BIOMETRICS 39, 675–682
September 1983

# Small-Sample Properties of Censored-Data Rank Tests

**A. M. Kellerer and D. Chmelevsky**

Institut für Medizinische Strahlenkunde der Universität Würzburg,
Versbacher Strasse 5, D-8700 Würzburg, West Germany

SUMMARY

Small-sample properties of censored-data rank tests, such as Mantel's logrank test, the Breslow test or the general class of tests proposed by Peto and Peto (1972, *Journal of the Royal Statistical Society, Series A* **135**, 185–206), are examined. It is found that the latter class of tests, in particular, can be substantially nonconservative when applied to small samples. The most serious inaccuracies occur in unbalanced trials, when higher event rates are inferred in the smaller sample.

## 1. Introduction

The Mantel–Haenszel or logrank test (Mantel, 1966; Peto and Peto, 1972; Cox, 1972) and various generalizations of the Wilcoxon test (Breslow, 1970; Peto and Peto, 1972; Prentice 1978) permit a comparison of hazard functions in samples with censored observations. The tests are based on the standard normal approximation and are valid asymptotically for event numbers that are sufficiently large. However, in actual cases it can be difficult to judge the applicability of the standard normal approximation. There have been, mostly in the context of the assessment of the power of censored-data tests, a number of investigations of the nominal and the exact sizes of these tests (Lee, Desu and Gehan, 1975; Muenz, Green and Byar, 1977; Peace and Flora, 1978; Lininger *et al.*, 1979; Latta, 1981). Substantially nonconservative behaviour of the logrank test and the generalized Wilcoxon tests has been found by Latta in the comparison of a sample of size 10 with a sample of size 50. The computations reported here will demonstrate, for a range of typical cases, in the absence of censoring, the discrepancies between exact significance levels and levels based on the normal approximation. In particular, it will be shown that, even with the usual continuity correction, the tests are not always conservative; applied to small samples they can yield approximate error levels that are considerably smaller than the exact levels under the null hypothesis.

## 2. Comparison of Exact and Approximate Probability Levels for the Logrank Test

If the individuals in two observed groups are subject to the stochastic occurrence of certain events, but can also be lost from observation for reasons unrelated to the events under study, one speaks of right-censored data. Suppose that events are observed at Times $t_i$, $i = 1, \ldots, I$, in two groups of initial sizes $N$ and $M$. Ideally, all events occur at distinct times. In practice the number of events, $k_i$, at Time $t_i$ may be larger than 1. The number of events in Group 1 at $t_i$ is designated by $n_i$, and the number of events in Group 2 by $m_i$. The number of individuals at risk in Group 1 just prior to Observation $i$ is designated by $N_i$, the total number of individuals still at risk in both groups by $K_i = N_i + M_i$, and the proportion of all individuals who are in Group 1 by $\eta_i = N_i/K_i$. The logrank test with the frequently employed

*Key words:* Censored-data rank tests; Mantel–Haenszel test; Logrank test; Breslow test; Generalized Wilcoxon test; Small-sample inaccuracies.

continuity correction uses the variate

$$z = \left\{ \left| \sum_{i=1}^{I} (n_i - E_i) \right| - \tfrac{1}{2} \right\} \Big/ \left( \sum_{i=1}^{I} V_i \right)^{\frac{1}{2}} \tag{1}$$

which is taken to follow the standard normal distribution under the null hypothesis of equality of the hazard functions in the two groups. If each observation entails only one event the expectations and variances are estimated as

$$\left.\begin{aligned} E_i &= \eta_i \\ \text{and} \qquad V_i &= \eta_i(1 - \eta_i). \end{aligned}\right\} \tag{2}$$

More generally one has

$$\left.\begin{aligned} E_i &= k_i\eta_i \\ \text{and} \qquad V_i &= \{k_i\eta_i(1 - \eta_i)(K_i - k_i)\}/(K_i - 1). \end{aligned}\right\} \tag{3}$$

That the test is not exact for small sample sizes requires no explanation, but it is of interest to investigate whether the commonly applied continuity correction suffices to keep the test conservative when it is applied to small samples.

For this purpose the approximate and exact $p$-values have been compared under the null hypothesis of equal hazard functions and in the absence of censoring. An approximate one-sided $p$-value is associated with a standard normal deviate, $z$. The exact one-sided $p$-value was obtained by exhaustive enumeration to see how many times the value from (1) equals or exceeds $z$. This number was divided by $\binom{N+M}{N}$ to obtain the exact significance level which corresponds to the approximate level $p$. Figure 1 gives the results for a first group of size $N$ tested against a second group of size $M$; the values $N$ and $M$ are given as parameters. The differences between actual and nominal levels, one-sided for higher event rates in Group 1, are plotted against the nominal levels. The computations are exact except for the cases (6, 24), (10, 20) and (20, 10), where Monte Carlo computations ($5 \times 10^4$ runs for each case) rather than exhaustive calculations were performed. Results for the limiting case of one infinite group are also exact, and are based on relations given in the Appendix.

Even the limited set of values of $N$ and $M$ suffices to show the trend and magnitude of the deviations from the exact significance levels. One concludes that the logrank test with the continuity correction is largely conservative if applied to small samples; however, it can be nonconservative if higher event rates are inferred in a small sample compared to a large sample. For high significance levels the discrepancies are substantial in such unbalanced comparisons. This type of error is to be expected if the standard normal approximation is applied to cases where the exact distribution must be skewed.

The logrank test is frequently applied without the continuity correction. Figure 2 gives results for this case, and demonstrates serious nonconservative discrepancies even in a balanced comparison with two samples each of size 10. In unbalanced trials the inaccuracies can be prohibitive; as can be seen in the lower panel of Fig. 2 they are substantial even for fairly large samples. The computations are exact except for the cases (10, 20), (20, 10), (10, 50), and (50, 10) where Monte Carlo computations were used ($5 \times 10^4$ runs for each case). For the case of one infinite group, see the Appendix. Earlier findings by Latta (1981) are consistent with the results. The diamonds in the upper panel of Fig. 2 show Latta's results and the standard errors for the cases (10, 50), (10, 10) and (50, 10).

## 3. Other Censored-Data Rank Tests

The logrank test has optimal power if the two compared groups have proportional-hazard functions (Peto and Peto, 1972). If the hazard functions differ only during part of the

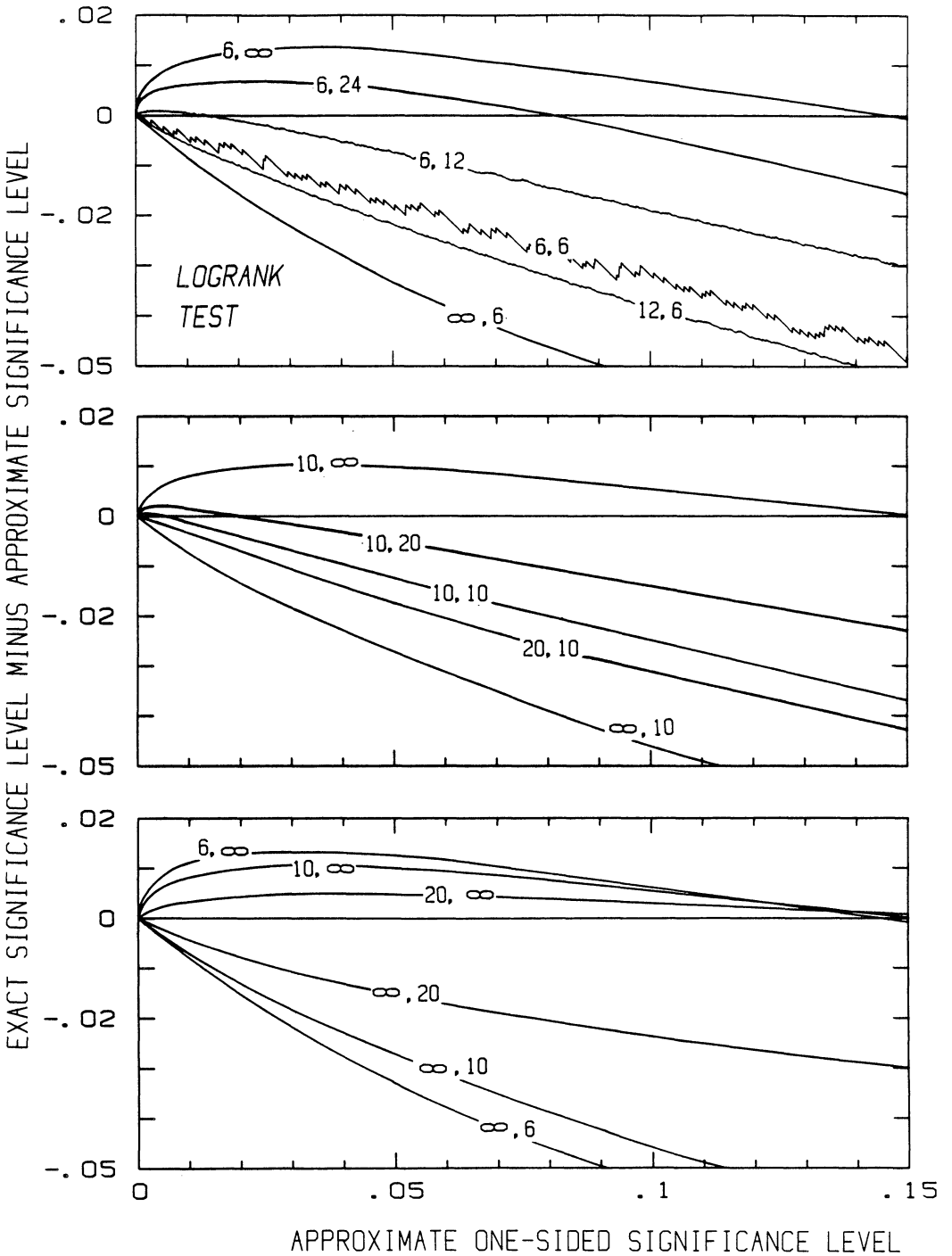


Figure 1. Differences between the exact and approximate significance levels under the null hypothesis for the logrank test with continuity correction, applied to small samples without censoring. The parameters on the curves give the sizes of the two samples; the significance levels are one-sided for higher event rates in the first sample.

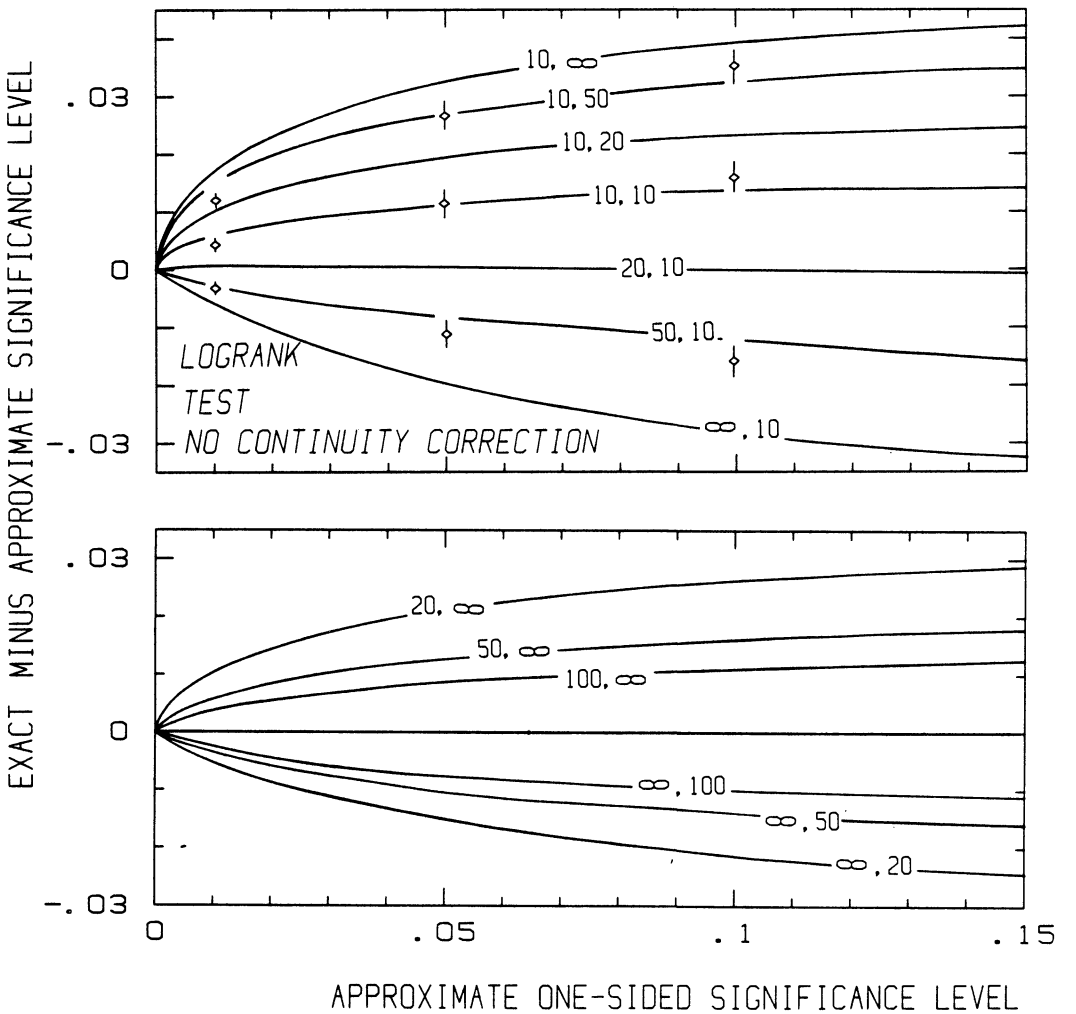


Figure 2. Differences between the exact and approximate significance levels under the null hypothesis for the logrank test without continuity correction, applied to small samples without censoring. The parameters on the curves give the sizes of the two samples; the significance levels are one-sided for higher event rates in the first sample.

observation period, rank tests that attribute enhanced weights to observations in this phase can be more suitable. Peto and Peto (1972) and Prentice (1978) have shown that a class of such tests can be generated if the statistic

$$z = \left\{\sum_{i=1}^{I} w_i(n_i - E_i)\right\} \Big/ \left(\sum_{i=1}^{I} w_i^2 V_i\right)^{\frac{1}{2}} \tag{4}$$

is used, the variables being defined as in (2) and (3). Different choices of the weights, $w_i$, result in different tests; the logrank test, without continuity correction, corresponds to the simple case $w_i = 1$.

The Gehan or Breslow test (Gehan, 1965; Breslow, 1970) utilizes the numbers at risk as weight factors, $w_i = K_i$. Prentice and Marek (1979) chose the survivor-function estimator as the weight factor,

$$w_i = \prod_{j=1}^{i} \left\{K_j \Big/ (K_j + k_j)\right\}, \tag{5}$$

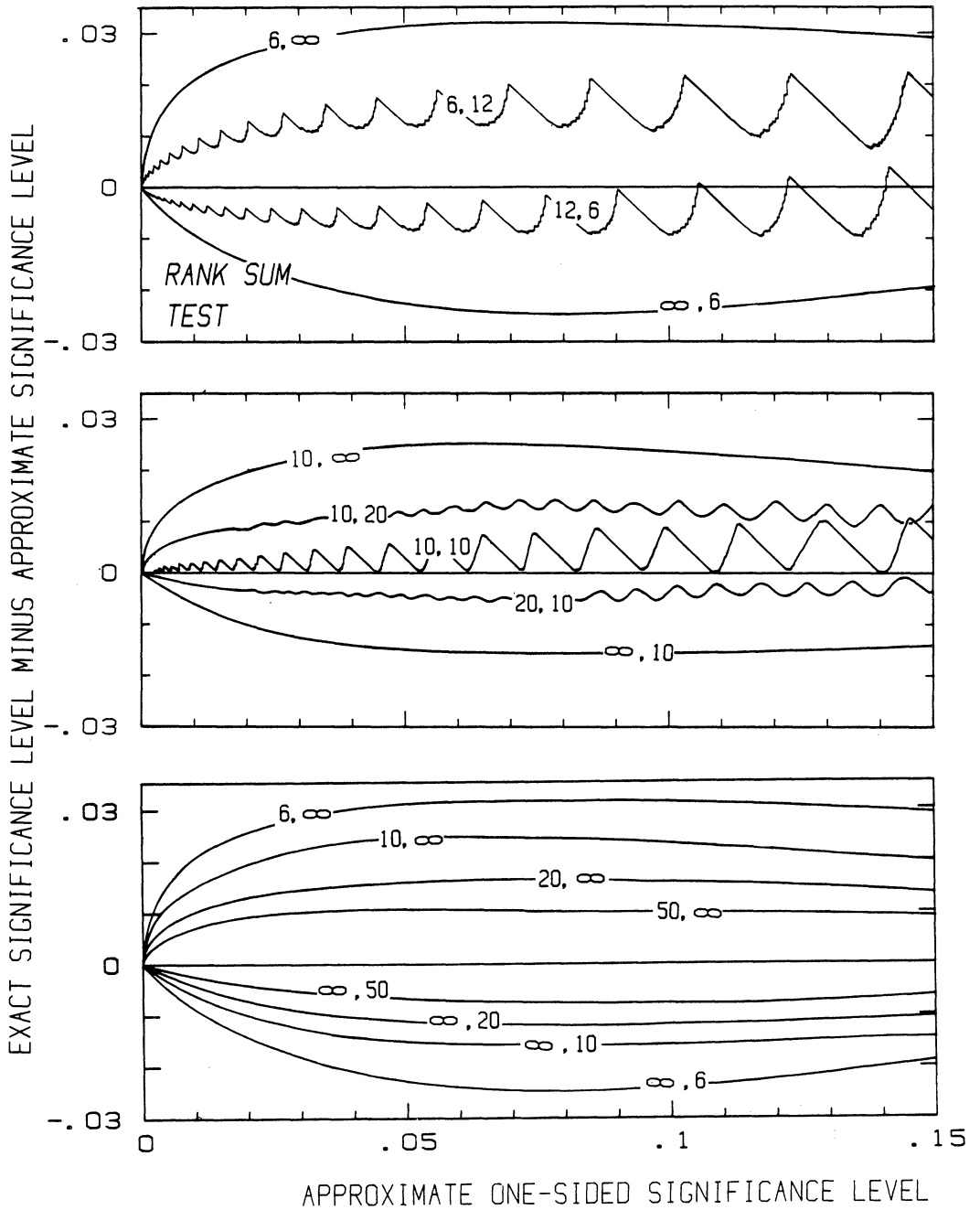


Figure 3. Difference between the exact and approximate significance levels under the null hypothesis for the rank-sum test (Wilcoxon generalization) applied to small samples without censoring. The parameters on the curves give the sizes of the two samples; the significance levels are one-sided for higher event rates in the first sample.

and referred to it as the Peto generalized Wilcoxon statistic. An explicit version of the test, based on successive convolutions (Kellerer, 1973), has been used with the weights set equal to the descending ranks of the exact observations. In the absence of censoring these three versions of the rank-sum test reduce equally to the Wilcoxon statistic. The subsequent computations for the uncensored case apply, therefore, to all three Wilcoxon generalizations.

It is clear that, in the absence of censoring, the Wilcoxon test is preferable; however, the limiting case of no censoring will serve to indicate the inaccuracies of the censored-data rank tests when they are applied to small samples and based on the normal approximation.

Figure 3 is analogous to Fig. 1 and compares one-sided actual and nominal error levels under the null hypothesis for the standard normal approximation. As in the earlier computations, the results were obtained by exhaustive computation of $z$ for all $\binom{N+M}{N}$ permutations of events in two uncensored groups. The calculations are exact except in the cases (10, 20) and (20, 10) where Monte Carlo computations were used ($5 \times 10^4$ runs for each case). The limiting case with one infinite group is also based on Monte Carlo simulations ($5 \times 10^4$ runs) and on relationships given in the Appendix. The results are in fair agreement with the values given by Latta (1981) for the cases (10, 10), (10, 50) and (50, 10).

One concludes that the tests can be substantially nonconservative if applied to small samples. As with the logrank test, and as expected for the standard normal approximation, the errors are most critical if larger event rates are inferred in the smaller sample in an unbalanced trial. Extreme nominal significance levels can then be mere artifacts caused by the standard normal approximation.

## 4. Conclusion

One-sided nominal error levels can deviate markedly from the exact levels when the logrank test or the generalizations of the Wilcoxon test are applied to small samples. The errors are generally less serious for the logrank test with continuity correction; however, in unbalanced trials when higher event rates are inferred in the smaller group, this test, too, can be highly nonconservative. On the other hand, the tests are overly conservative when lower effect rates are inferred in the smaller sample. The mere utilization of a sufficiently increased continuity correction does not adequately improve the small-sample properties of the tests.

These results confirm the statements of Prentice and Marek (1979) who, applying various rank tests to a highly unbalanced trial, expressed doubts concerning the validity of the significance levels for higher event rates in the smaller sample. Methods that are not based on the standard normal approximation are therefore desirable.

### Acknowledgement

This work was supported by Euratom Contract 208-76-7 BIO D.

### Résumé

Nous examinons les propriétés de petits échantillons pour des tests de rang de données censurées, telles que le test du logarithme de rang de Mantel, le test de Breslow ou la classe généralisée des tests proposés par Peto et Peto (1972, *Journal of the Royal Statistical Society, Series A* **135,** 185–206). Nous trouvons que cette dernière classe de tests, en particulier, peut être très fortement non conservative quand on l'applique à de petits échantillons. Les plus graves erreurs se produisent dans des essais déséquilibrés, quand on infére des taux élevés d'évènements dans l'échantillon le plus petit.

### References

Breslow, N. (1970). A generalized Kruskal–Wallis test for comparing $K$ samples subject to unequal patterns of censorship. *Biometrika* **57,** 579–594.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika* **52,** 203–223.

Kellerer, A. M. (1973). Comparison of effect rates in samples of time decreasing size. In *Annual Report on Research Project, Radiological Research Laboratory, Columbia University USAEC-COO*-3243-2, 214–222. Springfield: National Technical Information Service, U. S. Dept of Commerce.

Latta, R. B. (1981). A Monte Carlo study of some two-sample rank tests with censored data. *Journal of the American Statistical Association* **76,** 713–719.
Lee, E. T., Desu, M. M. and Gehan, E. A. (1975). A Monte Carlo study of the power of some two-sample tests. *Biometrika* **62,** 425–432.
Lininger, L., Gail, M. H., Green, S. B. and Byar, D. P. (1979). Comparison of four tests for equality of survival curves in the presence of stratification and censoring. *Biometrika* **66,** 419–428.
Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50,** 163–179.
Mantel, N. and Ciminera, J. L. (1979). Use of logrank-scores in the analysis of litter-matched data of time to tumor appearance. *Cancer Research* **38,** 4308–4315.
Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22,** 719–748.
Muenz, L. R., Green, S. B. and Byar, D. P. (1977). Applications of the Mantel–Haenszel statistic to the comparison of survival distributions. *Biometrics* **33,** 617–626.
Peace, K. and Flora, R. (1978). Size and power assessments of tests of hypotheses on survival parameters. *Journal of the American Statistical Association* **73,** 129–132.
Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A* **135,** 185–206.
Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65,** 167–179.
Prentice, R. L. and Marek, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics* **35,** 861–867.

*Received March* 1981; *revised November* 1981 *and March* 1982

## Appendix

### *Nominal and Actual Significance Levels in the Limit Case of One Infinite Group and in the Absence of Censoring*

Let $S(t)$ be the survival distribution of the infinite population and let $T_\lambda$, $\lambda = 1, 2, \ldots, N$, be the death times in the finite sample. Then, under the null hypothesis $\text{pr}(T_\lambda > t) = S(t)$, the variables $S(T_\lambda)$ are independent uniform $|\,0, 1\,|$ variates.

*Logrank Test*

Let $X$ be the limit of $\sum_1^I E_i$ and of $\sum_1^I V_i$ [see (2)] for $M \to \infty$:

$$X = \sum_{\lambda=1}^{N} \int_0^{T_\lambda} -\frac{dS(t)}{S(t)}$$

$$= -\sum_{\lambda=1}^{N} \ln S(T_\lambda). \tag{A.1}$$

The variables $-\ln S(T_\lambda)$ are independently exponentially distributed; accordingly, $X$ follows the gamma distribution of order $N$,

$$G_N(x) = \text{pr}(X < x)$$

$$= \int_0^x \tau^{N-1}\exp(-\tau)\, d\tau \Big/ (N-1)!. \tag{A.2}$$

For the specified nominal level, $p$, in the standard normal approximation and the corresponding value $z = (|\,x - N\,| - \frac{1}{2})/x^{\frac{1}{2}}$ or $z = (x - N)/x^{\frac{1}{2}}$, one derives $x$. From $x$, one obtains the exact one-sided significance levels $G_N(x)$ and $1 - G_N(x)$.

*Rank-Sum Test*

The weights, $w$, are set equal to the survival fractions, $S(T_\lambda)$. The distributions of the test statistic, $z$, are obtained by Monte Carlo simulation from the independently uniformly distributed variables $S(T_\lambda)$.

The following relations for $M \to \infty$ are used:

$$\left.\begin{aligned} &\sum_{i=1}^{I} w_i n_i \to \sum_{\lambda=1}^{N} S(T_\lambda), \\ &\sum_{i=1}^{I} w_i E_i \to \sum_{\lambda=1}^{N} \int_0^{T_\lambda} - dS(t) \\ &\qquad = \sum_{\lambda=1}^{N} \{1 - S(T_\lambda)\}, \end{aligned}\right\} \tag{A.3}$$

$$\begin{aligned} &\sum_{i=1}^{I} w_i^2 V_i \to \sum_{\lambda=1}^{N} \int_0^{T_\lambda} - S(t)\, dS(t) \\ &\qquad = \tfrac{1}{2} \sum_{\lambda=1}^{N} \{1 - S(T_\lambda)^2\}. \end{aligned}$$

Accordingly, one has

$$z = \left[\sum_{\lambda=1}^{N} \{2S(T_\lambda) - 1\}\right] \Big/ \left[\tfrac{1}{2} \sum_{\lambda=1}^{N} \{1 - S(T_\lambda)^2\}\right]^{\frac{1}{2}}. \tag{A.4}$$