A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk

Robert J. Gray

The Annals of Statistics, Vol. 16, No. 3. (Sep., 1988), pp. 1141-1154.

The Annals of Statistics
1988, Vol. 16, No. 3, 1141–1154

A CLASS OF K-SAMPLE TESTS FOR COMPARING THE CUMULATIVE INCIDENCE OF A COMPETING RISK[1]

BY ROBERT J. GRAY

Harvard School of Public Health and Dana-Farber Cancer Institute

In this paper, for right censored competing risks data, a class of tests developed for comparing the cumulative incidence of a particular type of failure among different groups. The tests are based on comparing weighted averages of the hazards of the subdistribution for the failure type of interest. Asymptotic results are derived by expressing the statistics in terms of counting processes and using martingale central limit theory. It is proposed that weight functions very similar to those for the G^ρ tests from ordinary survival analysis be used. Simulation results indicate that the asymptotic distributions provide adequate approximations in moderate sized samples.

1. Introduction. Consider the competing risks setting where the data consist of failure times for different subjects and where failure is categorized into several distinct and exclusive types. In this paper a method is given for comparing over time the probability of failures of a certain type being observed among different groups. To be precise, suppose there are K independent groups of subjects, and let T_{ik}^0 be the failure time of the ith subject in group k, $i = 1, \ldots, n_k$, and δ_{ik}^0 be the type of failure, $\delta_{ik}^0 = 1, \ldots, J$. The pairs $(T_{ik}^0, \delta_{ik}^0)$ from different subjects in a group are assumed to be independent and identically distributed. However, it is not assumed that the underlying processes leading to failures of different types are acting independently for a given subject. Rather, only quantities which can be identified from the observed data, regardless of whether or not the risks are independent, will be used. Thus quantities have a "crude" rather than a "net" interpretation, see Tsiatis (1975).

Denote the subdistribution function for failures of type j in group k by

$$F_{jk}(t) = P\big(T_{ik}^0 \leq t, \delta_{ik}^0 = j\big).$$

This will be called the cumulative incidence function for failures of type j here [Kalbfleisch and Prentice (1980), pages 168–169, use this term]. The main subject of this paper is to develop tests for the hypothesis

$$H_0\colon F_{1k} = F_1^0, \quad k = 1, \ldots, K, \tag{1.1}$$

where F_1^0 is an unspecified subdistribution function and where the failure type of special interest is taken to be type 1. To simplify the presentation, the $F_{jk}(t)$ are assumed to be continuous with subdensities $f_{jk}(t)$ with respect to Lebesgue measure.

Received January 1987; revised November 1987.

[1]This work was supported by Grants CA-39929 and CA-31247, awarded by the National Cancer Institute, DHHS, and by a grant from the Mellon Foundation.

AMS 1980 *subject classifications.* Primary 62G10; secondary 62E20.

Key words and phrases. Censored data, counting processes, martingales, G^ρ tests.

The motivation for this work came from the setting of clinical trials for the evaluation of cancer therapies. Investigators from the Eastern Cooperative Oncology Group were considering mounting a trial to investigate whether radiotherapy, when added to conventional therapy consisting of surgery and chemotherapy, would prolong the disease-free interval. The investigators wished to use data on conventionally treated patients from earlier studies to identify subgroups of patients where a benefit from radiotherapy was most likely to be observed. Since radiotherapy is only applied locally, the benefit, if any, should be most apparent in those subgroups with the largest number of isolated local failures. Thus methods for comparing the cumulative incidence of isolated local failures from different subgroups were needed. One such comparison is presented in Section 5.

Information on comparisons among treatments of the cumulative incidence of different types of failure could also be useful when selecting the appropriate treatment for a particular patient. For an adjuvant breast cancer patient there are a number of different possible types of treatment failure, including death from a toxic reaction to the therapy, an isolated local recurrence (which can often be successfully treated using only surgery or radiotherapy), appearance of distant metastases, development of a second type of cancer and so on. These different types of failure will not be of equal importance to the patient, and their likelihood may be different for different therapies. Thus, in addition to comparing treatments for time to failure, information on comparisons of the cumulative incidence of the different types of failure may also be useful.

Let $S_k(t) = P(T_{ik}^0 > t) = 1 - \Sigma_j F_{jk}(t)$ denote the survivor function for subjects in group k, and let

$$\lambda_{jk}(t) = f_{jk}(t)/S_k(t) \tag{1.2}$$

be the cause specific hazard for failures of type j in group k. Much of the previous work on analyzing the effect of factors on competing risks has concentrated on examining their effect on the λ_{jk}, see Prentice, Kalbfleisch, Peterson, Flournoy, Farewell and Breslow (1978) and Larson (1984). However, the effect of a factor on the cause specific hazard for a particular type of failure can be quite different than its effect on the cumulative incidence of that type of failure. As an example of this, suppose there are two types of failure, local and distant, and suppose all cause specific hazards are constant, with the cause specific hazards for both local and distant failure being $\lambda_{j1} = 3$ in group 1, while in group 2, $\lambda_{12} = 2$ for local failure and $\lambda_{22} = 1$ for distant failure. Then the cumulative incidence functions for local failure are $F_{11}(t) = (1 - e^{-6t})/2$ in group 1 and $F_{12}(t) = 2(1 - e^{-3t})/3$ in group 2, so $F_{12}(t) > F_{11}(t)$ for $t > (\log 3)/3$ even though $\lambda_{11} > \lambda_{12}$. Differences in the relationships of cause specific hazards and the relationships of cumulative incidences are also seen in the example in Section 5.

As a consequence, the hypothesis of equality of the cumulative incidence functions for failures of type 1 is not equivalent to the hypothesis of equality of the cause specific hazard functions for failures of type 1, except when the survival functions S_k are also equal under the null, see (1.2). Although in

principle the hypothesis (1.1) could be examined by looking at how the cause specific hazards for all causes of failure vary in the different groups, this would often be difficult in practice. The methods given here appear to be the first direct way to examine the hypothesis (1.1).

The form of the proposed test statistics is clearest when only two groups are being compared. For this case it is proposed that tests be based on a score of the form

$$(1.3) \qquad \int_0^\tau K(t)\Big\{\big[1-\hat{F}_{11}(t-)\big]^{-1}\,d\hat{F}_{11}(t)-\big[1-\hat{F}_{12}(t-)\big]^{-1}\,d\hat{F}_{12}(t)\Big\},$$

where $\hat{F}_{1k}$ is an estimate of F_{1k}, see (2.3), and where $K(t)$ is a suitably chosen weight function. Basically, (1.3) compares weighted averages of the "subdistribution hazards" $f_{1k}/(1-F_{1k})$. In Section 2 the class of K-sample tests, generalizations of (1.3), are developed and asymptotic results stated. In Section 3 consideration is given to the choice of the weight function $K(t)$, and a family of tests is proposed which is very similar to the G^ρ tests given by Harrington and Fleming (1982) for ordinary survival analysis. In Section 4 results of a limited simulation study are given, which indicate generally good performance of the tests. Derivations of the asymptotic results are given in Section 6. The derivations are based on a counting process formulation and martingale central limit theory.

2. Development of the K-sample test statistic. In the remainder of the paper, it is assumed that there are only two types of failure ($J = 2$). This does not place any restriction on the generality of the results, since when there are more than two types of failure, all types other than the type of interest can be combined into one "other" category while comparing the cumulative incidence of the type of interest.

Before proceeding with the development, it will be convenient to introduce some additional notation. In general, if F is a subdistribution function, then $G = 1 - F$. Define $n = n_\cdot = \sum_{k=1}^K n_k$. Throughout a subscript replaced by a "$\cdot$" will denote summation over that subscript. Also define $\gamma_{jk}(t) = f_{jk}(t)/G_{jk}(t)$ and $\Gamma_{jk}(t) = \int_0^t \gamma_{jk}(u)\,du$.

In general, the data will be right censored. Let U_{ik} be the censoring time for the (i, k)th subject, with U_{ik} independent of $(T_{ik}^0, \delta_{ik}^0)$. It is assumed that only $T_{ik} = (T_{ik}^0 \wedge U_{ik})$ and $\delta_{ik} = \delta_{ik}^0 I(T_{ik} \leq U_{ik})$ are observed, where $\wedge$ denotes minimum and $I(A)$ is the indicator function of the set A.

The development will be based on the theory of counting processes; see Aalen (1978b). Define

$$(2.1) \qquad N_{jk}(t) = \sum_{i=1}^{n_k} I(T_{ik} \leq t,\ \delta_{ik} = j)$$

and

$$(2.2) \qquad Y_k(t) = \sum_{i=1}^{n_k} I(T_{ik} \geq t).$$

Then $N_{jk}(t)$ is the number of failures of type j by t and $Y_k(t)$ is the number of subjects still at risk just prior to t in group k. An estimate of the cumulative incidence function is then given by

$$\hat{F}_{jk}(t) = \int_0^t \hat{S}_k(u-) Y_k^{-1}(u)\, dN_{jk}(u), \tag{2.3}$$

where $\hat{S}_k(t-)$ is the left-hand limit of the Kaplan–Meier (1958) estimate $\hat{S}_k(t)$ and where, to simplify the notation, $\hat{S}_k(t-)$ is defined to be 0 when $Y_k(t) = 0$ and the convention $0/0 = 0$ is employed.

Aalen (1978a) has given strong consistency and weak convergence results for (2.3). Although Aalen assumes independent risks, Tsiatis (1975) has shown that for dependent risks there is always a hypothetical setting with independent risks which gives the same distribution for the observed data; also see the beginning of Section 6. These results for (2.3) are also an immediate consequence of the more general results of Aalen and Johansen (1978); see also Mode (1976), Fleming (1978a, b) and Gill (1980b). Johansen (1978) showed that the estimators studied by Aalen and Johansen, and thus (2.3) as well, were nonparametric maximum likelihood estimates.

To motivate the form of the test statistic, define (improper) random variables by

$$\begin{aligned} X_{ik} &= T_{ik}^0, \qquad \delta_{ik}^0 = 1, \\ X_{ik} &= \infty, \qquad \delta_{ik}^0 > 1. \end{aligned} \tag{2.4}$$

Then $F_{1k}(t) = P(X_{ik} \le t)$ and $\gamma_{1k}(t)$ is the hazard function for X_{ik}. Thus the statistic (1.3) compares the hazard functions of the X_{ik}. The K-sample statistic will be defined by assigning a score to each group which compares this hazard for each group to a combined estimate of this hazard under the null.

The null subdistribution F_1^0 cannot be estimated by computing (2.3) from the combined data set, since the null hypothesis does not require that either the S_k or the Λ_{1k} be equal for different k. Defining

$$R_k(t) = I(\tau_k \ge t) Y_k(t) \hat{G}_{1k}(t-)/\hat{S}_k(t-)$$

gives

$$\begin{aligned} \hat{\Gamma}_{1k}(t) &= \int_0^t \left[\hat{G}_{1k}(u-)\right]^{-1} d\hat{F}_{1k}(u) \\ &= \int_0^t \left[R_k(u)\right]^{-1} dN_{1k}(u), \quad \text{for } t \le \tau_k, \end{aligned}$$

where the last equality follows from (2.3). The quantities τ_k are fixed times satisfying conditions given in the statement of Theorem 1. In the convergence arguments it will be convenient to have defined $R_k(t) = 0$ for $t > \tau_k$. The expression for $\hat{\Gamma}_{1k}$ suggests taking

$$\hat{\Gamma}_1^0(t) = \int_0^t \left[R_\cdot(u)\right]^{-1} dN_{1\cdot}(u) \tag{2.5}$$

as an estimator for Γ_1^0, the null value of Γ_{1k}. This estimator is consistent under

the null, which can be seen by noting that

$$\hat{\Gamma}_1^0(t) = \sum_k \int_0^t [R_k(u)/R_\cdot(u)]\hat{G}_{1k}^{-1}(u-)\,d\hat{F}_{1k}(u),$$

and recalling that the $\hat{F}_{1k}$ all consistently estimate F_1^0 under the null. K-sample tests thus can be based on scores of the form

$$z_k = \int_0^{\tau_k} K_k(t)\{d\hat{\Gamma}_{1k} - d\hat{\Gamma}_1^0\}, \tag{2.6}$$

where again the $K_k(t)$ are suitably chosen weight functions.

Further motivation for the estimator (2.5) comes from noting $R_k(t)/n_k$ estimates $P(X_{ik} \wedge U_{ik} \geq t)$, so $R_k(t)$ estimates the expected number of X_{ik} still at risk at time t in group k when they are censored by the U_{ik}. Then $R_\cdot(t)$ denotes this same quantity in the pooled sample, so $dN_{1\cdot}(t)/R_\cdot(t)$ is essentially of the form number of events at t divided by the number at risk at t.

In practice the weight functions $K_k(t)$ in (2.6) will generally be of the form $L(t)R_k(t)$, for some function $L(t)$. With this definition for K_k, and setting $K(t)$ in (1.3) equal to $L(t)R_1(t)R_2(t)/[R_1(t) + R_2(t)]$, it is easily verified that (2.6) has the desirable property of reducing to (1.3) when only two groups are being compared.

The asymptotic distribution of the z_k will be given under a sequence of local alternatives where the subdistributions F_{jk}^n are all absolutely continuous with respect to Lebesgue measure, and have densities satisfying

$$\begin{aligned} f_{1k}^n(t) &\to f_1^0(t), \\ f_{2k}^n(t) &\to f_{2k}^0(t) \end{aligned} \tag{2.7}$$

uniformly in t, and

$$n^{1/2}[\gamma_{1k}^n(t) - \gamma_{1r}^n(t)] \to \beta_{kr}(t) \tag{2.8}$$

uniformly in t, where the $\beta_{kr}(t)$ are bounded functions. Note that β_{kk} is identically 0.

THEOREM 1. *Assume* $0 < \alpha_k = \lim n_k/n$ *for each* k. *Let* τ_k, $k = 1, \ldots, K$, *be fixed times satisfying* $\Pi_k^0(\tau_k) > 0$, *where* $\Pi_k^0(t) = \alpha_k P(T_{ik} \geq t)$ *under the null hypothesis. Let* $K_k(t)$ *be predictable processes on* $[0, \tau_k]$ *such that*

$$n^{-1}K_k(t) \to K_k^0(t)$$

uniformly in probability, where each K_k^0 *is bounded on* $[0, \tau_k]$. *Let* $Z = (z_1, \ldots, z_K)'$. *Then under a sequence of local alternatives satisfying* (2.7) *and* (2.8),

$$n^{-1/2}Z \to_D N_k(\mu, \Sigma),$$

where $\to_D$ *denotes convergence in distribution, and where the components of* μ *are*

$$\mu_k = \sum_{r \neq k} \int_0^{\tau_k} K_k^0(t)\beta_{kr}(t)\left[h_r(t)/h_\cdot(t)\right] dt \tag{2.9}$$

and the components of Σ *are*

$$\begin{aligned}\sigma_{kk'}^2 = &\sum_{r=1}^{K} \int_0^{\tau_k \wedge \tau_{k'}} a_{kr}(t)a_{k'r}(t)h_r^{-1}(t)\, dF_1^0(t) \\ &+ \sum_{r=1}^{K} \int_0^{\tau_k \wedge \tau_{k'}} b_{2kr}(t)b_{2k'r}(t)h_r^{-1}(t)\, dF_{2r}^0(t),\end{aligned} \tag{2.10}$$

where

$$\begin{aligned}a_{kr}(t) &= d_{1kr}(t) + b_{1kr}(t), \\ b_{jkr}(t) &= \left[I(j=1) - G_1^0(t)/S_r^0(t)\right]\left[c_{kr}(\tau_k) - c_{kr}(t)\right], \\ c_{kr}(t) &= \int_0^t d_{1kr}(u)\, d\Gamma_1^0(u), \\ d_{jkr}(t) &= I(j=1)K_k^0(t)\left[I(k=r) - h_r(t)/h_\cdot(t)\right]/G_1^0(t)\end{aligned}$$

and

$$h_r(t) = I(t \leq \tau_r)\Pi_r^0(t)/S_r^0(t).$$

The proof of this theorem is outlined in Section 6. A consistent estimate of (2.10) under the null can be obtained by estimating $h_r(t)$ with $\hat{h}_r(t) = n^{-1}I(t \leq \tau_r)Y_r(t)/\hat{S}_r(t-)$, F_{2r}^0 with (2.3), $S_r^0(t)$ with $\hat{S}_r^0(t-)$, K_k^0 with $n^{-1}K_k$ and $F_1^0(t)$ with

$$\hat{F}_1^0(t) = n^{-1}\int_0^t \hat{h}_\cdot^{-1}(u)\, dN_\cdot(u). \tag{2.11}$$

When the functions $K_k(t)$ are of the form $L(t)R_k(t)$, then $\Sigma z_k = 0$, so only $K-1$ of the scores are linearly independent. An appropriate K-sample test statistic can then be formed by using a quadratic form consisting of $K-1$ components of Z and the inverse of their estimated variance–covariance matrix, which asymptotically will have a chi-square distribution with $K-1$ degrees of freedom under the null hypothesis. A stratified version of the test can also be given by computing contributions to the z_k and the $\hat{\sigma}_{kk'}^2$ within each stratum, adding the contributions over strata and proceeding as before.

As a further extension, note that if the risks are assumed to be independent, then the test can easily be modified to test equality of the partial transition probabilities in the multiple decrement model studied by Aalen (1978a). Essentially the only change is to treat transitions to states not in the partial chain as censored failure times.

In the absence of censoring, the entire development is much simpler, as discussed at the end of Section 6. In particular, $\Pi_k^0(t) = \alpha_k S_k^0(t)$, and (3.7)

becomes

$$\sigma^2_{kk'} = \left[\alpha_k^{-1} I(k = k') - 1\right] \int_0^{\tau_k \wedge \tau_{k'}} K_k^0(t) K_{k'}^0(t) \left[G_1^0(t)\right]^{-2} dF_1^0(t).$$

3. A specific class of tests. In this section the choice of the weight function in the scores (2.6) is considered. The discussion will be limited to the two-sample problem. The test then is based on the single score z_1, and only weight functions of the form $L(t)R_1(t)$ will be considered, where $L(t)$ is a predictable process converging uniformly in probability to a bounded function $L^0(t)$. From Theorem 1, the asymptotic efficacy of the test against a sequence of local alternatives satisfying (2.7) and (2.8) is

$$\sigma_{11}^{-2}\left[\int_0^\tau L^0(t)\left[\beta_{12}(t)/\gamma_1^0(t)\right] h_1(t) h_2(t)/h_\cdot(t)\, dF_1^0(t)\right]^2, \tag{3.1}$$

where σ_{11}^2 is given by (2.10), with $K_1^0(t) = L^0(t) G_1^0(t) h_1(t)$ and where $\tau = \tau_1 \wedge \tau_2$.

In general, it does not appear possible to solve for the function L^0 which maximizes (3.1) for a particular alternative. Exceptions to this are when there is no censoring or when there is no competing cause of failure. In these cases the formula simplifies and standard arguments, see Gill (1980a), especially his Lemma 5.2.1, and Schoenfeld (1981), can be applied to show (3.1) is maximized by $L^0 = \beta_{12}/\gamma_1^0$.

For general use, one attractive possibility is to take

$$L(t) = \left[\hat{G}_1^0(t)\right]^\rho, \tag{3.2}$$

where $1 - \hat{G}_1^0$ is defined by (2.11). Then taking ρ large will give more weight to early differences and taking ρ negative will give more weight to later differences. Note that since (3.2) is a function only of $\hat{G}_1^0$, the resulting test will still be invariant to monotone transformations of the data. It is shown in Section 6 that the weight function resulting from (3.2) meets the conditions of Theorem 1.

Further motivation for (3.2) comes from considering the family of alternatives

$$\begin{aligned} G(t;\theta) = 1 - F(t;\theta) &= \left\{1 + \left(\left[G_1^0(t)\right]^{-\rho} - 1\right) e^\theta\right\}^{-1/\rho}, && \rho \neq 0, \\ &= \left[G_1^0(t)\right]^{\exp(\theta)}, && \rho = 0, \end{aligned} \tag{3.3}$$

where the null is $\theta = 0$. For a sequence of local alternatives from this family $\beta_{12}/\gamma_1^0 = [G_1^0]^\rho$, so with either no censoring or no competing cause of failure the test using (3.2) is optimal for the alternative (3.3). Harrington and Fleming (1982) showed this for ordinary survival data, and in fact the test using (3.2) is asymptotically equivalent to their G^ρ test when there is no competing cause of failure.

To give a clearer interpretation of the alternative (3.3), note that under this alternative

$$\frac{G^\rho(t;\theta_1)\left[1 - G^\rho(t;\theta_2)\right]}{\left[1 - G^\rho(t;\theta_1)\right] G^\rho(t;\theta_2)} = e^{\theta_2 - \theta_1},$$

TABLE 1
Empirical sizes of a nominal 5% level test

Censoring distribution	K = 2 test statistic			K = 3 test statistic			K = 5 test statistic		
	$\rho = 1$	$\rho = 0$	$\rho = -1$	$\rho = 1$	$\rho = 0$	$\rho = -1$	$\rho = 1$	$\rho = 0$	$\rho = -1$
0	5.1	4.8	4.6	5.2	4.5	4.5	4.5	4.6	4.8
25	3.9	4.4	4.3	5.4	6.1	5.5	4.1	4.6	4.1
50	4.9	4.9	4.7	4.4	4.3	4.7	4.0	3.7	3.5

for all t. Thus taking $\rho = 1, 0, -1$ specifies that the odds ratio, the hazard ratio $\gamma_1(t; \theta_1)/\gamma_1(t; \theta_2)$ and the cumulative risk ratio $F(t; \theta_2)/F(t; \theta_1)$, respectively, are constant over time.

4. Simulation results. In the simulations the weight functions (3.2), with $\rho = -1, 0, 1$, were used, and all data was used in calculating the statistics. In all cases the number of subjects per group was $n_k = 50$, and there were two types of failure. The first set of simulations, given in Table 1, examined the size of the tests. The number of groups used was $K = 2$, 3 and 5. The probability of each type of failure was 1/2, with the conditional failure distributions unit exponential. The censoring distributions used were no censoring and uniform $(0, C)$ censoring with $C = 3.9207$ (25% censored) and $C = 1.59362$ (50% censored).

The second set of simulations, given in Table 2, compares the power of the tests using the three different values of ρ. In all cases the subdistribution for failures of type 1 in group 1 was $G_1^0(t) = 0.5(1 - e^{-t})$, with the subdistribution for failures of type 1 in group 2 given by (3.3), with $(\rho, \theta) = (-1, 1.5), (0, 2), (1, 3)$. The values of θ were chosen so that 75% of the failures in group 2 would be type 1 in the absence of censoring. The censoring distributions used here were identical to those used in the first set of simulations, and the conditional distributions of failures of type 2 were again taken to be unit exponential in each group.

TABLE 2
Empirical powers

Censoring distribution	Alternative $(\rho,\theta) = (-1,1.5)$ test statistic			Alternative $(\rho,\theta) = (0,2)$ test statistic			Alternative $(\rho,\theta) = (1,3)$ test statistic		
	$\rho = 1$	$\rho = 0$	$\rho = -1$	$\rho = 1$	$\rho = 0$	$\rho = -1$	$\rho = 1$	$\rho = 0$	$\rho = -1$
0	58.4	65.6	66.8	72.6	75.7	73.1	84.3	82.0	74.2
25	43.9	49.8	52.4	64.0	66.4	63.4	80.4	79.3	73.0
50	27.4	29.8	30.2	47.1	49.2	48.7	71.9	71.0	67.0

In each case 1000 simulated samples were generated, and the percent of samples where the test exceeded the upper 5% critical value of the appropriate chi-square distribution was calculated for each test. Thus binomial standard errors can be used for the entries in Tables 1 and 2, although it should be noted that in each case the three tests are computed from the same samples. Uniform random numbers were generated using IMSL routine GGUBFS, and then transformed using the inverse cumulative distribution functions.

The simulations with $K = 2$ were repeated with a log-logistic distribution for the conditional distribution of the failures of type 2 in group 1, to investigate the effect of having different failure distributions in the two groups for the competing cause. The results were very similar to the results in Tables 1 and 2 and are omitted.

The empirical sizes in Table 1 appear adequate, with only one of the entries more than 2 standard errors from the nominal size of 5%. For the powers in Table 2, two features stand out. One is that the test with $\rho = m$ had the best power for the alternative with $\rho = m$ in all cases. The second is that the differences in power are quite small. Although differences for ordinary survival data are not much larger, see Latta (1981), this does suggest that in applications where $\hat{G}_1^0(\tau)$ is fairly large, as in the example in the following section, one may need to consider values of ρ more extreme than ± 1 to seriously alter where the power of the test is focused.

5. Example. The data are taken from two adjuvant breast cancer trials conducted by the Eastern Cooperative Oncology Group. Here the effect of number of positive nodes, a major prognostic factor in breast cancer, is examined. As discussed in the Introduction, the goal is to identify patients who are at higher risk of developing isolated local recurrences, with distant recurrences being the competing type of failure. Table 3 gives the number of patients and the percents with isolated local and distant recurrences by number of positive nodes. Patients with both local and distant involvement at recurrence are included in the distant category. To get an idea of the amount of follow-up at the time of this analysis, there were 430 patients at risk at 3 years of follow-up, 138 at risk at 5 years and the maximum follow-up was 7 years.

TABLE 3
Summary of breast cancer data

	Number of positive nodes			
	1–3	4–7	> 7	Total
Number of patients	388	223	163	774
Percent with isolated local recurrence	11.3	17.9	19.6	15.0
Percent with distant recurrence	24.0	30.5	52.8	31.9

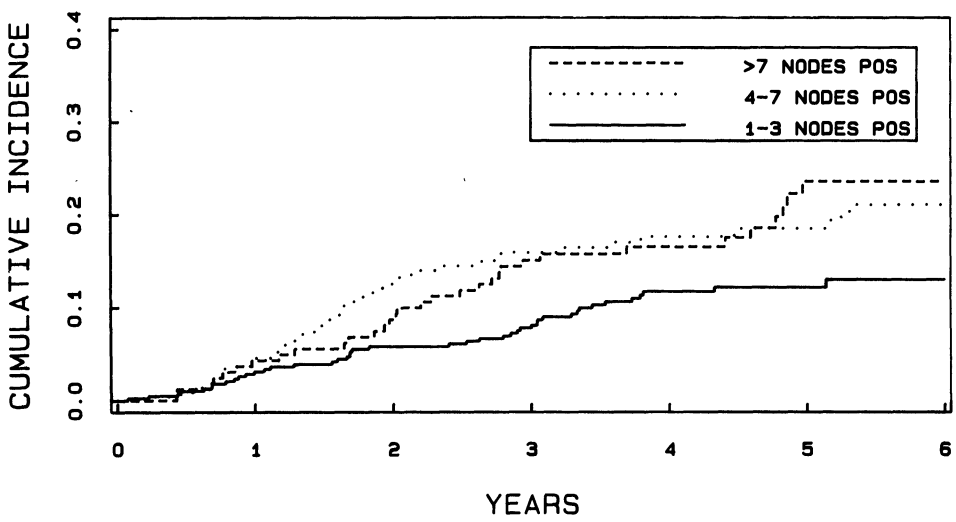

FIG. 1. *Cumulative incidence of local failure by number of positive nodes.*

Figure 1 gives the cumulative incidence of isolated local failure by nodal status. *P*-values using the test with weight function (3.2) with $\rho = 0$ are 0.02 for the overall three-way comparison, 0.02 for the pairwise comparisons of the 1–3 node positive group to either the 4–7 group or the > 7 group and 0.89 for the pairwise comparison of 4–7 to > 7. Thus patients with 4 or more nodes positive appear to be more likely to have isolated local recurrences.

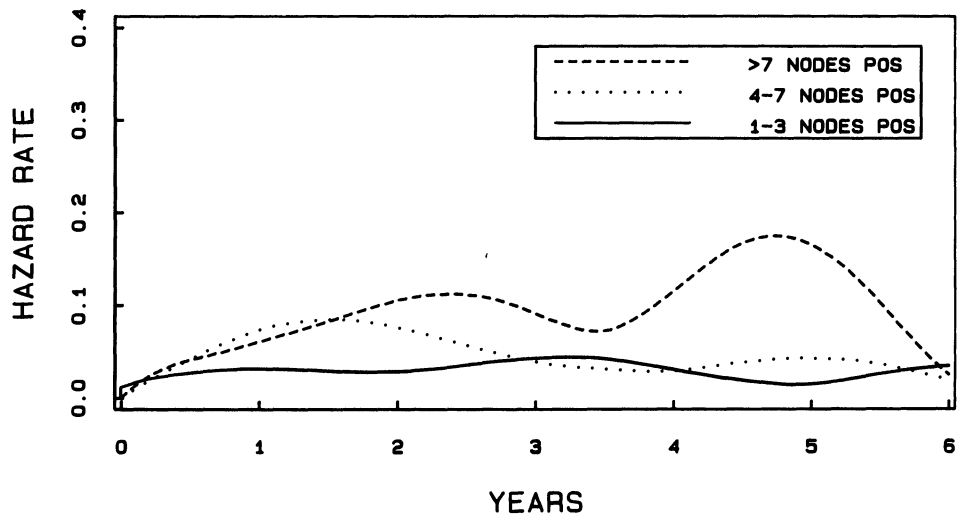

FIG. 2. *Cause specific hazard for local failure by number of positive nodes.*

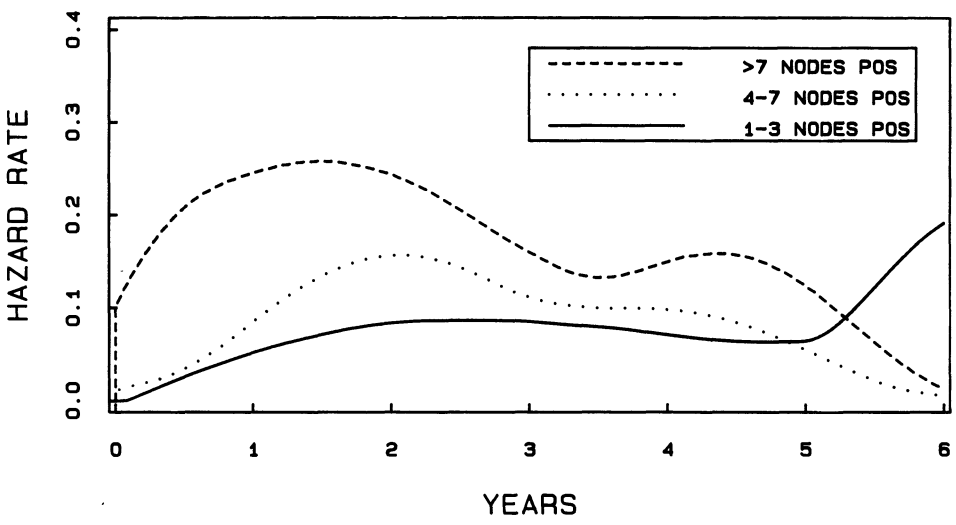

FIG. 3. *Cause specific hazard for distant failure by number of positive nodes.*

Estimated cause specific hazards for local and distant failures are given in Figures 2 and 3. The cumulative hazard estimators studied by Aalen (1976, 1978b) were used, and the smoothed hazard estimate calculated using an approach similar to that given by Ramlau-Hansen (1983), with a bi-weight kernel and a window radius of 1.5 years. As discussed in the Introduction, relationships between cause specific hazards and cumulative incidence functions can be very different. Here the hazard for local failure in Figure 2 is larger in the > 7 group than in the 4–7 group, while the cumulative incidences for the two groups are nearly equal. This is due to the large difference in distant hazards in Figure 3. However, given the complexity of the relationships between the cause specific hazards, it hardly seems possible to infer the equality of the cumulative incidence functions for these two groups directly from the hazards.

6. Derivation of the asymptotic results. Let N_{jk} and Y_k be as defined by (2.1) and (2.2), and set

$$M_{jk}^n(t) = N_{jk}(t) - \int_0^t Y_k(u)\, d\Lambda_{jk}^n(u).$$

Then M_{jk}^n are orthogonal square integrable martingales with predictable variance processes

$$\langle M_{jk}^n(t), M_{jk}^n(t)\rangle = \int_0^t Y_k(u)\, d\Lambda_{jk}^n(u).$$

The filtration assumed here is the one generated by the processes N_{jk} and Y_k. This result will follow from Theorem 3.1.1 of Gill (1980). To put the current problem into Gill's setting, we can think of the failure times as being the minimum of latent failure times for each cause. Although the risks are not

assumed to be independent here, Tsiatis (1975) has shown that regardless of the distribution of the observed data, there are hypothetical independent latent failure times which give the same distribution for the observed data. The set of hypothetical latent failure times for a given subject, each censored by the other hypothetical latent failure times and by the U_{ik}, then meet the conditions of Gill's theorem.

The first result given here is that the estimator $\hat{F}_1^0$ defined by (2.11) converges uniformly in probability to F_1^0 on $[0, \tau_m]$, where $\tau_m = \max\{\tau_k\}$. This will establish that the weighting functions proposed at (3.2) meet the conditions of Theorem 1. Now

$$\hat{F}_1^0(t) - F_1^0(t) = \sum_k \left\{ \int_0^t \frac{\hat{h}_k(u)}{\hat{h}_\cdot(u)} \frac{\hat{S}_k(u-)}{Y_k(u)} \, dM_{1k}^n(u) + \int_0^t \left[\frac{\hat{h}_k(u)}{\hat{h}_\cdot(u)} \hat{S}_k(u-)\lambda_{1k}^n(u) - \frac{h_k(u)}{h_\cdot(u)} S_k^0(u)\lambda_{1k}^0(u) \right] du \right\}.$$

The second integral converges uniformly to 0 in probability on $[0, \tau_m]$ because the integrand does, since each component function on the left converges uniformly to the corresponding function on the right, which in each case is bounded, and because $\tau_m < \infty$. Convergence of the first integral can be established using Lenglart's (1977) inequality [see Gill (1980a), page 18].

Consistency of the variance estimate proposed in Section 2 can be established using very similar methods and will not be given. Next the proof of Theorem 1 is outlined. Further details are given in a technical report available from the author.

PROOF OF THEOREM 1. Setting

$$\eta_k^n = \sum_r \int_0^{\tau_k} n^{-1} K_k(t) R_r(t) R_\cdot^{-1}(t) \sqrt{n} \{\gamma_{1k}^n(t) - \gamma_{1r}^n(t)\} \, dt,$$

it is easily verified that η_k^n converges in probability to μ_k, so it remains to show that the vector W, whose components are

$$w_k = n^{-1/2} z_k - \eta_k^n = \sum_{r=1}^K \int_0^{\tau_k} \frac{K_k(t)}{n} \left[I(k=r) - \frac{R_r(t)}{R_\cdot(t)} \right] \times \sqrt{n} \left\{ \left[\frac{d\hat{F}_{1r}(t) - dF_{1r}^n(t)}{\hat{G}_{1r}(t-)} \right] + \left[\frac{1}{\hat{G}_{1r}(t-)} - \frac{1}{G_{1r}^n(t)} \right] dF_{1r}^n(t) \right\}, \tag{6.1}$$

converges in distribution to a $N_K(0, \Sigma)$ distribution.

Using algebraic manipulations, integration by parts, (2.3) and formula 3.2.12 of Gill (1980a), it can be verified that (6.1) can be expressed as

$$n^{-1/2}z_k - \eta_k^n = \sum_{r=1}^{K}\sum_{j=1}^{2}\left\{A_{jkr}^n(\tau_k) + C_{kr}^n(\tau_k)B_{jr}^n(\tau_k)\right\},$$

where

$$A_{jkr}^n(t) = \int_0^t \left[D_{jkr}^n(u) - E_{jr}^n(u)C_{kr}^n(u)\right]\hat{h}_r^{-1}(u)n^{-1/2}\,dM_{jr}^n(u),$$

$$B_{jr}^n(t) = \int_0^t E_{jr}^n(u)\hat{h}_r^{-1}(u)n^{-1/2}\,dM_{jr}^n(u),$$

$$C_{kr}^n(t) = \int_0^t D_{1kr}^n(u)\,dF_{1r}^n(u),$$

$$D_{jkr}^n(t) = I(j=1)n^{-1}K_k(t)\left[I(k=r) - R_r(t)/R_\cdot(t)\right]/\hat{G}_{1r}(t-),$$

$$E_{jr}^n(t) = I(j=1) - G_{1r}^n(t)/S_r^n(t).$$

The joint asymptotic normality of the $A_{jkr}(\tau_k)$ and $B_{jr}(\tau_k)$ follows from Theorem 2.1 of Andersen, Borgan, Gill and Keiding (1982). The conditions in Theorem 1 have been given so that the conditions of the theorem of Andersen, Borgan, Gill and Keiding can easily be verified, by showing that the integrands converge uniformly in probability. The covariance calculations are also straightforward. The result then follows from the continuous mapping theorem [see, e.g., Billingsley (1968), page 34]. □

In the absence of censoring, the X_{ik} defined by (2.4) are observed, and the tests introduced here reduce to standard survival analysis tests for comparing the hazards of the X_{ik}. A much simpler development can then be given, using counting processes defined from the X_{ik} and many of the results of Aalen (1978b), Gill (1980a) and Andersen, Borgan, Gill and Keiding (1982) can be applied directly. Note that even though the X_{ik} are improper random variables, this creates no problems for the counting process formulation, and Gill specifically allows improper random variables. The reason the results are more complicated with censoring is that when a subject fails from a competing cause, so that $X_{ik} = \infty$, the censoring time U_{ik} is not observed, so that appropriate risk sets cannot be defined.

Acknowledgments. The author wishes to thank David Harrington for helpful discussions during the preparation of this manuscript, the referees and Associate Editor for helpful comments and the Eastern Cooperative Oncology Group for permission to use their data.

REFERENCES

AALEN, O. (1976). Nonparametric inference in connection with multiple decrement models. *Scand. J. Statist.* **3** 15–27.

AALEN, O. (1978a). Nonparametric estimation of partial transition probabilities in multiple decrement models. *Ann. Statist.* **6** 534–545.

AALEN, O. (1978b). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6** 701–726.

AALEN, O. and JOHANSEN, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scand. J. Statist.* **5** 141–150.

ANDERSEN, P. K., BORGAN, O., GILL, R. and KEIDING, N. (1982). Linear nonparametric tests for comparisons of counting processes, with applications to censored survival data. *Internat. Statist. Rev.* **50** 219–258.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

FLEMING, T. R. (1978a). Nonparametric estimation for nonhomogeneous Markov processes in the problem of competing risks. *Ann. Statist.* **6** 1057–1070.

FLEMING, T. R. (1978b). Asymptotic distribution results in competing risks estimation. *Ann. Statist.* **6** 1071–1079.

GILL, R. D. (1980a). *Censoring and Stochastic Integrals. Math. Centre Tracts* **124**. Math. Centrum, Amsterdam.

GILL, R. D. (1980b). Nonparametric estimation based on censored observations of a Markov renewal process. *Z. Wahrsch. verw. Gebiete* **53** 97–116.

HARRINGTON, D. P. and FLEMING, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69** 553–566.

JOHANSEN, S. (1978). The product limit estimator as a maximum likelihood estimator. *Scand. J. Statist.* **5** 195–199.

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

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

LARSON, M. G. (1984). Covariate analysis of competing-risks data with log-linear models. *Biometrics* **40** 459–469.

LATTA, R. B. (1981). A Monte Carlo study of some two-sample rank tests with censored data. *J. Amer. Statist. Assoc.* **76** 713–719.

LENGLART, E. (1977). Relation de domination entre deux processus. *Ann. Inst. H. Poincaré Sect. B (N.S.)* **13** 171–179.

MODE, C. J. (1976). A large sample investigation of a multiple decrement life-table estimator. *Math. Biosci.* **32** 111–123.

PENTICE, R. L., KALBFLEISCH, J. D., PETERSON, A. V., JR., FLOURNOY, N., FAREWELL, V. T. and BRESLOW, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34** 541–554.

RAMLAU-HANSEN, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11** 453–466.

SCHOENFELD, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68** 316–319.

TSIATIS, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proc. Nat. Acad. Sci. U.S.A.* **72** 20–22.

DIVISION OF BIOSTATISTICS AND EPIDEMIOLOGY
DANA-FARBER CANCER INSTITUTE
44 BINNEY STREET
BOSTON, MASSACHUSETTS 02115