Statistica Sinica **9**(1999), 985-997

COMPARING TWO TREATMENTS WITH MULTIPLE COMPETING RISKS ENDPOINTS

Xiaolong Luo and Bruce W. Turnbull

Bristol-Myers Squibb Co. and Cornell University

Abstract: The problem of testing for differences between two treatments with respect to multiple competing risk time-to-event endpoints is considered. Using a nonparametric approach, a multivariate two-sample statistic is proposed for simultaneous testing for differences in cumulative incidence of the multiple competing survival endpoints.

Establishment of the statistical properties of the test requires consideration of covariation processes among martingales associated with all cause-specific counting processes. The procedure is illustrated by a simulation study and by an application to data from a recent large randomized cancer prevention clinical trial.

Key words and phrases: Cause-specific cumulative incidence, cause-specific failure probability, cause-specific hazard, clinical trial, counting process, covariation process, martingale, multiple outcomes, nonparametric estimates, simultaneous inference, survival data.

1. Introduction

Design and evaluation of clinical cancer trials often involve several time-to-event endpoints. The regulatory decision regarding the recommendation of a new drug or treatment regimen can be difficult to make when effects of a new drug vary among study endpoints — Huque and Sankoh (1997). For example, use of estrogen replacement therapy in post-menopausal women can lead to benefits of prevention of onset of heart disease, osteoporosis and colon cancer but at a cost of increased risks of breast, endometrial and ovarian cancer, e.g. see Bilezikian (1994), Calle, Miracle-McMahill, Thun and Heath (1995) and Jick (1993).

Various statistical methods have been proposed to assess the overall effect of treatment and aid in the two-decision problem of whether or not to go forward with the new treatment when there are multiple outcome variables. One such method is to consider quality of life adjusted survival (Gelber, Gelman and Goldhirsch (1989)). For sequential monitoring of a clinical trial, Jennison and Turnbull (1993) and Thall, Simon and Estey (1993, 1996) propose specification of *minimum acceptable tradeoffs* for multiple normal and binary responses, respectively. In particular, Jennison and Turnbull (1993, Sec. 2) propose dividing the parameter space for each outcome variable into acceptable, indifference and unacceptable regions. For K outcome variables, this leads to 3^K regions and non-statistical considerations are used to assign an "Accept" or "Reject" decision to each. This will be the paradigm we will use when considering the situation with competing time-to-event outcomes on each subject.

Competing risk models with failure times to several endpoints have been intensively studied, see, for example, Kalbfleisch and Prentice (1980, Chap. 6) and Cox and Oakes (1984, Chap. 9). Some interesting examples from clinical oncology data are presented in detail by Gaynor, Feuer, Tan, Wu, Little, Straus, Clarkson and Brennan (1993). For comparing groups with respect to cumulative incidence of a particular single failure type in the presence of several competing risks, Gray (1988) develops a test statistic based on integrated weighted differences of the hazard rates corresponding to the cumulative incidence functions for the failure type of interest in each group. (The cumulative incidence for a given cause at time t is defined as the probability that failure is due to that cause and that this occurs at or before time t ($t > 0$). It has also been termed the "absolute cause-specific risk" in Benichou and Gail (1990) and the "crude incidence curve" in Korn and Dorey (1992).) An approach similar to that of Gray (1988) was taken by Pepe and Mori (1993) but using weighted differences of the cumulative incidence functions themselves, or alternatively of conditional probability functions. (The latter is defined as the probability of failure due to the given cause by time t conditional that no other cause has led to failure prior to t.) In the context of a regression setup, Cheng, Fine and Wei (1998) have recently considered the problem of predicting the cumulative incidence function under a proportional hazards model.

Pepe (1991) develops a very general class of procedures for comparing groups using estimates of functionals of survival, cumulative incidence and cumulative hazard functions. However Pepe uses a univariate (or "global") summary statistic for the difference in the multiple competing risk experience between the two groups. Our methodological approach is to retain the multivariate structure and base inferences on a multivariate statistic, in which each component corresponds to a respective failure type. Also our analytic approach differs from that of Pepe (1991). We utilize a martingale technique to obtain asymptotic joint distributions of various test statistics for use with simultaneous inference procedures involving multiple survival endpoints in a dependent competing risks setup. Previous authors have considered variation processes of martingales associated with individual cause-specific counting processes. By calculating covariation processes among martingales associated with all cause-specific counting processes, we are able to develop a simultaneous approach for the dependent competing risks problem. This development is described in Section 2. In Section 3 we apply the theory to the construction of a two-decision testing procedure, as described above, with given operating characteristics. In Section 4, we describe the results of a simulation study designed to investigate the finite sample properties of the procedure. Finally we give an application to data from the Nutritional Prevention of Cancer (NPC) trial (Clark (1996)).

2. Theoretical Background

2.1. The one sample problem

We start by describing some large sample results for the one sample problem. Suppose data from a competing risk model are i.i.d. random vectors $\{(Y_i, \delta_i), 1 \leq i \leq n\}$ with distribution equal to a random vector (Y, δ). We assume that the failure time $Y \geq 0$ is a continuous random variable and failure type δ can take on one of the values $1, \ldots, k$. Thus, when $\delta = j$, we say that the subject fails for cause j at time Y. A subject cannot fail from two causes at the same time. We may consider censoring as one of the failure types. For each subject i and failure type j, we denote cause-specific counting processes as

$$N_i^j(t) = 1_{(Y_i \leq t, \delta_i = j)}, \qquad t \geq 0.$$

Denote all information observed up to time t by the filtration

$$\mathcal{F}_t = \sigma\{N_i^j(u), 1 \leq j \leq k, 1 \leq i \leq n; 0 \leq u \leq t\}, \quad t \geq 0. \tag{1}$$

Note that, when $k = 2$, this filtration is consistent with the one defined in Theorem 1.3.1 of Fleming and Harrington (1991, p.26). We define the cause-specific hazard rate as

$$\lambda_j^{\#}(t) = -\frac{\frac{d}{dt}P(Y \geq t, \delta = j)}{P(Y \geq t)}, \quad \text{for} \quad 1 \leq j \leq k.$$

Fleming and Harrington (1991, p.29) show that, for each j, $1 \leq j \leq k$,

$$M_i^j(t) = N_i^j(t) - \int_0^t \lambda_j^{\#}(s) 1_{(Y_i \geq s)} ds, \quad 1 \leq i \leq n, \tag{2}$$

are locally square integrable F_t-martingales with respect to the filtration (1).

To facilitate the description, we need additional notation. Denote the cause-specific cumulative hazard $\Lambda_j^{\#}(t) = \int_0^t \lambda_j^{\#}(s)ds$, marginal survival function $F(t) = P(Y \geq t)$, and cumulative incidence $I_j(t) = P(Y < t, \delta = j) = \int_0^t F(s)\lambda_j^{\#}(s)ds$. We will use corresponding estimators $\hat{\Lambda}_{jn}^{\#}(t) = \int_0^t (1/\bar{Y}_n(s)) d\bar{N}_n^j(s)$, $\hat{F}_n(t) = \exp\{-\sum_{j=1}^k \int_0^t (1/\bar{Y}_n(s)) d\bar{N}_n^j(s)\}$, and $\hat{I}_{jn}(t) = \int_0^t (\hat{F}_n(s)/\bar{Y}_n(s)) d\bar{N}_n^j(s)$, respectively. In these expressions, the sample means are defined as $\bar{N}_n^j(s) = \frac{1}{n}\sum_{i=1}^n N_i^j(s)$,

$1 \leq j \leq k$ and $\bar{Y}_n(s) = \frac{1}{n}\sum_{i=1}^n 1_{(Y_i \geq s)}$. The asymptotic properties of these estimators have been described by several authors, including Kalbfleisch and Prentice (1980) and Pepe (1991).

2.2. The two sample problem

Suppose there are two groups of i.i.d. random vectors $\{(Y_i^{(m)}, \delta_i^{(m)}),\ 1 \leq i \leq n_m\}, m = 1, 2$ with distributions equal to random vectors $(Y^{(m)}, \delta^{(m)}), m = 1, 2$, respectively. In general, the superscript (m) attached to any quantity defined in Section 2.1, such as $\lambda_j^{(m),\#}(s)$, $n^{(m)}$, etc., refers to that same quantity for the corresponding group, $m = 1, 2$. In particular, we let $I_j^{(m)}(t)$ be the cumulative incidence function for group m and cause j and let $\hat{I}_j^{(m)}(t)$ be its estimator as described above. We propose to use the cause-specific cumulative incidence function as the basis to compare the two groups. This permits the comparison of probabilities of failure of a specific type in the setting where the competing risks are acknowledged to exist and thus has a direct clinical interpretation (Pepe and Mori (1993), Sec. 1). This is not the case for the cause-specific hazard function, for example. As Gray (1988, p.1142) points out, a test of equality of cumulative incidence functions for a given failure type is not equivalent to one of equality of corresponding cause-specific hazard functions for that type, unless the null hypothesis also specifies that the overall survival functions are the same in both groups.

From Pepe (1991, p.772), we have that $\{\sqrt{n^{(m)}}(\hat{I}_j^{(m)}(t) - I_j^{(m)}(t)),\ 1 \leq j \leq k\}$ converges in distribution to a k-variate normal distribution for each $m = 1, 2$. We propose to use the integrated weighted difference

$$X_j(\tau) = \sqrt{\frac{n^{(1)}n^{(2)}}{n^{(1)} + n^{(2)}}} \int_0^\tau \hat{W}_j(s) d\left[\hat{I}_j^{(1)}(s) - \hat{I}_j^{(2)}(s)\right]$$

as a test statistic to compare cause-specific cumulative incidences $I_j^{(1)}(t)$ and $I_j^{(2)}(t)$, where $\hat{W}_j(t)$ is a weight function converging to some $W_j(t)$ in probability. This statistic is in the class considered by Pepe (1991, Sec. 4) and related to the one proposed by Gray (1988, Eqn 1.3). However, here we are interested in simultaneous inference on cumulative incidences of all causes. Accordingly, we need the joint distribution of $(X_1(\tau), \ldots, X_k(\tau))$.

Denote weighted cumulative incidences as

$$B_j^{(m)}(t) = \int_0^t W_j(s)F^{(m)}(u)\lambda_j^{(m),\#}(s)ds, \quad 1 \leq j \leq k, \quad m = 1, 2$$

and, for each fixed $\tau > 0$, define

$$A_l^{(m)}(s;j,\tau,n^{(m)}) = \begin{cases} \left[B_j^{(m)}(\tau) - B_j^{(m)}(s)\right] \dfrac{\hat{F}^{(m)}_{n^{(m)}}(s)}{F^{(m)}(s)\bar{Y}^{(m)}_{n^{(m)}}(s)}, & l \neq j \\ \left[B_j^{(m)}(\tau) - B_j^{(m)}(s)\right] \dfrac{\hat{F}^{(m)}_{n^{(m)}}(s)}{F^{(m)}(s)\bar{Y}^{(m)}_{n^{(m)}}(s)} - \hat{W}_j(s)\dfrac{\hat{F}^{(m)}_{n^{(m)}}(s)}{\bar{Y}^{(m)}_{n^{(m)}}(s)}, & l = j. \end{cases}$$

With a standard martingale manipulation argument, we can show that, under a null hypothesis that the cumulative incidence functions for cause j are the same for both groups, i.e., $I_j^{(1)}(t) = I_j^{(2)}(t)$ for $t > 0$, we can write

$$X_j(\tau) = \sqrt{\frac{n^{(1)}n^{(2)}}{n^{(1)}+n^{(2)}}} \left[-\sum_{l=1}^{k} \int_0^{\tau} A_l^{(1)}(s;j,\tau,n^{(1)}) d\bar{M}^{(1),l}_{n^{(1)}}(s) \right.$$
$$\left. + \sum_{l=1}^{k} \int_0^{\tau} A_l^{(2)}(s;j,\tau,n^{(2)}) d\bar{M}^{(2),l}_{n^{(2)}}(s) \right] + o(1). \tag{3}$$

Note that for each fixed τ, $A_l^{(m)}(s;j,\tau,n^{(m)})$ is F_s-measurable. Under standard regularity conditions, variances and covariances of all $\{X_j(\tau)\}$ are convergent and any linear combination of the $\{X_j(\tau)\}$ has an asymptotic normal distribution. Hence by the Cramér-Wold device (e.g. Billingsley (1995), Theorem 29.4), $(X_1(\tau), \ldots, X_k(\tau))$ converges to a multivariate normal distribution. The considerations concerning the choice of weight function W are analogous to those for for weighted Kaplan-Meier or weighted logrank statistics for comparing survival distributions — Pepe and Fleming (1989, 1991). A decreasing function gives more weight to early differences resulting in greater power against corresponding alternative hypotheses. The opposite would apply to increasing weight functions. In our example, we examine sensitivity to the choice of W by considering several weight functions in the classes considered by Gray (1988) and by Pepe and Mori (1993). There has been less guidance in the literature as to the choice of horizon τ. For the asymptotic theory to hold, τ must be fixed and should be prespecified based on historical experience. Typically τ should be chosen so that it is expected that most or all of the events are included in the interval $[0, \tau]$; however, if the weight function is not chosen to be decreasing sufficiently, increased variability of the estimates of the cumulative incidence function at later times can tend to dilute the power of the test. Fortunately, limited sensitivity analyses that we have performed have shown results that are fairly insensitive to the choices of weight function and horizon, at least in the examples we present.

To illustrate an application of the result (3), we consider simultaneous inference for two of k competing risks. The same method can be used for simultaneous inference for any number of competing risks.

2.3. Simultaneous inference for two competing risks

Suppose we are interested in failure causes 1 and 2 out of the k competing causes. From the previous section we have that the joint distribution of $(X_1(\tau), X_2(\tau))$ is asymptotically bivariate normal. From the martingale expression (3), we have variances that are asymptotically equivalent to

$$\begin{aligned}\mathrm{Var}\,(X_1(\tau)) \overset{a}{=} & \frac{n^{(1)}n^{(2)}}{n^{(1)}+n^{(2)}}\sum_{l=1}^{k}\int_0^\tau \Big[\frac{A_l^{(1)}(s;1,\tau,n^{(1)})^2\bar{Y}_{n^{(1)}}(s)\lambda_l^{(1),\#}(s)}{n^{(1)}} \\ & +\frac{A_l^{(2)}(s;1,\tau,n^{(2)})^2\bar{Y}_{n^{(2)}}(s)\lambda_l^{(2),\#}(s)}{n^{(2)}}\Big]ds, \\ \mathrm{Var}\,(X_2(\tau)) \overset{a}{=} & \frac{n^{(1)}n^{(2)}}{n^{(1)}+n^{(2)}}\sum_{l=1}^{k}\int_0^\tau \Big[\frac{A_l^{(1)}(s;2,\tau,n^{(1)})^2\bar{Y}_{n^{(1)}}(s)\lambda_l^{(1),\#}(s)}{n^{(1)}} \\ & +\frac{A_l^{(2)}(s;2,\tau,n^{(2)})^2\bar{Y}_{n^{(2)}}(s)\lambda_l^{(2),\#}(s)}{n^{(2)}}\Big]ds\end{aligned}$$

and the asymptotic covariance

$$\begin{aligned}&\mathrm{Cov}\,(X_1(\tau), X_2(\tau)) \overset{a}{=} \\ &\frac{n^{(1)}n^{(2)}}{n^{(1)}+n^{(2)}}\sum_{l=1}^{k}\int_0^\tau \Big\{A_l^{(1)}(s;1,\tau,n^{(1)})A_l^{(1)}(s;2,\tau,n^{(1)})\frac{\bar{Y}_{n^{(1)}}(s)\lambda_l^{(1),\#}(s)}{n^{(1)}} \\ &\quad +A_l^{(2)}(s;1,\tau,n^{(2)})A_l^{(2)}(s;2,\tau,n^{(2)})\frac{\bar{Y}_{n^{(2)}}(s)\lambda_l^{(2),\#}(s)}{n^{(2)}}\Big\}ds.\end{aligned}$$

Denote

$$\hat{A}_l^{(m)}(s;j,t,n^{(m)}) = \begin{cases}\Big[\hat{B}_j^{(m)}(t)-\hat{B}_j^{(m)}(s)\Big]\frac{1}{\bar{Y}_{n^{(m)}}^{(m)}(s)}, & l \neq j, \\ \Big[\hat{B}_j^{(m)}(t)-\hat{B}_j^{(m)}(s)\Big]\frac{1}{\bar{Y}_{n^{(m)}}^{(m)}(s)} - \hat{W}_j(s)\frac{\hat{F}_{n^{(m)}}^{(m)}(s)}{\bar{Y}_{n^{(m)}}^{(m)}(s)}, & l = j.\end{cases}$$

where

$$\hat{B}_j^{(m)}(u) = \int_0^u \hat{W}_j(s)\frac{\hat{F}_{n^{(m)}}^{(m)}(s)}{\bar{Y}_{n^{(m)}}^{(m)}(s)}d\bar{N}_n^{(m),j}(s).$$

We can estimate the variances by

$$\begin{aligned}\hat{\sigma}_1^2 &= \tfrac{n^{(1)}n^{(2)}}{n^{(1)}+n^{(2)}}\sum_{l=1}^{k}\int_0^\tau \Big[\tfrac{\hat{A}_l^{(1)}(s;1,\tau,n^{(1)})^2}{n^{(1)}}d\bar{N}_{n^{(1)}}^{(1),l}(s)+\tfrac{\hat{A}_l^{(2)}(s;1,\tau,n^{(2)})^2}{n^{(2)}}d\bar{N}_{n^{(2)}}^{(2),l}(s)\Big], \\ \hat{\sigma}_2^2 &= \tfrac{n^{(1)}n^{(2)}}{n^{(1)}+n^{(2)}}\sum_{l=1}^{k}\int_0^\tau \Big[\tfrac{\hat{A}_l^{(1)}(s;2,\tau,n^{(1)})^2}{n^{(1)}}d\bar{N}_{n^{(1)}}^{(1),l}(s)+\tfrac{\hat{A}_l^{(2)}(s;2,\tau,n^{(2)})^2}{n^{(2)}}d\bar{N}_{n^{(2)}}^{(2),l}(s)\Big]\end{aligned}$$

and the covariance by

$$\hat{C} = \frac{n^{(1)}n^{(2)}}{n^{(1)}+n^{(2)}} \sum_{l=1}^{k} \int_0^{\tau} \Big\{ \hat{A}_l^{(1)}(s;1,\tau,n^{(1)})\hat{A}_l^{(1)}(s;2,\tau,n^{(1)}) \frac{d\bar{N}_{n^{(1)}}^{(1),l}(s)}{n^{(1)}}$$
$$+ \hat{A}_l^{(2)}(s;1,\tau,n^{(2)})\hat{A}_l^{(2)}(s;2,\tau,n^{(2)}) \frac{d\bar{N}_{n^{(2)}}^{(2),l}(s)}{n^{(2)}} \Big\}.$$

Denoting

$$T_1 = \frac{X_1(\tau)}{\hat{\sigma}_1}, \quad T_2 = \frac{X_2(\tau)}{\hat{\sigma}_2}, \quad \hat{\rho} = \frac{\hat{C}}{\hat{\sigma}_1\hat{\sigma}_2}, \tag{4}$$

we will use (T_1, T_2) as a bivariate test statistic that has an asymptotic bivariate normal distribution with correlation coefficient consistently estimated by $\hat{\rho}$.

3. A Testing Procedure for a Two-Decision Problem

We use the statistic (T_1, T_2) and, by the above, we may suppose it is approximately bivariate normally distributed with variances $\sigma_1^2 = \sigma_2^2 = 1$, covariance ρ and means (μ_1, μ_2), say. We suppose that the event “failure” is an unfavorable one, sample 1 is from a standard treatment and sample 2 is a new treatment. Since $T_j,\ j = 1, 2$, is an observed measure of the standardized difference in cumulative incidence for cause j between the control and treatment group, positive values of $\mu_j,\ j = 1, 2$, would imply the new treatment is beneficial. A reasonable procedure might be to recommend the new treatment if it is not worse than the control in both cause-specific cumulative incidences and it is better in at least one. That is, we recommend the treatment if $\min\{\mu_1, \mu_2\} \geq 0$ and $\max\{\mu_1, \mu_2\} > 0$. These criteria can easily be adapted to any of the alternate situations described in Section 2 of Jennison and Turnbull (1993). Following that paper, a reasonable shape for an acceptance region for an overall favorable treatment outcome is:

$$C(a, b) = \{(t_1, t_2) : \min\{t_1, t_2\} \geq a, \ \ \max\{t_1, t_2\} \geq b\} \tag{5}$$

for some a, b. The choice of (a, b) can be made so that the procedure satisfies prespecified requirements on the operating characteristic (OC) function. This will depend on non-statistical considerations such as the relative importance (risk/benefit tradeoff) of the different types of failure outcome. In the absence of such input, an arbitrary *ad hoc* rule might be to first choose $a = -0.3551$ so that $T \leq a$ would marginally imply a negative mean (95% power at $\mu = -2$); and then choose a ρ-dependent b so that $\Pr[(T_1, T_2) \in C(a, b)|\mu_1 = \mu_2 = 0] = \alpha$, a given significance level. For example, if $\hat{\rho} = 0.9$ and $\alpha = 0.05$, we might choose $b = 1.798$ so that $P((T_1, T_2) \in C(a, b)) = 0.05$, when $\mu_1 = \mu_2 = 0$. We employ this procedure in the simulation study used to illustrate the methodology in Section 4.

Alternatively, one may avoid the direct choice of a and b by simply stating a P-value. The simultaneous bivariate P-value for the two-decision problem is given by the probability of the event given in (5) with a and b set equal to the observed values of $\min(T_1, T_2)$ and $\max(T_1, T_2)$, respectively, and assuming a bivariate normal distribution with zero means, unit variances and correlation given by (4). This idea is easily generalized to $k \geq 2$ competing risks. The resulting P-value can be viewed as a univariate summary statistic of the difference in the multiple competing risk experience between the two groups. We now apply this procedure in two examples.

4. Example 1: Simulated Data Sets

We describe a simulation study carried out for the purpose of demonstrating the practicality of the methodology of Section 5 and to illustrate its finite sample properties. One thousand data sets were generated for sample sizes and parameter values that might be typical of a large disease prevention trial, such as the one mentioned in Section 1 and analyzed further in the next section.

First, a binary group assignment variable X is generated which has a Bernoulli distribution $B(1, \frac{1}{2})$; $X = 0$ implies assignment to the standard or control (group $m = 1$), $X = 1$ implies assignment to the new treatment (group $m = 2$). Subjects are equally likely to be assigned to either group, independent of other subjects. Conditional on X, three latent failure times, (Z_1, Z_2, Z_3), are generated from independent exponential distributions with constant hazard rates $p_1 = (2 - d_1 X)/6$, $p_2 = (2 - d_2 X)/6$, and $p_3 = 1 - p_1 - p_2$, respectively. Here, the parameters d_1 and d_2 (with $-4 \leq d_1 \leq 2,\ -4 \leq d_2 \leq 2,\ d_1 + d_2 \geq -4$) specify the amount by which the cumulative incidence is decreased (for $d_j > 0$) or increased ($d_j < 0$) by the new treatment over the standard for failures of type $j = 1$ and $j = 2$, respectively. Finally the failure time Y and failure type are defined as $Y = \min\{Z_1, Z_2, Z_3\}$ and $\delta = j^*$, if $Y = Z_{j^*}$. Note that this is equivalent to generating a failure time Y from an exponential distribution with mean 1 and generating the failure type indicator, independently of Y, from a trinomial distribution with $P(\delta = j) = p_j$, $j = 1, 2, 3$. A random sample of $n = 1000$ group assignments, failure times and failure types were generated. We simulated 1000 such data sets for each of the three parameter settings: $(d_1, d_2) = (0, 0), (1, 0)$, and $(1, 1)$.

To assess the sensitivity of the test statistics, we explored three choices of horizon, namely $\tau = 2, 3, 4$, and two classes of weight function. The first class, suggested by Gray (1988, Eqn 3.2), takes the form:

$$W_j^G(t) = (1 - \hat{I}_{jn}^{(0)}(t))^r, \quad j = 1, 2,$$

using the notation of Section 3, and $\hat{I}^{(0)}_{jn}(t)$ represents the estimated cumulative incidence function for cause j when the data from both groups 1 and 2 are combined. The second class of weight functions considered generalizes that of Pepe and Mori (1993, Sec. 3):

$$W_1(t) = W_2(t) = \left[\frac{\hat{C}_0(t)\hat{C}_1(t)}{(n^{(0)}\hat{C}_0(t) + n^{(1)}\hat{C}_1(t))/(n^{(0)} + n^{(1)})}\right]^r,$$

where

$$\hat{C}_m(t) = \exp\{-\hat{\Lambda}^{(m),\#}_{3n}(t)\}, \quad m = 1, 2,$$

is the estimated survivor function for failures *not* of types 1 or 2 (e.g. includes censoring). Both classes of weight function are indexed by a parameter r. When $r = 0$, both weight functions reduce simply to a constant. When $r > 0$, less weight is put on differences at later time points where there is more variability in the incidence estimates due to smaller numbers at risk. Conversely, a choice of $r < 0$ will lead to more sensitive test when differences at later time points may be more important. For their respective forms of weight function, Gray (1988) proposed choices of $r = 1, 0, -1$, whereas Pepe and Mori (1993) used $r = 1$. We will use the Gray-type weight functions with $r = 1, 0, -1$, and the Pepe/Mori-type weight functions with $r = 0, \pm\frac{1}{2}, \pm 1$.

Table 1. Proportions[1] of 1000 simulated data sets which recommend the new treatment.

	Horizon	Weight function						
Difference		Constant	"Gray"		"Pepe/Mori"			
parameters	τ	$r = 0$	$r = -1$	$r = 1$	$r = -1$	$r = -\frac{1}{2}$	$r = \frac{1}{2}$	$r = 1$
Null	2	0.048	0.070	0.026	0.045	0.047	0.047	0.054
$d_1 = d_2 = 0$	3	0.049	0.076	0.027	0.042	0.045	0.049	0.054
	4	0.048	0.077	0.026	0.058	0.050	0.054	0.053
	2	0.624	0.606	0.653	0.609	0.617	0.637	0.636
$d_1 = 1, d_2 = 0$	3	0.638	0.601	0.659	0.620	0.627	0.628	0.636
	4	0.628	0.607	0.665	0.636	0.638	0.627	0.635

[1] Standard errors for rejection probability estimates around 0.05 are approximately 0.007; for rejection probabilities around 0.60, they are approximately 0.015.

For each of the thousand data sets generated, and for each combination of weight function and horizon τ, the statistics $(T_1, T_2, \hat{\rho})$ were computed. First, the approximate normality (with unit variance) of the marginal distributions of T_1 and of T_2 was examined by construction of smoothed histograms and probability plots using the thousand values generated for each combination of parameter settings, weight function and time horizon. These plots, not shown here, all confirmed the adequacy of the normal approximation. This finding was supported by the non-significance in every case of the corresponding Shapiro-Wilk (1965) test. Next, Table 1 shows the proportions of the 1000 data sets in which the standard treatment was rejected in favor of the new one using the procedure of Section 5 with size $\alpha = 0.05$, for differences $d_1 = d_2 = 0$ and $d_1 = 1, d_2 = 0$. When $d_1 = d_2 = 1$, the null hypothesis was rejected in all 1000 cases (rejection rate = 100%). For the particular situation shown in Table 1, the test based on Gray-type weights with $r = 1$ appears to do best, having lower empirical sizes and slightly higher empirical powers. However, the differences may possibly be due to sampling error. The tests using constant and Pepe/Mori-type weight functions maintain close to nominal size and exhibit good power.

5. Example 2: Application to NPC Trial Data

We now apply the procedure to data from the Nutritional Prevention of Cancer (NPC) trial (Clark *et al.* (1996)), as mentioned in Section 1. In this trial, starting in 1983, 1312 subjects were randomized to either a daily supplement of 200 mg of selenium or a placebo. The patients were followed for up to ten years. The primary endpoint was squamous cell carcinoma of the skin, but incidence of internal cancer (lung, colorectal, prostate etc.) was also of interest as a designated secondary endpoint. Daily dietary supplementation of selenium might be recommended to the general population if it could be demonstrated that it delayed the onset of one (or both) endpoints compared to placebo, provided that the other endpoint was not adversely affected. The two endpoints of interest are (a) internal cancer incidence and (b) incidence of squamous cell carcinoma of the skin. Failure time is measured in months from date of entry into the study to date of first diagnosis. The data are summarized in Table 2 below.

Table 2. Summary data for NPC Trial.

Group	1: Placebo	2: Se Supplement
Patients entered	659	653
(a) Cancer incidence	52	41
(b) SCC incidence	184	211
(c) Incidence Free	423	401
Mean followup (months at risk)	53.4 (Range: 0 - 124.3)	

These frequencies are lower than those reported in Clark *et al.* (1996, Tables 2 and 3) because here we use only competing risk data of time up to the first diagnosis of either endpoint (a) or (b) for each subject.

Initially we used a test with a constant weight function and a horizon $\tau = 60$ months, approximately equal to the median followup. Computing the statistics

(4), we obtain standardized weighted differences between placebo and Se groups $T_1 = 1.93$ for cumulative cancer incidence, $T_2 = -1.45$ for cumulative SCC incidence, and correlation coefficient $\hat{\rho} = -0.13$. These findings agree qualitatively with the results of the standard logrank tests used in Clark *et al.* (1996) . There they reported a statistically significant benefit from Se supplementation for endpoint (a), and a moderate but non-significant increased risk in the less important SCC incidence – endpoint (b). As remarked in Section 5, the simultaneous bivariate P-value for the two-decision problem is given by the probability of the event given in (5) with a and b set equal to the observed values of $\min(T_1, T_2) = -1.45$ and $\max(T_1, T_2) = 1.93$, respectively, and assuming a bivariate normal distribution with zero means, unit variances and correlation $\hat{\rho} = -0.13$. This is computed to be 0.047 and suggests that there is no significant negative treatment effect in both cancer and SCC incidence but there may be significant positive treatment effect in at least one of the endpoints, here endpoint (a), the internal cancers.

Table 3. Sensitivity of P-values for NPC trial data to various horizons and weight functions.

Horizon	Weight function						
τ	Constant	"Gray"		"Pepe/Mori"			
(months)	$r = 0$	$r = -1$	$r = 1$	$r = -1$	$r = -\frac{1}{2}$	$r = \frac{1}{2}$	$r = 1$
50	0.062	0.057	0.066	0.064	0.064	0.063	0.063
60	0.047	0.043	0.050	0.046	0.046	0.046	0.047
70	0.097	0.092	0.100	0.130	0.120	0.111	0.108
80	0.211	0.210	0.210	0.338	0.303	0.269	0.260

It is important to check the sensitivity of these findings to alternate choices of horizon τ (50,60,70,80 months, for example), and to alternate choices of weight function. Such sensitivity analyses should be viewed as exploratory or, alternatively, as providing more credibility to the confirmatory analysis with the prespecified choice of weight function and τ. The results are displayed in Table 3. The longer horizons include later periods where the cumulative incidence function estimates have higher variability thus diluting power, which is reflected in the table by higher P-values. For a given τ, the P-value is relatively stable here for different choices of weight function, except at the highest value of $\tau = 80$. We might argue on statistical grounds that it is preferable to use a decreasing weight function that downweights later time periods because heavy censoring increases the variability and dilutes power. On the other hand, it might be argued on medical grounds that, in a prevention trial, it is unlikely that an intervention will have an immediate effect (see Luo, Turnbull and Clark (1997)), and so early differences in tumor rates should be discounted, thus calling for an increasing weight function. We compromised on a constant weight function.

Acknowledgements

The authors are extremely grateful to Larry C. Clark for providing the data used in Example 2 of Section 5. They are also grateful to a Referee for a careful reading and many helpful comments. The research of the second author was supported in part by a grant from the U.S. National Institutes of Health.

References

Benichou, J. and Gail, M. H. (1990). Estimates of absolute cause-specific risk in cohort studies. *Biometrics* **46**, 813-826.

Bilezikian, J. P. (1994). Major issues regarding estrogen replacement therapy in postmenopausal women. *J. Women's Health* **3**, 273-282.

Billingsley, P. (1995). *Probability and Measure.* 3rd edition. Wiley, New York.

Calle, E. E., Miracle-McMahill, H. L., Thun, M. J. and Heath, C. W. (1995). Estrogen replacement therapy and risk of fatal colon cancer in a prospective cohort of postmenopausal women. *J. Nat. Cancer Inst.* **87**, 517-23.

Cheng, S. C., Fine, J. P. and Wei, L. J. (1998). Prediction of cumulative incidence function under the proportional hazards model. *Biometrics* **54**, 219-228.

Clark, L. C., Combs, G. F., Turnbull, B. W., Slate, E. H., Chalker, D. K., Chow, J., Davis, L. S., Glover, R. A., Graham, G. F., Gross, E. G., Krongrad, A., Lesher, J. L., Park, H. K., Sanders, B. B., Smith, C. L., Taylor, J. R. and the Nutritional Prevention of Cancer Study Group. (1996). Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin: A randomized clinical trial. *J. Amer. Med. Assoc.* **276**, 1957-1963.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, New York.

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

Gaynor, J. J., Feuer, E. J., Tan, C. C., Wu, D. H., Little, C. R., Straus, D. J., Clarkson, B. D. and Brennan, M. F. (1993). On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *J. Amer. Statist. Assoc.* **88**, 400-409.

Gelber, R. D., Gelman, R. S. and Goldhirsch, A. (1989). A quality of life oriented endpoint for comparing therapies. *Biometrics* **45**, 781-795.

Gray, R. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann. Statist.* **16**, 1141-1154.

Huque, M. F. and Sankoh, A. J. (1997). A reviewer's perspective on multiple endpoint issues in clinical trials. *J. Biopharm. Statist.* **7**, 545-564.

Jennison, C. and Turnbull, B. W. (1993). Group sequential tests for bivariate responses: interim analysis of clinical trials with both efficacy and safety endpoints. *Biometrics* **49**, 741-752.

Jick, S. S. (1993). Combined estrogen and progesterone use and endometrial cancer. *Epidemiology* **4**, 384.

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

Korn, E. L. and Dorey, F. J. (1992). Applications of crude incidence curves. *Statistics in Medicine* **11**, 813-829.

Luo, X., Turnbull, B. W. and Clark, L. C. (1997). Likelihood ratio tests for a changepoint with survival data. *Biometrika* **84**, 555-565.

Pepe, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies. *J. Amer. Statist. Assoc.* **86**, 770-778.

Pepe, M. S. and Fleming, T. R. (1989). Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics* **45**, 497-507.

Pepe, M. S. and Fleming, T. R. (1991). Weighted Kaplan-Meier statistics: large sample and optimality considerations. *J. R. Statist. Soc. B* **53**, 341-352.

Pepe, M. S. and Mori, M. (1993). Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine* **12**, 737-751.

Shapiro, S. S. and Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591-611.

Thall, P. F., Simon, R. M. and Estey, E. H. (1993). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* **14**, 357-379.

Thall, P. F., Simon, R. M. and Estey, E. H. (1996). New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *J. Clin. Oncology* **14**, 296-303.

Department 703/91, Bristol-Myers Squibb, P.O. Box 5100, 5 Research Parkway, Wallingford, CT 06492-7660, U.S.A.

E-mail: Xiaolong_Luo@ccmail.bms.com

School of Operations Research and Industrial Engineering, 227 Rhodes Hall, Cornell University, Ithaca, NY 14853-3801, U.S.A.

E-mail: turnbull@orie.cornell.edu

(Received June 1997; accepted January 1999)