

Smooth tests of equality of cumulative incidence functions in two samples¹

BY DAVID KRAUS

*Institute of Information Theory and Automation,
Pod Vodárenskou věží 4, 182 08 Prague 8, Czech Republic*

david.kraus@matfyz.cz

SUMMARY

In this paper, a method is developed for comparison of two samples of survival data with competing risks. In competing risks data the probability that a failure from a particular cause in the presence of other risks of failure occurs by some time is summarised by the cumulative incidence function. A new test is proposed for the hypothesis that cumulative incidence functions for a particular type of failure are equal in two samples. The procedure is based on Neyman's idea of smooth tests. The new test has stable power against a wider spectrum of alternatives than tests previously proposed in the literature. In particular, the method exhibits much better power in situations with crossing curves. Asymptotic results, simulations and a real example are presented.

Some key words: Competing risks; Cumulative incidence function; Neyman's smooth test; Two-sample test.

1. INTRODUCTION

In competing risks situations, individuals may fail from one of K causes. Observations consist of the failure time and the cause of failure. Formally, let $R \geq 0$ be the survival time and let $\varepsilon \in \{1, \dots, K\}$ be the cause of death. There are two main quantities describing the occurrence of events of type k : the cause-specific hazard rate (crude transition intensity)

$$\alpha(t, k) = \lim_{\Delta \rightarrow 0} \frac{\text{pr}(t \leq R < t + \Delta, \varepsilon = k | R \geq t)}{\Delta},$$

and the cumulative incidence function

$$F(t, k) = \text{pr}(R \leq t, \varepsilon = k) = \int_0^t S(s) \alpha(s, k) ds,$$

where $S(t) = \text{pr}(R > t)$ is the overall survival function. Observations are allowed to be right-censored, that is, we actually observe (T, δ) , $T = R \wedge C$, $\delta = \varepsilon 1[R \leq C]$, where the censoring time C is independent of R and ε .

I consider two samples of such data: $(T_{j,i}, \delta_{j,i})$, $j = 1, 2$, $i = 1, \dots, n_j$. The goal is to compare the occurrence of failures from one particular cause, say 1. Without loss of generality I assume that the number of possible endpoints K is 2; all event types different from 1, which are not of interest, may be merged in type 2. The comparison of two samples can be done either in terms of the cause-specific hazards $\alpha_j(\cdot, 1)$ or in terms of the cumulative incidence

¹This is Research Report 2197, Institute of Information Theory and Automation, Prague, 9 October 2007.

functions $F_j(\cdot, 1)$. Note that the hypotheses $\alpha_1(\cdot, 1) = \alpha_2(\cdot, 1)$ and $F_1(\cdot, 1) = F_2(\cdot, 1)$ are not equivalent, which is explained, e.g., by Gray (1988) or Lin (1997) and vividly illustrated by simulations of Bajorunaite & Klein (2007). Cause-specific hazard rates can be compared by standard methods (e.g., the logrank test) by working with failures from the other causes as with censored observations. On the other hand, the comparison of cumulative incidence curves requires special methods. In this paper, I develop a method for testing the nonparametric null hypothesis $F_1(\cdot, 1) = F_2(\cdot, 1)$ against the alternative that these functions differ.

Several tests have been previously proposed for this task. Gray (1988) developed a class of tests based on weighted integrals with respect to the difference of estimated cumulative subdistribution hazard functions corresponding to the subdistributions $F_j(t, 1)$, defined as $\Gamma_j(t, k) = -\log(1 - F_j(t, k))$. The test statistic follows the form

$$\int_0^\tau L(t)(d\hat{\Gamma}_2(t, 1) - d\hat{\Gamma}_1(t, 1)),$$

where $L(t)$ is a weight function (a predictable process), $\hat{\Gamma}_j(t, 1)$ are consistent estimates of $\Gamma_j(t, 1)$, and $\tau < \infty$ is the end of the observation period $[0, \tau]$. These tests are good for detection of ordered subdistribution hazards $\gamma_j(t, 1) = d\Gamma_j(t, 1)/dt$ but may fail to detect crossing subdistribution hazards. Note that these statistics are similar to weighted logrank statistics for the traditional situation with one type of failure. For this similarity, I call this test the logrank-type test, which should not be confused with the ordinary logrank test applied to cause-specific hazards: the ordinary logrank test compares cause-specific hazards whereas Gray's logrank-type test compares subdistribution hazards.

Another test was proposed by Pepe (1991) who used the integrated difference of estimates of the cumulative incidence functions

$$\int_0^\tau (\hat{F}_2(t, 1) - \hat{F}_1(t, 1))dt.$$

This test often possesses good power against ordered cumulative incidence functions but may be less powerful when these curves cross. Note that this test is not a rank test.

Pepe's integral test and Gray's logrank-type test lose some power against alternatives with crossing curves (cumulative incidences or subdistribution hazards) because positive differences in some part of the observation period are negated by negative differences in another part. Lin (1997) suggested to use a Kolmogorov–Smirnov type test based on the supremum of the absolute value of the difference of estimated cumulative incidence functions $\sup_{t \in [0, \tau]} |\hat{F}_2(t, 1) - \hat{F}_1(t, 1)|$. While such a test is theoretically consistent against any alternative, its power is low quite often.

Therefore, it is desirable to develop a test which is good at detecting a spectrum of practically relevant alternatives (including crossing situations not covered by tests of Gray (1988) and Pepe (1991)) with better performance than the supremum test proposed by Lin (1997). This paper deals with the class of Neyman's smooth tests.

Situations with complicated departures from the hypothesis (such as crossing functions) occur in real applications. As an example I consider data from a bone marrow transplant study discussed by Bajorunaite & Klein (2007). The treatment of leukaemia by the bone marrow transplantation may fail from one of two causes: recurrence of the disease (relapse), and death

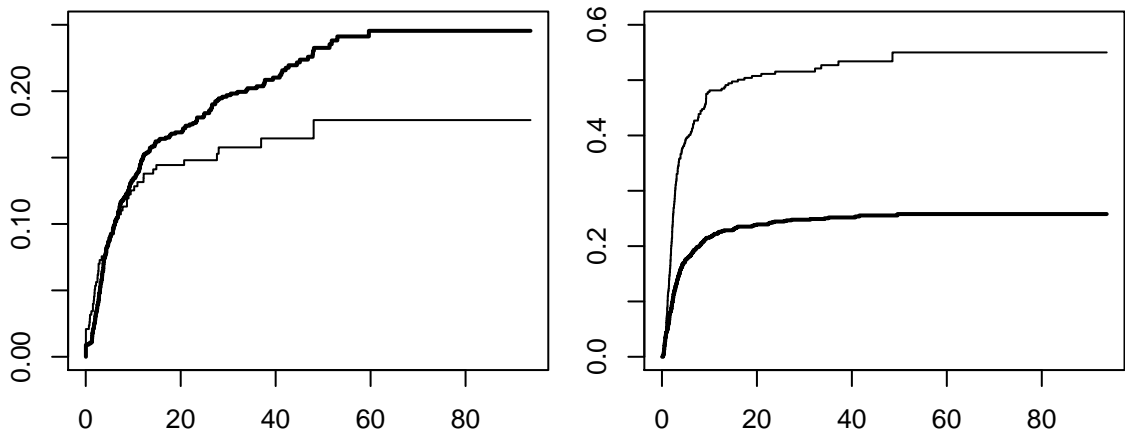


FIGURE 1. Cumulative incidence functions for relapse (left panel) and death in remission (right panel) for HLA-identical sibling donors (thick lines) and HLA-matched unrelated donors (thin lines). Time from the bone marrow transplantation is in months.

in remission (treatment-related death). There are two groups of patients to be compared: 1224 individuals with a human leukocyte antigen (HLA) identical sibling donor, and 383 with an HLA-matched unrelated donor. Figure 1 shows estimates of cumulative incidence functions for both samples for each type of treatment failure. Numerical results of Section 4 show that (some of) tests sensitive against ordered alternatives do not detect the difference between relapse cumulative incidence curves, which contradicts the visual impression. The reason is that the relapse cumulative incidence functions for these two groups cross.

The structure of this paper is as follows. In Section 2 the Neyman-type smooth test is constructed. Section 3 presents results of a simulation study. Results for the bone marrow transplant data are reported in Section 4. Asymptotic results are formulated and proved in Appendix.

2. NEYMAN'S EMBEDDING AND DEVELOPMENT OF THE SCORE TEST

In this section I show how Neyman's embedding idea can be applied in the two-sample competing risks situation. Neyman's smooth goodness-of-fit procedure is based on embedding the null hypothesis in a 'smooth' alternative model described by a finite number of parameters.

Traditionally (Rayner & Best, 1989), the embedding is formulated in terms of densities. In the goodness-of-fit problem of testing the simple hypothesis that the data come from a distribution with density $f = f_0$ the null hypothesis is embedded into the d -dimensional alternative

$$f(x; \theta) = f_0(x) \exp\{\theta^\top \varphi(F_0(x)) - c(\theta)\}, \quad x \in \mathbb{R}, \quad (1)$$

where $\varphi(u) = (\varphi_1(u), \dots, \varphi_d(u))^T$, $u \in [0, 1]$ are some square integrable basis functions, $c(\theta) = \log \int_{\mathbb{R}} f_0(x) \exp\{\theta^T \varphi(F_0(x))\} dx$ is a normalising constant and F_0 the distribution function corresponding to f_0 . The general alternative $f \neq f_0$ is replaced by $\theta \neq 0$. Neyman's test is the score test of $\theta = 0$ in the above model. Note that the function $f(x; \theta)$ is properly normalised, i.e., any value θ gives rise to a possible alternative distribution.

In the standard (single endpoint) survival context the embedding is most conveniently achieved in terms of hazard functions. The alternative takes the form

$$\alpha(t; \theta) = \alpha_0(t) \exp\{\theta^T \varphi(F_0(t))\}, \quad t \geq 0. \quad (2)$$

See Peña (1998) for details and extensions to composite hypotheses, and Kraus (2007) for an application in regression. Notice that in this formulation there is no normalising constant. The integral of a hazard function may be arbitrary positive. Therefore, any value of θ yields a well-defined hazard rate.

Let us turn to competing risks problems. First, consider a one sample situation with a simple hypothesis of the full specification of the cumulative incidence function for failures of type 1, that is $F(\cdot, 1) = F_0(\cdot, 1)$. Recall that $F(t, 1) = \int_0^t f(s, 1) ds$, where $f(t, 1) = S(t)\alpha(t, 1)$. The first idea is to formulate a smooth alternative in terms of $f(t, 1)$. This strategy is, however, infeasible because $f(t, 1)$ is a subdistribution density. That is, its integral $F(\infty, 1)$ is neither fixed nor unbounded ($F(t, 1)$ is a subdistribution function, hence $F(\infty, 1)$ may be anything between 0 and 1). The smooth alternative cannot be expressed in the form (1) since there is no normalising constant. On the other hand, we cannot use an unnormalised form like (2) because of the upper bound 1 for the integral of a subdensity. The cause-specific hazard $\alpha(t, 1)$ could be embedded similarly to (2) but hypotheses about cause-specific hazards are not equivalent to hypotheses about cumulative incidence functions. However, there is a characteristic suitable for embedding: the subdistribution hazard function $\gamma(t, 1)$ defined as

$$\gamma(t, k) = \frac{d}{dt} \Gamma(t, k) = \frac{f(t, k)}{1 - F(t, k)},$$

where $\Gamma(t, k) = -\log(1 - F(t, k))$. The functions $\gamma(t, 1)$ and $\Gamma(t, 1)$ may be seen as the hazard rate and the cumulative hazard function of the subdistribution $F(t, k)$ of the improper random variable $\tilde{R}^{(k)}$ defined by $\tilde{R}^{(k)} = R$ if $\varepsilon = k$, $\tilde{R}^{(k)} = \infty$ otherwise. Due to the one-to-one correspondence between $F(t, k)$ and $\gamma(t, k)$ hypotheses about $F(t, 1)$ and $\gamma(t, 1)$ are equivalent.

Now consider the two-sample hypothesis $F_1(\cdot, 1) = F_2(\cdot, 1) = F_0(\cdot, 1)$, equivalently $\gamma_1(\cdot, 1) = \gamma_2(\cdot, 1) = \gamma_0(\cdot, 1)$. The null model is viewed as a submodel of

$$\gamma_2(t, 1) = \gamma_1(t, 1) \exp\{\theta^T \psi(t)\}.$$

The logarithm of the subdistribution hazard ratio is expressed as a linear combination of some functions. Here $\psi_l(t)$, $t \in [0, \tau]$, $l = 1, \dots, d$ are of the form $\psi_l(t) = \varphi_l(F_0(t, 1)/F_0(\tau, 1))$, where $\varphi_l(u)$, $u \in [0, 1]$ are some linearly independent basis functions (for example, orthogonal Legendre polynomials of order $0, 1, \dots, d-1$, or cosines $\sqrt{2} \cos((l-1)\pi u)$).

I shall develop a score test of the hypothesis $\theta = 0$ versus $\theta \neq 0$.

The observations $(T_{j,i}, \delta_{j,i})$ can be represented by marked point processes as follows. For $k \in \{1, 2\}$ denote $N_{j,i}(t, k) = 1[T_{j,i} \leq t, \delta_{j,i} = k]$, the counting process counting events

of type k up to time t on the i th individual of the j th sample. Its intensity process is $\lambda_{j,i}(t, k) = Y_{j,i}(t)\alpha_j(t, k)$, where $Y_{j,i}(t) = 1[T_{j,i} \geq t]$ is the risk indicator process.

Estimators used in the following derivations are

$$\hat{F}_j(t, k) = \int_0^t \hat{S}_j(s-) \frac{d\bar{N}_j(s, k)}{\bar{Y}_j(s)}, \quad \hat{\Gamma}_j(t, k) = \int_0^t \frac{d\hat{F}_j(s, k)}{1 - \hat{F}_j(s-, k)} = \int_0^t \frac{d\bar{N}_j(s, k)}{\bar{R}_j(s, k)},$$

where \hat{S}_j is the Kaplan–Meier estimator of S_j , $\bar{N}_j = \sum_{i=1}^{n_j} N_{j,i}$, $\bar{Y}_j = \sum_{i=1}^{n_j} Y_{j,i}$, and $\bar{R}_j(t, k) = \bar{Y}_j(t)(1 - \hat{F}_j(t-, k))/\hat{S}_j(t-)$. Under the null hypothesis, there are consistent pooled sample estimators

$$\hat{F}_0(t, 1) = \int_0^t \frac{d\bar{N}_1(s, 1) + d\bar{N}_2(s, 1)}{\bar{Y}_1(s)/\hat{S}_1(s-) + \bar{Y}_2(s)/\hat{S}_2(s-)}, \quad \hat{\Gamma}_0(t, 1) = \int_0^t \frac{d\bar{N}_1(s, 1) + d\bar{N}_2(s, 1)}{\bar{R}_1(s, 1) + \bar{R}_2(s, 1)}$$

introduced by Gray (1988, formulae (2.11) and (2.5)).

The logarithm of the likelihood takes the form

$$\begin{aligned} & \sum_{j=1}^2 \sum_{i=1}^{n_j} \int_0^\tau \sum_{k=1}^2 \log(\lambda_{j,i}(t, k)) dN_{j,i}(t, k) - \sum_{j=1}^2 \sum_{i=1}^{n_j} \int_0^\tau \sum_{k=1}^2 \lambda_{j,i}(t, k) dt \\ &= \sum_{j=1}^2 \int_0^\tau \sum_{k=1}^2 \log(\alpha_j(t, k)) d\bar{N}_j(t, k) - \sum_{j=1}^2 \int_0^\tau \bar{Y}_j(t) \sum_{k=1}^2 \alpha_j(t, k) dt. \end{aligned}$$

Using the relation $\gamma_j(t, k) = S_j(t)\alpha_j(t, k)/(1 - F_j(t, k))$ we get

$$\begin{aligned} & \sum_{j=1}^2 \int_0^\tau \sum_{k=1}^2 \log(\gamma_j(t, k)) d\bar{N}_j(t, k) + \sum_{j=1}^2 \int_0^\tau \sum_{k=1}^2 \log\left(\frac{1 - F_j(t, k)}{S_j(t)}\right) d\bar{N}_j(t, k) \\ & \quad - \sum_{j=1}^2 \int_0^\tau \bar{Y}_j(t) \gamma_j(t, 1) \frac{1 - F_j(t, 1)}{S_j(t)} dt - \sum_{j=1}^2 \int_0^\tau \bar{Y}_j(t) \gamma_j(t, 2) \frac{1 - F_j(t, 2)}{S_j(t)} dt. \end{aligned}$$

In the above expression not only $\gamma_j(t, 1)$ but also $F_j(t, k)$ and $S_j(t)$ depend on the parameter θ . However, things simplify when we replace $F_j(t, k)$ and $S_j(t)$ by their consistent estimators $\hat{F}_j(t, k)$ and $\hat{S}_j(t)$. Then only the first and third term contain θ . Taking derivatives with respect to θ we arrive at

$$\int_0^\tau \psi(t) \left[d\bar{N}_2(t, 1) - \bar{Y}_2(t) \frac{1 - \hat{F}_j(t, 2)}{\hat{S}_j(t)} \exp\{\theta^\top \psi(t)\} \gamma_0(t, 1) dt \right].$$

Since $\gamma_0(t, 1)$ is unknown, we use its null Breslow-type estimator $\hat{\Gamma}_0(t, 1)$. Finally, we obtain the score vector

$$\begin{aligned} U(\tau) &= \int_0^\tau \psi(t) \left[d\bar{N}_2(t, 1) - \bar{R}_2(t, 1) \exp\{\theta^\top \psi(t)\} \frac{d\bar{N}_1(t, 1) + d\bar{N}_2(t, 1)}{\bar{R}_1(t, 1) + \bar{R}_2(t, 1)} \right] \\ &= \int_0^\tau L(t) (d\hat{\Gamma}_2(t, 1) - d\hat{\Gamma}_1(t, 1)), \end{aligned}$$

where

$$L(t) = \psi(t) \frac{\bar{R}_1(t, 1) \bar{R}_2(t, 1)}{\bar{R}_1(t, 1) + \bar{R}_2(t, 1)}.$$

This vector resembles a partial likelihood score vector but here the risksets are reweighted. It is a vector of weighted logrank-type statistics for comparing subdistribution hazard functions. When $U(\tau)$ is one-dimensional ($d = 1$) and $\varphi_1(t) = 1$, it agrees with the statistic of Gray (1988).

In practice, the time transformation in $\psi(t) = \varphi(F_0(t, 1)/F_0(\tau, 1))$ must be estimated, i.e., $\hat{F}_0(\cdot, 1)$ replaces $F_0(\cdot, 1)$.

In Theorem 1 in Appendix, I show that under the null hypothesis the score vector $n^{-1/2}U(\tau)$ (where $n = n_1 + n_2$) is asymptotically normal with mean zero and variance matrix which is consistently estimated by $n^{-1}\hat{\sigma}(\tau, \tau)$ given in that theorem. Consequently, the quadratic score statistic $T = U(\tau)^\top \hat{\sigma}(\tau, \tau)^{-1}U(\tau)$ is asymptotically χ^2 distributed with d degrees of freedom. Significantly large values of T contradict the hypothesis.

Theorem 2 provides a condition for consistency of the test. Unlike the Kolmogorov–Smirnov test of Lin (1997), this test is not consistent against an arbitrary alternative. Alternatives that will be rejected with probability converging to 1 are given by the choice of the basis functions. The consistency condition essentially says that the test is consistent unless the basis functions are orthogonal to the true distribution in certain sense. If we take three or four basis functions, the true difference between subdistributions would have to be quite unusually complicated for the test to be inconsistent. For instance, Legendre polynomials of order 0, 1, 2 will be able to detect proportional subdistribution hazards as well as monotone and nonmonotone (convex or concave) subdistribution hazard log-ratios.

3. SIMULATIONS

I conducted a simulation study to investigate properties of the proposed test and compare them with other existing tests both under the null hypothesis and under alternatives. Datasets of size 100 (50 in each sample) are generated. The number of Monte Carlo runs for each model is 20 000 under the hypothesis and 5000 under alternatives. The data generating procedure in the j th sample is as follows: first, the failure type is set to k with probability $p_k = F_j(\infty, k)$, $k \in \{1, 2\}$, then the failure time is drawn from the conditional distribution $F_j(t, k)/p_k$, and, finally, the observation is possibly censored. Censoring times are generated from the uniform distribution on $[0, c]$ (values of c are reported below).

Neyman-type tests proposed in this paper are performed with $d = 3$ Legendre polynomials (of order 0, 1, 2). Lin’s Kolmogorov–Smirnov-type test uses 1000 resampled test processes (see Lin (1997) for the description of the simulation procedure), with pooled sample null estimates of $F_0(t, 1)$ which was found by Bajorunaite & Klein (2007) to give a more accurate approximation than with individual samples estimators. In Pepe’s integral test I use the asymptotic normal approximation with the martingale-based variance estimator derived by Bajorunaite & Klein (2007). Lebesgue integrals involved in this statistic are computed from 0 to $\tau = c$.

In the first set of simulations I investigate the behaviour of tests under H_0 . Cumulative incidence functions take the form $F_0(t, 1) = p_1(1 - e^{-t})$, $F_j(t, 2) = (1 - p_1)(1 - e^{-t})$. Probability p_1 of failure type 1 is 0.25, 0.5 and 0.75. The parameter of the censoring distribution is $c = 7$ (about 15 % censored in all of the situations) and $c = 2.5$ (about 37 %). Rejection probabilities

TABLE 1. Empirical levels on the nominal level of 5%. Figures based on 20 000 Monte Carlo repetitions (standard deviation 0.0015).

	$c = 7$ (15% censored)			$c = 2.5$ (37% censored)		
	p_1			p_1		
	0.25	0.5	0.75	0.25	0.5	0.75
Neyman	0.0608	0.0579	0.0450	0.0562	0.0682	0.0602
KS	0.0300	0.0160	0.0196	0.0491	0.0495	0.0557
Pepe	0.0333	0.0205	0.0172	0.0468	0.0464	0.0476
Gray	0.0582	0.0590	0.0623	0.0550	0.0562	0.0572

on the nominal level 5% are reported in Table 1. The accuracy of the level of the smooth test and Gray's logrank-type test appears acceptable. The Kolmogorov-Smirnov-type test and integral tests tend to be slightly conservative when the censoring rate is low (see Bajorunaite & Klein (2007) for a detailed analysis).

Next I consider five alternative configurations. Figure 2 shows subdistribution characteristics for type 1 events.

Configuration A.

$$\begin{aligned} F_1(t, 1) &= 0.5(1 - e^{-t}), & F_2(t, 1) &= 1 - (1 - F_1(t, 1))^2, \\ F_1(t, 2) &= 0.5(1 - e^{-t}), & F_2(t, 2) &= 0.25(1 - e^{-t}). \end{aligned}$$

This situation was considered by Gray (1988). Subdistribution hazard rates for events of type 1 are proportional. With $c = 4$ there is 23% censored observations.

Configuration B.

$$\begin{aligned} F_1(t, 1) &= \frac{\pi(1 - e^{-t})}{1 - \pi + \pi(1 - e^{-t})}, & F_2(t, 1) &= \frac{\pi\theta(1 - e^{-t})}{1 - \pi + \pi\theta(1 - e^{-t})}, \\ F_1(t, 2) &= (1 - \pi)(1 - e^{-t}), & F_2(t, 2) &= (1 - \pi)(1 - e^{-t})/(1 - \pi + \pi\theta). \end{aligned}$$

With $\pi = 0.5$, $\theta = e^{0.75}$, this is Model 1 of Bajorunaite & Klein (2007). Set $c = 3$ (24% censoring).

Configuration C.

$$F_j(t, 1) = p_{j1}(1 - e^{-t/p_{j1}}), \quad F_j(t, 2) = (1 - p_{j1})(1 - e^{-t/p_{j1}}).$$

In this situation considered by Bajorunaite & Klein (2007, Model 3) cause-specific intensities of events of type 1 are the same (equal to 1) in both samples. I take $(p_{11}, p_{21}) = (0.3, 0.7)$ and $c = 2$ (giving 28% censoring).

Configuration D.

$$\begin{aligned} F_1(t, 1) &= \frac{2}{3}(1 - e^{-t}), & F_2(t, 1) &= \frac{2}{3}(1 - e^{-t^{0.5}}), \\ F_1(t, 2) &= \frac{1}{3}(1 - e^{-0.8t}), & F_2(t, 2) &= \frac{1}{3}(1 - e^{-1.2t}). \end{aligned}$$

Peng & Fine (2007) used this situation in which both cumulative incidence curves and subdistribution hazards cross. For $c = 4$, 26% is censored.

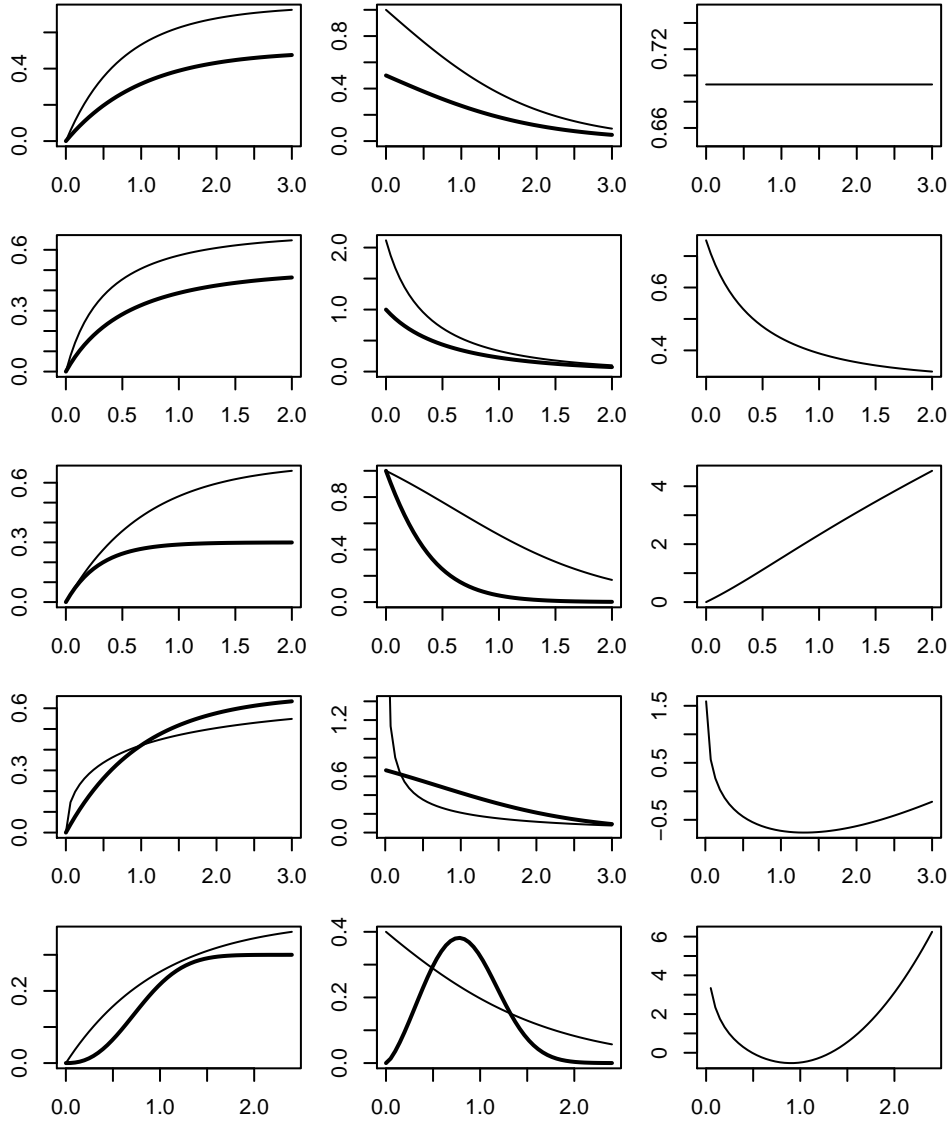


FIGURE 2. Alternative configurations A–E (from top to bottom). On each row: Left plot: cumulative incidence functions $F_1(t, 1)$ (thick line) and $F_2(t, 1)$ (thin line). Middle plot: corresponding subdistribution hazards. Right plot: logarithm of subdistribution hazard ratios.

Configuration E.

$$\begin{aligned}
 F_1(t, 1) &= 0.3(1 - e^{-1.3t^{2.4}}), & F_2(t, 1) &= 0.4(1 - e^{-t}), \\
 F_1(t, 2) &= 0.7(1 - e^{-t}), & F_2(t, 2) &= 0.6(1 - e^{-t}).
 \end{aligned}$$

In this model subdistribution hazards cross but cumulative incidence functions are ordered. The censoring proportion is 25% with $c = 4$.

TABLE 2. Estimated powers on the nominal level 5%. Based on 5000 Monte Carlo repetitions (standard deviation 0.007).

	A	B	C	D	E
Neyman	0.507	0.334	0.672	0.531	0.521
KS	0.344	0.338	0.477	0.130	0.199
Pepe	0.593	0.417	0.581	0.053	0.161
Gray	0.690	0.505	0.550	0.058	0.235

Empirical powers are summarised in Table 2. In Configurations A and B, Gray’s logrank-type test performs best which is not surprising as the subdistribution hazards differ largely (at least at the beginning) and do not cross. The Neyman-type smooth test and Pepe’s integral test do not lose much. In Configuration C, the difference of the subdistributions appears later in time which makes the smooth test slightly more powerful but the difference is not dramatic. In Configuration D, the logrank-type test and the integral test fail because functions they are based on cross. Similarly, in the last Configuration E, Neyman’s test outperforms the other methods (here the bad performance of Pepe’s test is rather unexpected as the subdistribution functions are ordered). Interestingly, Lin’s Kolmogorov–Smirnov-type test has also low power for D and E despite its ‘omnibus’ property (consistency against any alternative).

The proposed smooth tests appear to have stable power over a wide range of realistic alternatives and thus seem to be virtually ‘omnibus’. They did not ‘completely fail’ in any of practically relevant situations considered in this simulation. Smooth test procedures can be recommended as an alternative to the supremum-type procedure for their better performance in complicated situations. Moreover, in simple situations, they do not lose much compared to other existing methods.

4. REAL EXAMPLE

In the bone marrow transplant study introduced in Section 1 there were two competing risks (relapse and death in remission) and two groups of patients (with an HLA-identical sibling donor and with an HLA-matched unrelated donor).

First consider the risk of relapse. Gray’s logrank-type test does not lead to rejection of the hypothesis of equal relapse cumulative incidence functions for the two kinds of donors: the test statistic is -1.66 with p -value 0.098. Pepe’s test statistic is -2.09 with $p = 0.036$, Lin’s Kolmogorov–Smirnov test statistic equals 0.0672 with p -value 0.027 (based on 5000 simulated processes). In contrast to these marginally significant results, the Neyman-type smooth test with $d = 3$ basis functions strongly rejects the hypothesis with the test statistic 14.1, $p = 0.0028$. This conclusion agrees with the visual impression drawn from Figure 1.

Cumulative incidence functions for the risk of death in remission are ordered and well separated. All of the tests discussed here reject with $p < 0.0001$.

ACKNOWLEDGEMENTS

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).

APPENDIX: ASYMPTOTIC RESULTS

The following results are derived under the assumption that $n^{-1}\bar{Y}_1, n^{-1}\bar{Y}_2$ converge uniformly on $[0, \tau]$ to some functions \bar{y}_1, \bar{y}_2 , respectively, which are bounded away from zero. By the Glivenko–Cantelli theorem, this basically holds with $\bar{y}_j(t) = a_j S_j(t)(1 - G_j(t))$, where $G_j(t)$ is the distribution function of censoring times in the j th group and $a_j = \lim_{n \rightarrow \infty} n_j/n$, provided $S_j(\tau) > 0$, $1 - G_j(\tau) > 0$ and $a_j \in (0, 1)$. Denote by $\bar{r}_j(t, 1) = \bar{y}_j(t)(1 - F_j(t, 1))/S_j(t)$ uniform limits in probability of $n^{-1}\bar{R}_j(t, 1)$ (the limits exist by uniform consistency of $\hat{S}_j(t)$ and $\hat{F}_j(t, 1)$).

Theorem 1 (Asymptotic distribution). *The score vector $n^{-1/2}U(\tau)$ is under the null hypothesis $F_1(\cdot, 1) = F_2(\cdot, 1)$ asymptotically (as $n \rightarrow \infty$) distributed as a zero-mean Gaussian vector with covariance matrix whose consistent estimator is $n^{-1}\hat{\sigma}(\tau, \tau) = n^{-1}(\hat{\sigma}_1(\tau, \tau) + \hat{\sigma}_2(\tau, \tau))$ with $\hat{\sigma}_j(s, t)$ given by (3) below.*

Proof. Under the null hypothesis $\Gamma_1(t, 1) = \Gamma_2(t, 1)$ the score process may be expressed as $U(t) = U_2(t) - U_1(t)$ with

$$U_j(t) = \int_0^t L(s)(d\hat{\Gamma}_j(s, 1) - d\Gamma_j(s, 1)) = \int_0^t L(s) \left(\frac{d\hat{F}_j(s, 1)}{1 - \hat{F}_j(s-, 1)} - \frac{dF_j(s, 1)}{1 - F_j(s-, 1)} \right).$$

Hence the process U_j is a function of $\hat{F}_j(\cdot, 1)$, and thus the asymptotic distribution of $n^{-1/2}U_j$ can be inferred from that of $n^{1/2}(\hat{F}_j(\cdot, 1) - F_j(\cdot, 1))$ by a use of the functional delta method (see Section II.8 of Andersen et al. (1993), or Chapter 3.9 of van der Vaart & Wellner (1996)).

Asymptotic results for $n^{1/2}(\hat{F}_j(\cdot, 1) - F_j(\cdot, 1))$ were derived by Lin (1997). He found a martingale representation in the form $n^{1/2}(\hat{F}_j(\cdot, 1) - F_j(\cdot, 1)) = n^{1/2}V_j + o_P(1)$, where

$$\begin{aligned} n^{1/2}V_j(t) = n^{1/2} \int_0^t \frac{1 - F_j(s, 2)}{\bar{Y}_j(s)} d\bar{M}_j(s, 1) + n^{1/2} \int_0^t \frac{F_j(s, 1)}{\bar{Y}_j(s)} d\bar{M}_j(s, 2) \\ - n^{1/2}F_j(t, 1) \int_0^t \frac{d\bar{M}_j(s, 1) + d\bar{M}_j(s, 2)}{\bar{Y}_j(s)} \end{aligned}$$

(beware of misprints in eq. (2) in Lin’s paper) with counting process martingales $\bar{M}_j(t, k) = \bar{N}_j(t, k) - \int_0^t \bar{Y}_j(s)\alpha_j(s, k)ds$. By the martingale central limit theorem (Andersen et al., 1993, Theorem II.5.1) this process converges weakly to a zero-mean continuous Gaussian process with covariance function consistently estimated by $n\hat{\rho}(s, t)$ with

$$\begin{aligned} \hat{\rho}(s, t) = \int_0^{s \wedge t} \frac{(1 - \hat{F}_j(u, 2))^2}{\bar{Y}_j(u)^2} d\bar{N}_j(u, 1) + \int_0^{s \wedge t} \frac{\hat{F}_j(u, 1)^2}{\bar{Y}_j(u)^2} d\bar{N}_j(u, 2) \\ + \hat{F}_j(s, 1)\hat{F}_j(t, 1) \int_0^{s \wedge t} \frac{d\bar{N}_j(u, 1) + d\bar{N}_j(u, 2)}{\bar{Y}_j(u)^2} \end{aligned}$$

$$- (\hat{F}_j(s, 1) + \hat{F}_j(t, 1)) \left(\int_0^{s \wedge t} \frac{1 - \hat{F}_j(u, 2)}{\bar{Y}_j(u)^2} d\bar{N}_j(u, 1) + \int_0^{s \wedge t} \frac{\hat{F}_j(u, 1)}{\bar{Y}_j(u)^2} d\bar{N}_j(u, 2) \right).$$

The delta method (together with the chain rule, and Proposition II.8.6 of Andersen et al. (1993) or Lemma 3.9.17 of van der Vaart & Wellner (1996)) yields that $n^{-1/2}U_j(\cdot)$ is asymptotically equivalent to

$$\int_0^\cdot n^{-1}L(s) \frac{1}{1 - F_j(s, 1)} n^{1/2} dV_j(s) + \int_0^\cdot n^{-1}L(s) \frac{n^{1/2}V_j(s)}{(1 - F_j(s, 1))^2} dF_j(s, 1).$$

By integration by parts this equals

$$\int_0^\cdot n^{-1}Q_j(s) n^{1/2} dV_j(s) + n^{-1}H_j^L(\cdot) n^{1/2} V_j(\cdot),$$

where

$$Q_j(t) = \frac{L(t)}{1 - F_j(t, 1)} - H_j^L(t),$$

$$H_j^L(t) = \int_0^t L(s) H_j(s, 1), \quad H_j(t, 1) = \int_0^t \frac{dF_j(s, 1)}{(1 - F_j(s, 1))^2}$$

($H_j(t, 1)$ is equal to $F_j(t, 1)/(1 - F_j(t, 1))$ and can be viewed as a subdistribution odds function). Using the fact that $n^{-1}L$ uniformly converges in probability (by the assumption of the theorem and by consistency of estimators involved in L) we obtain that $n^{-1/2}U_j$ is asymptotically a zero-mean continuous Gaussian process. Its limiting covariance matrix function can be consistently estimated by

$$n^{-1}\hat{\sigma}_j(s, t) = n^{-1} \int_0^s \int_0^t \hat{Q}_j(u) \hat{Q}_j(v) \hat{\rho}_j(du, dv) + n^{-1} \int_0^s \hat{Q}_j(u) \hat{\rho}_j(du, t) \hat{H}_j^L(t)^\top$$

$$+ n^{-1} \hat{H}_j^L(s) \int_0^t \hat{Q}_j(v)^\top \hat{\rho}_j(s, dv) + n^{-1} \hat{H}_j^L(s) \hat{\rho}_j(s, t) \hat{H}_j^L(t)^\top \quad (3)$$

(\hat{Q}_j and \hat{H}_j^L are defined like Q_j and H_j^L with $\hat{F}_j(\cdot, 1)$ in place of $F_j(\cdot, 1)$).

The weak convergence result achieved above holds jointly for $n^{-1/2}(U_1, U_2)$ with U_1, U_2 asymptotically independent (by independence of the two samples). Finally, the score vector $n^{-1/2}U(\tau) = n^{-1/2}U_2(\tau) - n^{-1/2}U_1(\tau)$ converges in distribution to a zero-mean normal vector with variance matrix which is consistently estimated $n^{-1}(\hat{\sigma}_1(\tau, \tau) + \hat{\sigma}_2(\tau, \tau))$ \square

Theorem 2 (Consistency). *Assume that $\gamma_1(t, 1) \neq \gamma_2(t, 1)$ on a non-null set. Denote by $F_0^*(\cdot, 1)$ the limit in probability of $\hat{F}_0(\cdot, 1)$ under this alternative, and set $\psi^*(t) = \varphi(F_0^*(t, 1)/F_0^*(\tau, 1))$. Then the test is consistent provided the condition*

$$\int_0^\tau \psi^*(t) \frac{\bar{r}_1(t, 1) \bar{r}_2(t, 1)}{\bar{r}_1(t, 1) + \bar{r}_2(t, 1)} (\gamma_2(t, 1) - \gamma_1(t, 1)) dt \neq 0 \quad (4)$$

(at least one component is nonzero) is satisfied.

Proof. We have

$$n^{-1}U(\tau) = n^{-1} \int_0^\tau L(t) (d\hat{\Gamma}_2(t, 1) - d\Gamma_2(t, 1)) - n^{-1} \int_0^\tau L(t) (d\hat{\Gamma}_1(t, 1) - d\Gamma_1(t, 1))$$

$$+ n^{-1} \int_0^T L(t)(d\Gamma_2(t, 1) - d\Gamma_1(t, 1)).$$

The first two terms on the right-hand side converge in probability to zero by the previous proof, and the last term converges in probability to the left-hand side of (4). Therefore, $n^{-1}T$ converges in probability to a nonzero quantity, and the test rejecting for large T is consistent. \square

REFERENCES

- ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. & KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.
- BAJORUNAITE, R. & KLEIN, J. P. (2007). Two-sample tests of the equality of two cumulative incidence functions. *Comput. Statist. Data Anal.* 51 4269–4281.
- GRAY, R. J. (1988). A class of k -sample tests for comparing the cumulative incidence of a competing risk. *Ann. Statist.* 16 1141–1154.
- KRAUS, D. (2007). Data-driven smooth tests of the proportional hazards assumption. *Life-time Data Anal.* 13 1–16.
- LIN, D. Y. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Stat. Med.* 16 901–910.
- PEÑA, E. A. (1998). Smooth goodness-of-fit tests for composite hypothesis in hazard based models. *Ann. Statist.* 26 1935–1971.
- PENG, L. & FINE, J. P. (2007). Nonparametric quantile inference with competing-risks data. *Biometrika* 94 735–744.
- PEPE, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies. *J. Amer. Statist. Assoc.* 86 770–778.
- RAYNER, J. C. W. & BEST, D. J. (1989). *Smooth Tests of Goodness of Fit*. Oxford University Press, New York.
- VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer, New York.