$ converges under the alternative to a finite matrix, namely

$$\int_0^\tau \psi^*(t)^{\otimes 2}\frac{\bar{y}_1(t)\bar{y}_2(t)}{\bar{y}(t)}\left[\frac{\bar{y}_1(t)}{\bar{y}(t)}\alpha_1(t)\,dt+\frac{\bar{y}_2(t)}{\bar{y}(t)}\alpha_2(t)\,dt\right].$$

Therefore, the limit of $n^{-1}T_d$ is nonzero, thus $T_d\to\infty$ in probability and consistency of the fixed-dimensional test follows. To see consistency of data-driven tests it remains to realise that for any subset $C\in\mathscr{S}$ containing at least one index corresponding to a nonzero component of (3) it holds that $T_C-|C|\log n\to\infty$ in probability (because $n^{-1}T_C$ converges to a positive number). Thus some of the subsets with the score statistic converging to infinity will be selected with probability converging to 1. Hence the data-driven test statistic T_S converges in probability to infinity which proves the assertion. □

Condition (3) may be interpreted as follows. Our working model is (2). The true form of the hazard functions is, however, more general: it may be rewritten as $\lambda_{j,i}(t)=Y_{j,i}(t)\alpha(t)\exp\{\eta(t)Z_{j,i}\}$, where the function η is nonzero on a nonnull set. Thus we work with a (possibly) misspecified Cox model. Struthers and Kalbfleisch (1986, Theorem 2.1) (see also Lin and Wei, 1989) show that the maximum partial likelihood estimator in a misspecified proportional hazards model converges to the solution to a limiting estimating equation. In our situation this limiting equation for θ is

$$\int_0^\tau \psi^*(t)\frac{\bar{y}_1(t)\bar{y}_2(t)}{\bar{y}_1(t)+\bar{y}_2(t)\exp\{\theta^{\mathsf{T}}\psi^*(t)\}}(\alpha_2(t)-\alpha_1(t)\exp\{\theta^{\mathsf{T}}\psi^*(t)\})\,dt=0.$$

Condition (3) just means that $\theta=0$ is not the solution to the limiting estimating equation, i.e., the estimate in the smooth model does not asymptotically fall to the null model. In other words, (3) says that the basis functions $\varphi_1,\dots,\varphi_d$ are not chosen completely wrong in the sense that at least some of them contributes to the approximation of η.

5. Behaviour under local alternatives

The aim of this section is to investigate the limit distribution of the test statistics under a sequence of local alternatives. Consider local alternatives of the form $\alpha_2(t)=\alpha_1(t)\exp\{n^{-1/2}\eta(t)\}$, where η is a bounded function.

Theorem 3. *Under the sequence of local alternatives*

$$\lambda_{j,i}(t)=Y_{j,i}(t)\alpha(t)\exp\{n^{-1/2}\eta(t)Z_{j,i}\}$$

the logrank process $n^{-1/2}U_0(t)$ converges weakly in $D[0,\tau]$ to the Gaussian process $\mu_0(t)+V_0(t)$, where the martingale V_0 is given in Section 2 and the mean function is

$$\mu_0(t)=\int_0^t\eta(s)\frac{\bar{y}_1(s)\bar{y}_2(s)}{\bar{y}(s)}\alpha(s)\,ds=\int_0^t\eta(s)\,d\sigma_0(s).$$

The process $n^{-1/2}U(t)$ converges to $\mu(t)+V(t)$ with $\mu(t)=\int_0^t\psi(s)\,d\mu_0(s)$, and, consequently, the statistic T_d is asymptotically distributed as a chi-squared variable with d degrees of freedom and noncentrality parameter $\mu(\tau)^{\mathsf{T}}\sigma(\tau)^{-1}\mu(\tau)$. The statistic T_S of the adaptive test converges weakly to

$$\max\{(\mu_C(\tau)+V_C(\tau))^{\mathsf{T}}\sigma_{CC}(\tau)^{-1}(\mu_C(\tau)+V_C(\tau)):C\in\mathscr{S},|C|=d^*\}.$$

Proof. The convergence of the logrank process is shown in Andersen et al. (1993, Section V.2.3). The convergence of $n^{-1/2}U$ and T_d is an immediate consequence. The results for data-driven tests follow from the fact that also along the sequence of local alternatives all variants of Schwarz's rule asymptotically concentrate in sets of the minimal dimension d^*. □

For nested subsets with $d_0=0$ the test behaves asymptotically under local alternatives like the directional test based on the first basis functions. (Note, however, that the test is consistent against the same alternatives as tests with all d functions.) This local behaviour was the motivation of Janssen (2003) for including high priority basis functions. Such tests (both with nested subsets and all subsets) behave asymptotically like the smooth test with d_0.

6. Numerical study

6.1. General information

We conducted simulations in order to examine the behaviour of the proposed tests and compare them with some of the existing two-sample procedures. We considered one situation satisfying the null hypothesis and several alternative configurations with hazard differences of various kind.

Table 1
Estimated sizes of fixed-dimensional and adaptive tests on the nominal level 5% with asymptotic critical values.

(n_1, n_2)	$d=4, d_0=0$				$d=7, d_0=4$	
	T_d	T_S^{nested}	T_S^{nested}	T_S^{all}	T_S^{nested}	T_S^{all}
	(χ_4^2)	(χ_1^2)	(two-term)	$(\max \chi_1^2)$	(χ_4^2)	(χ_4^2)
Censoring U(0, 10) (10%)						
(25, 25)	0.0664	0.1265	0.0695	0.0701	0.0945	0.1167
(50, 50)	0.0608	0.0960	0.0560	0.0600	0.0860	0.1084
(100, 100)	0.0600	0.0766	0.0554	0.0528	0.0772	0.0987
(200, 200)	0.0537	0.0656	0.0528	0.0516	0.0662	0.0848
(15, 35)	0.0769	0.1359	0.0770	0.0740	0.1158	0.1368
(30, 70)	0.0698	0.0960	0.0586	0.0586	0.0986	0.1215
(60, 140)	0.0636	0.0814	0.0604	0.0548	0.0832	0.1026
(120, 280)	0.0609	0.0695	0.0550	0.0519	0.0760	0.0944
Censoring U(0, 2) (43%)						
(25, 25)	0.0512	0.1132	0.0554	0.0620	0.0717	0.0898
(50, 50)	0.0548	0.0911	0.0536	0.0602	0.0710	0.0915
(100, 100)	0.0516	0.0701	0.0512	0.0542	0.0664	0.0854
(200, 200)	0.0522	0.0642	0.0490	0.0508	0.0632	0.0792
(15, 35)	0.0654	0.1238	0.0664	0.0734	0.0948	0.1129
(30, 70)	0.0572	0.0916	0.0542	0.0616	0.0785	0.0978
(60, 140)	0.0560	0.0762	0.0569	0.0566	0.0726	0.0899
(120, 280)	0.0534	0.0668	0.0535	0.0518	0.0654	0.0815

The distribution of survival times is unit exponential. Estimates based on 20 000 replications (standard deviation 0.0015).

Table 2
Estimated selection probabilities for subsets with dimension d^*, d^*+1, d^*+2 (three smallest dimensions) under the null hypothesis (unit exponential).

n	$d=4, d_0=0$			$d=7, d_0=4$		
	$\lvert S\rvert=1$	$\lvert S\rvert=2$	$\lvert S\rvert=3$	$\lvert S\rvert=4$	$\lvert S\rvert=5$	$\lvert S\rvert=6$
Nested subsets						
50	0.9366	0.0518	0.0094	0.9482	0.0444	0.0063
100	0.9607	0.0325	0.0060	0.9662	0.0294	0.0038
200	0.9770	0.0198	0.0026	0.9758	0.0219	0.0021
400	0.9848	0.0134	0.0016	0.9844	0.0145	0.0008
All subsets						
50	0.9804	0.0176	0.0014	0.8857	0.1101	0.0038
100	0.9891	0.0099	0.0009	0.9184	0.0796	0.0020
200	0.9938	0.0056	0.0003	0.9428	0.0559	0.0012
400	0.9967	0.0030	0.0003	0.9594	0.0400	0.0006

Censoring U(0, 2), various sample sizes $n = n_1 + n_2$ (with $n_1 = n_2$). Estimates based on 20 000 replications (standard deviation at most 0.0035).

Random numbers were generated using the Mersenne–Twister generator implemented in R (version 2.1.0). Twenty thousand Monte Carlo runs were performed under the null hypothesis, and 5000 for alternative situations. Smooth tests were used with the Legendre polynomial basis; the time transformation g was based on the distribution function.

6.2. Results on level

The behaviour of the test procedures under the null hypothesis is examined. We repeatedly generated two samples of unit exponential variables, censored them by independently generated uniform variables and performed the fixed-dimensional smooth test and both nested subsets and all subsets adaptive tests with and without specifying high priority basis functions. Various sample sizes n_1, n_2 and two censoring distributions (U(0, 10) and U(0, 2)) were considered. The tests were performed on the nominal level 5% using asymptotic critical values.

Table 1 provides empirical sizes. It is seen that the tests often exceed the nominal level. There are two sources of inaccuracy: bad performance of the asymptotic normal approximation for the score vector and slow convergence of selection criteria to the smallest dimension.

First, we may observe that when the censoring is light the fixed-dimensional test T_d is anticonservative even for rather large samples. A similar phenomenon could be observed for $G^{0,\gamma}$ tests (especially with $\gamma > 0$). Like these tests, our tests give some weight to late differences too.

A second, apparently more serious problem concerns data-driven tests. It is mainly seen for the nested subsets test with $d_0 = 0$ and for both variants with $d_0 > 0$ (here $d_0 = 4$) that the $\chi^2_{d^*}$ approximation is unacceptable even for the sample size 400. The reason of inaccuracy is the slow convergence of the selection criterion to the smallest dimension. Table 2 reports estimated selection probabilities for sets of three smallest dimensions. It shows that the concentration of $|S|$ in d^* is insufficient for small samples. There is an exception: the criterion with all subsets with $d_0 = 0$ is more concentrated in smallest (one-dimensional) sets and the asymptotic distribution (i.e., the maximum of χ_1^2 variables) performs much better (the size is comparable to the size of the test with fixed dimension). This is not so surprising because in this case the selection rule is asymptotically concentrated in d one-dimensional sets, whereas with the other classes of subsets the rule asymptotically selects one set (of dimension d^*). For the nested subsets criterion with $d_0 = 0$ we have the two-term approximation mentioned in Section 3. It successfully removes the problem of the slow convergence of S, the size is then similar to the size of the fixed-dimensional test (see Table 1).

To make the inference valid we use the permutation principle (Neuhaus, 1993). It assumes that the pairs $(N_{j,i}, Y_{j,i})$ (or $(T_{j,i}, \delta_{j,i})$ for survival data) are independent identically distributed under the null hypothesis and the distribution of the test statistic is exchangeable (permutation invariant). Hence, in the survival context, the censoring distributions in the two samples should be equal. The test is then exact. If the censoring distributions differ, the permutation procedure is valid asymptotically. Neuhaus (1993) and Heller and Venkatraman (1996) show that the permutation method remains reliable when the assumption of equal censoring distributions is not satisfied. Note that although Neuhaus (1993) focused on weighted logrank and Kolmogorov–Smirnov tests, the theoretical results of his Theorem 3.2 (asymptotic equivalence of distributions of the observed and permutation logrank process) justify the use of the permutation method also for the tests proposed in the present paper.

The permutation test was used with 2000 random permutations which seemed enough as the rejection probability was between 0.0470 and 0.0530 for all of the situations of Table 1. Note that alternatively instead of permutations (sampling without replacement) one may use the bootstrap (sampling with replacement); bootstrap results both under the null and under alternatives were very similar to permutation results.

Detailed results on the null behaviour of other two-sample tests are not reported. Just note that they often do not have size close to the nominal level when the asymptotic distribution is used. The weighted logrank $G^{0,\gamma}$ tests with normal approximation seriously exceed the level especially with light or without censoring. On the contrary, the Kolmogorov–Smirnov and other tests based on the logrank process are conservative. For these tests one may alternatively use the simulation approximation of Lin et al. (1993) which removes the conservatism to some extent (but not so well as permutations). Therefore, hereafter in simulations of power, all of the tests are performed using the permutation principle (with 2000 permutations).

6.3. *Alternative configurations*

Several configurations found in the literature were analysed. We report mainly situations previously studied by other authors not to be suspect of designing the study to favour the tests we propose. We investigated many other situations with similar conclusions. Configurations I–IV correspond to I–IV of Fleming et al. (1987), Configuration V corresponds to IV of Lee (1996). We admit that some of these alternatives may look somewhat peculiar when written in terms of hazard functions, they, however, do not look so when survival functions are plotted (see Fig. 1).

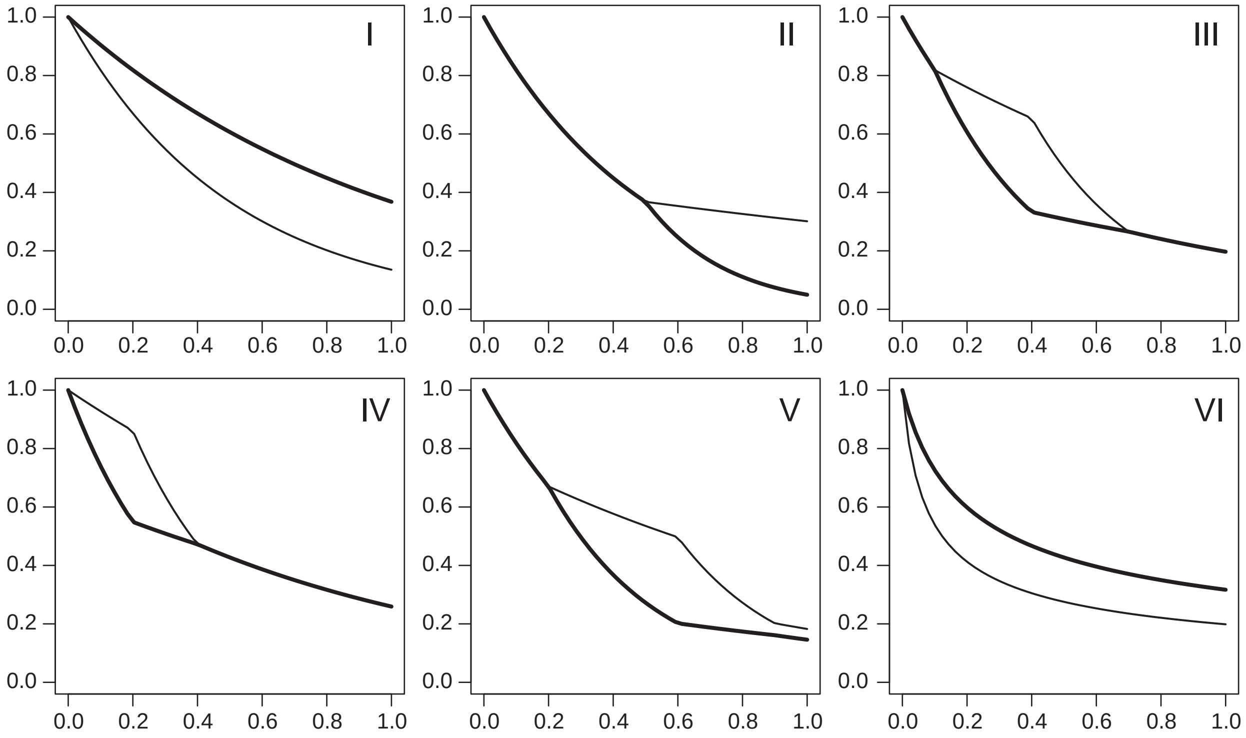

Fig. 1. Survival functions S_1 (thick lines) and S_2 (thin) under the simulation scenarios I–VI.

Table 3
Comparison of power for fixed-dimensional and various data-driven tests with various values of d and d_0.

d	T_d	$d_0 = 0$		$d_0 = 4$	
		T_S^{nested}	T_S^{all}	T_S^{nested}	T_S^{all}
Configuration I					
4	0.603	0.677	0.639	–	–
6	0.553	0.680	0.585	0.565	0.555
8	0.526	0.677	0.544	0.562	0.515
10	0.502	0.677	0.508	0.564	0.478
12	0.488	0.675	0.480	0.562	0.458
14	0.465	0.678	0.459	0.563	0.433
Configuration III					
4	0.713	0.507	0.678	–	–
6	0.817	0.539	0.763	0.771	0.789
8	0.804	0.542	0.751	0.767	0.764
10	0.784	0.542	0.734	0.770	0.750
12	0.753	0.541	0.721	0.770	0.729
14	0.736	0.541	0.707	0.767	0.708

Censoring U(0, 2), sample sizes $n_1 = n_2 = 50$, nominal level 5% (permutation test). Estimates based on 5000 replications (standard deviation at most 0.007).

Configuration I (proportional hazards):

$$\alpha_1(t) = 1, \quad \alpha_2(t) = 2.$$

Configuration II (late difference):

$$\alpha_1(t) = 2 \times 1_{[0,0.5)}(t) + 4 \times 1_{[0.5,\infty)}(t), \quad \alpha_2(t) = 2 \times 1_{[0,0.5)}(t) + 0.4 \times 1_{[0.5,\infty)}(t).$$

Configuration III (middle/early difference):

$$\alpha_1(t) = 2 \times 1_{[0,0.1)}(t) + 3 \times 1_{[0.1,0.4)}(t) + 0.75 \times 1_{[0.4,0.7)}(t) + 1_{[0.7,\infty)}(t),$$
$$\alpha_2(t) = 2 \times 1_{[0,0.1)}(t) + 0.75 \times 1_{[0.1,0.4)}(t) + 3 \times 1_{[0.4,0.7)}(t) + 1_{[0.7,\infty)}(t).$$

Configuration IV (early difference):

$$\alpha_1(t) = 3 \times 1_{[0,0.2)}(t) + 0.75 \times 1_{[0.2,0.4)}(t) + 1_{[0.4,\infty)}(t),$$
$$\alpha_2(t) = 0.75 \times 1_{[0,0.2)}(t) + 3 \times 1_{[0.2,0.4)}(t) + 1_{[0.4,\infty)}(t).$$

Configuration V (middle difference):

$$\alpha_1(t) = 2 \times 1_{[0,0.2)}(t) + 3 \times 1_{[0.2,0.6)}(t) + 0.75 \times 1_{[0.6,0.9)}(t) + 1_{[0.9,\infty)}(t),$$
$$\alpha_2(t) = 2 \times 1_{[0,0.2)}(t) + 0.75 \times 1_{[0.2,0.6)}(t) + 5 \times 1_{[0.6,0.9)}(t) + 1_{[0.9,\infty)}(t).$$

Configuration VI (proportional generalised odds):

$$\alpha_j(t) = e^{\beta_j}/(1 + 2e^{\beta_j}t), \quad j = 1, 2 \text{ with } \beta_1 = 1.5, \ \beta_2 = 2.5.$$

(The survival functions are $S_j(t) = H(\log t + \beta_j)$ with $H(x) = (1 + 2e^x)^{-1/2}$, $x \in \mathbb{R}$. For this situation the $G^{2,0}$ test is efficient.)

The sample size was always 100 (each group 50), the censoring distribution was uniform on (0, 2) giving censoring rates from 28% to 38%.

6.4. Comparison of fixed-dimensional and various data-driven tests

In Table 3, several variants of smooth tests are compared in two of the considered situations. For Configuration I (proportional hazards) the best test is with $d = 1$, hence it is not surprising that increasing d decreases the power (other basis functions than $\varphi_1 \equiv 1$ are superfluous). In Configuration III we can see that the power decreases for $d > 6$ ($d = 6$ gives the best power because the hazard ratio is rather complicated and its description requires several functions). Now let us see how the data-driven tests with various classes of subsets behave.

First consider $d_0 = 0$ (no basis functions of primary interest). The all subsets version seems to suffer from the same problem as the test with a fixed dimension: when d is too high, the power decays. This is caused by the dependence of the null distribution of the test statistic on d. The nested subsets criterion gives stable power for various values of d in both configurations. In Configuration I, this test has higher power than the other tests because the selection rule mostly selects the smallest set containing only the intercept which describes the data well. On the other hand, in Configuration III, the concentration in the set {1} negatively affects the power because the constant basis function cannot catch the true shape of the hazard ratio.

Table 4
Comparison of power of various two-sample tests.

	Configuration						Robustness of power
	I	II	III	IV	V	VI	
$G^{0,0}$	0.794	0.338	0.234	0.133	0.307	0.466	0.160
$G^{2,0}$	0.655	0.056	0.357	0.562	0.171	0.581	0.064
$G^{0,2}$	0.515	0.875	0.069	0.097	0.122	0.115	0.086
$G^{2,2}$	0.676	0.306	0.239	0.134	0.590	0.233	0.161
T^{sum}	0.779	0.499	0.217	0.141	0.343	0.389	0.169
T^{max}	0.733	0.795	0.316	0.471	0.458	0.476	0.393
KS-W	0.770	0.273	0.556	0.557	0.470	0.519	0.312
KS-B	0.721	0.193	0.617	0.807	0.450	0.547	0.221
CM-W	0.701	0.057	0.482	0.512	0.320	0.571	0.065
CM-B	0.623	0.051	0.425	0.739	0.191	0.575	0.058
AD-W	0.643	0.052	0.434	0.696	0.227	0.578	0.059
AD-B	0.577	0.051	0.356	0.767	0.150	0.558	0.058
T_d $(d=4)$	0.599	0.854	0.713	0.762	0.546	0.363	0.625
T_d $(d=8)$	0.525	0.796	0.803	0.832	0.669	0.278	0.478
T_S^{nested} $(d=8, d_0=0)$	0.677	0.803	0.539	0.734	0.418	0.411	0.624
T_S^{all} $(d=8, d_0=0)$	0.541	0.684	0.750	0.790	0.608	0.280	0.481
T_S^{nested} $(d=8, d_0=4)$	0.563	0.825	0.770	0.795	0.634	0.316	0.544
T_S^{all} $(d=8, d_0=4)$	0.518	0.794	0.766	0.787	0.622	0.277	0.476

Censoring U(0, 2), sample sizes $n_1 = n_2 = 50$, nominal level 5% (permutation test). Estimates based on 5000 replications (standard deviation at most 0.007).

Now let $d_0 = 4$. Again, the power of the test with nested subsets is stable. For Configuration I it is now lower than with $d_0 = 0$ (because now the BIC concentrates in the set $\{1, 2, 3, 4\}$ instead of $\{1\}$), for Configuration III the power is higher (four basis functions catch the difference better than one). The power of the all subsets test decreases with increasing d which is somewhat surprising since, unlike with $d_0 = 0$, the null distribution now does not depend on d. It perhaps may be explained by the slow convergence of the criterion to the four-dimensional set as already seen in Table 2 (with the all subsets criterion the limiting set $\{1, 2, 3, 4\}$ has much more competitors than with nested sets).

To summarise, mainly the nested subsets approach helps to avoid the use of too many components.

6.5. *Comparison with other tests*

Let us compare Neyman's smooth tests with other two-sample methods. Firstly, we consider weighted logrank tests with $G^{0,0}$, $G^{2,0}$, $G^{0,2}$ and $G^{2,2}$ weights and tests combining these four statistics (the statistic T^{sum} equals the sum of absolute values of these four standardised statistic while T^{max} equals their maximum). Secondly, functionals of the whole path of the logrank process leading to the Kolmogorov–Smirnov, Cramér–von Mises and Anderson–Darling type tests are considered. They are of two kinds: those using the untransformed process (denoted KS-W, CM-W, AD-W) and those transformed in the Hall–Wellner way (denoted KS-B, CM-B, AD-B) (the former process is asymptotically a Brownian motion, whereas the latter converges to a Brownian bridge, both in transformed time); for details see Section V.4.1 of Andersen et al. (1993). All of these tests are performed as two-sided since Neyman's tests are naturally two-sided. In all situations, the permutation approach is employed.

Table 4 presents estimated powers for the above situations I–VI. Before looking at the results we should realise what we expect from versatile tests. Certainly, it is impossible to hope that they will outperform all other methods in all situations. Rather, one may wish to have tests whose behaviour is not bad under a broad range of situations, that is, one seeks tests with robust power.

To assess robustness of power we computed a quantity, presented in the last column of the table, as follows. For each situation (each column) the ratio of the power of each test and the power of the best test (in the column) is computed. Then for each test (each row) the minimum of these ratios is presented as a measure of robustness of power. In other words, the last column contains row minima of standardised powers (where standardisation means division by column maxima).

As expected, directional tests $G^{\rho,\gamma}$ have very low robustness scores because they perform excellent in situations they are designed for but often do very bad for other situations. Among versatile tests previously proposed in the literature the test T^{max} as well as the Kolmogorov–Smirnov type tests (mainly KS-W) appear to have more stable power. The behaviour of power of smooth tests proposed in this paper seems much better (regarding stability over various alternatives) than of the other versatile tests. Smooth tests of course often lose against some of the other tests but not so much as the other tests sometimes do. These conclusions should be looked at with caution as they are based on the limited set of Configurations I–VI. However, we studied various other situations but never found smooth tests completely failing.

A closer look at results for various configurations reveals several findings. Tests employing the untransformed logrank process detect late differences better than those with the transformed process and vice versa (because the Hall–Wellner transformation downweights late differences). Surprisingly, the integral type tests completely failed in Configuration II, thus they do not appear as versatile as expected (this may be probably explained by the Karhunen–Loève decomposition of the test process and principal components analysis of the integral statistics but I will not pursue this examination here).

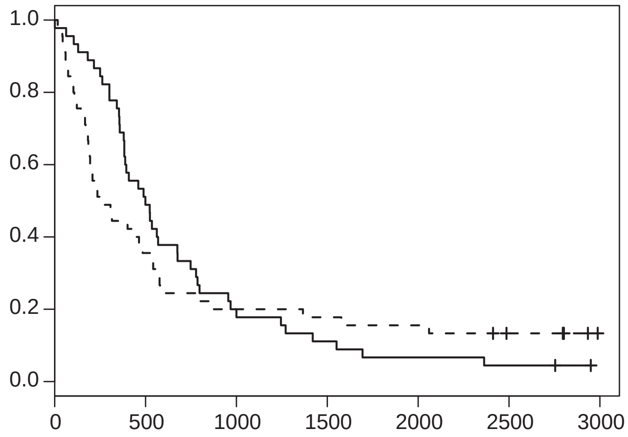

Fig. 2. Kaplan–Meier estimates for chemotherapy (solid) and chemotherapy plus radiotherapy (dashed) for the gastric cancer data. Survival times in days.

The behaviour of $T^{\max}$ in Configurations III and V is interesting. Situation III was termed a middle difference in Fleming et al. (1987) but it rather seems to be something between a middle and early difference as is seen from powers of $G^{\rho,\gamma}$ tests. None of $G^{0,0}$, $G^{2,0}$, $G^{2,2}$ tests clearly dominates but $T^{\max}$ must choose one of them. Hence, the $T^{\max}$ test loses some power compared to smooth tests which can combine more than one direction (no matter that directions are given by different functions for the two approaches). In Configuration V studied by Lee (1996) the difference is more clear in the middle ($G^{2,2}$ is much better than the other $G^{\rho,\gamma}$ tests), so $T^{\max}$ does better. In this regard, the behaviour of the adaptive test of Pecková and Fleming (2003) will be similar as this test is also forced to select one of the weighted logrank statistics (simulations in their paper show that the power of this test in most cases lies above the power of $T^{\max}$ and below the best power in the cluster).

6.6. *Summary*

All versions of smooth tests provide a procedure with power that seems to be more stable than power of other methods. The test with a relatively small fixed dimension (e.g., 3 or 4) does well quite often. Among data-driven tests, mainly the nested subsets approach helps to avoid the use of too many components. The adaptive dimension selection with nested subsets with $d_0 = 1$ (or perhaps $d_0 = 2$) slightly prefers simpler alternatives, which reflects the natural idea that simple situations occur in reality more often than complicated ones. The nested subsets selection procedure is not sensitive with respect to the choice of the maximum dimension d if d is large enough to cover realistic departures from the hypothesis (in practice, d equal to, say, 6 should be enough in most situations). Regarding the basis functions, results not reported here suggest that their choice does not affect the behaviour of the tests significantly.

7. Illustration

Stablein and Koutrouvelis (1985) studied data from a trial comparing two types of treatment of gastric cancer: chemotherapy versus chemotherapy combined with radiotherapy. There were 45 patients in each group (2 and 6 were censored, respectively). Fig. 2 displays crossing survival curves. It is not obvious against which alternative we should test, hence a versatile test is handy. On the conventional level 5% all of them reject the hypothesis of no difference between the two treatments.

The test statistic of Neyman's smooth test with $d = 8$ Legendre polynomials (of order $0, \ldots, 7$) is 17.55 with p-value 0.023 (based on 5000 permutations). The selection rule with $d_0 = 4$ selects the smallest possible set $\{1, 2, 3, 4\}$ for both nested and all subsets search. The test statistic equals 13.59 with $p = 0.018$ for nested subsets and $p = 0.03$ for all subsets. If no functions of primary interest are specified ($d_0 = 0$) then the nested subsets criterion selects $\{1, 2\}$ with the statistic 13.45 and p-value 0.005 while the all subsets rule gives the set $\{2\}$, statistic 13.32 and p-value 0.01.

The tests $G^{0,0}$, $G^{2,0}$, $G^{0,2}$, $G^{2,2}$ have statistics 0.47 ($p = 0.637$), 2.59 (0.009), 1.99 (0.053), 0.41 (0.684), respectively. The p-value of the maximal statistic 2.59 is 0.021. The value of the KS-W statistic is 2.20 with $p = 0.047$, the KS-B statistic equals 1.58 with $p = 0.008$.

Acknowledgements

I would like to thank two anonymous referees and the coordinating editor for their detailed constructive reviews which made me improve the paper significantly. I also thank the executive editor and the coordinating editor for their patience during the revision period.

The work has been supported by the GAČR Grant 201/05/H007, GAAV Grant IAA101120604 and MŠMT Project 1M06047. Computations have been carried out in METACentrum (Czech academic supercomputer network).

References

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

Claeskens, G., Hjort, N.L., 2004. Goodness of fit via non-parametric likelihood ratios. Scand. J. Statist. 31, 487–513.

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

Fleming, T.R., Harrington, D.P., O'Sullivan, M., 1987. Supremum versions of the log-rank and generalized Wilcoxon statistics. J. Amer. Statist. Assoc. 82, 312–320.

Gill, R.D., 1980. Censoring and Stochastic Integrals. Mathematical Centre Tracts, vol. 124. Mathematisch Centrum, Amsterdam.

Harrington, D.P., Fleming, T.R., 1982. A class of rank test procedures for censored survival data. Biometrika 69, 553–566.

Heller, G., Venkatraman, E.S., 1996. Resampling procedures to compare two survival distributions in the presence of right-censored data. Biometrics 52, 1204–1213.

Inglot, T., Kallenberg, W.C.M., Ledwina, T., 1997. Data driven smooth tests for composite hypotheses. Ann. Statist. 25, 1222–1250.

Janssen, A., 2003. Which power of goodness of fit tests can really be expected: intermediate versus contiguous alternatives. Statist. Decisions 21, 301–325.

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

Ledwina, T., 1994. Data-driven version of Neyman's smooth test of fit. J. Amer. Statist. Assoc. 89, 1000–1005.

Lee, J.W., 1996. Some versatile tests based on the simultaneous use of weighted log-rank statistics. Biometrics 52, 721–725.

Lin, D.Y., Wei, L.J., 1989. The robust inference for the Cox proportional hazards model. J. Amer. Statist. Assoc. 84, 1074–1078.

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

Neuhaus, G., 1993. Conditional rank tests for the two-sample problem under random censorship. Ann. Statist. 21, 1760–1779.

Pecková, M., Fleming, T.R., 2003. Adaptive test for testing the difference in survival distributions. Lifetime Data Anal. 9, 223–238.

Peña, E.A., 1998a. Smooth goodness-of-fit tests for composite hypothesis in hazard based models. Ann. Statist. 26, 1935–1971.

Peña, E.A., 1998b. Smooth goodness-of-fit tests for the baseline hazard in Cox's proportional hazards model. J. Amer. Statist. Assoc. 93, 673–692.

Schumacher, M., 1984. Two-sample tests of Cramér–von Mises- and Kolmogorov–Smirnov-type for randomly censored data. Internat. Statist. Rev. 52, 263–281.

Stablein, D.M., Koutrouvelis, I.A., 1985. A two-sample test sensitive to crossing hazards in uncensored and singly censored data. Biometrics 41, 643–652.

Struthers, C.A., Kalbfleisch, J.D., 1986. Misspecified proportional hazard models. Biometrika 73, 363–369.