

Exact Log-Rank Tests for Unequal Follow-Up

Georg Heinze,^{1,*} Michael Gnant,² and Michael Schemper¹

¹Department of Medical Computer Sciences, Section of Clinical Biometrics,
University of Vienna Medical School, Spitalgasse 23, A-1090 Vienna, Austria

²Department of Surgery, Section of General Surgery, University of Vienna Medical School,
Währinger Gürtel 18-20, A-1090 Vienna, Austria

* *email*: Georg.Heinze@akh-wien.ac.at

SUMMARY. The asymptotic log-rank and generalized Wilcoxon tests are the standard procedures for comparing samples of possibly censored survival times. For comparison of samples of very different sizes, an exact test is available that is based on a complete permutation of log-rank or Wilcoxon scores. While the asymptotic tests do not keep their nominal sizes if sample sizes differ substantially, the exact complete permutation test requires equal follow-up of the samples. Therefore, we have developed and present two new exact tests also suitable for unequal follow-up. The first of these is an exact analogue of the asymptotic log-rank test and conditions on observed risk sets, whereas the second approach permutes survival times while conditioning on the realized follow-up in each group. In an empirical study, we compare the new procedures with the asymptotic log-rank test, the exact complete permutation test, and an earlier proposed approach that equalizes the follow-up distributions using artificial censoring. Results confirm highly satisfactory performance of the exact procedure conditioning on realized follow-up, particularly in case of unequal follow-up. The advantage of this test over other options of analysis is finally exemplified in the analysis of a breast cancer study.

KEY WORDS: Censored data; Conditional tests; Generalized Wilcoxon tests; Linear-rank tests; Permutation tests; Survival analysis; Unequal censoring.

1. Introduction

The log-rank test (Mantel, 1966; Peto and Peto, 1972) and two variants of the generalized Wilcoxon test (by Breslow, 1970, and by Prentice, 1978) currently are the standard nonparametric tests for comparing samples of possibly censored survival times. These tests are *conditional permutation tests* in which conditioning is on the individuals at risk at each distinct survival time. The asymptotic validity of the conditional permutation tests in the presence of unequal distributions of potential follow-up times of the samples (“unequal follow-up” in the sequence) has contributed to their popularity.

Complete permutation tests for censored data such as the early generalized Wilcoxon test by Gehan (1965) or the Savage (1956) test modified for censored data (Schemper, 1984) do not enjoy this property, as they assume that survival and censoring times can be simultaneously permuted under the null hypothesis of identical survival distributions for the samples to be compared. These tests therefore require equal follow-up which, as Mantel (1985) has reemphasized, rarely can be assumed. Attempting to satisfy this requirement with complete permutation tests, Jennrich (1984) introduced an artificial mechanism to equalize censorship between the groups, but for the price of a possible loss of power. Furthermore, the idea of conditioning on the follow-up occurring in a study of survival, as adopted by the conditional permutation tests or

by Cox’s (1972) model, is closer in spirit to the survival study actually conducted.

Nevertheless, the preferred asymptotic versions of conditional permutation tests suffer from one shortcoming: for small and for unbalanced sample sizes, they can be substantially anticonservative (Latta, 1981; Kellerer and Chmelevsky, 1983). Obviously, this shortcoming of the conditional permutation tests is due to marked skewness and discreteness of the distribution of the test statistic in small and unbalanced samples. Therefore, exact versions of these tests would be useful, but have not been available. Only for some complete permutation tests for censored data, with their known limitations, have exact versions been developed and implemented in StatXact (Mehta and Patel, 2000), a software package focused on exact inference.

In short, while asymptotic conditional and unconditional tests are available, as well as exact unconditional ones, exact conditional tests are missing. It is the purpose of this article to present *exact conditional tests* for possibly censored survival times.

Recently, we compared the survival of breast cancer patients who had primary treatment at the Department of Surgery of the University Hospital in Vienna between 1982 and 2001, and had either been enrolled in clinical trials (“trial” group) or not (“nontrial” group). It was assumed that trial patients have better prognosis than nontrial

patients. All patients in our study were premenopausal, node negative, and had tumor sizes ≤ 1 cm and tumor gradings GX, G1, or G2. The group sizes were 38 (all of them censored) for the trial group and 90 (80 of them censored) for the nontrial group. Clearly, as we compare patient data from a prospective study to patient data collected retrospectively, follow-up must be assumed to differ. This is reflected in the medians (quartiles) of follow-up time of 9.5 (5.8, 24.3) and 79.1 (56.0, 98.7) months for the trial and the nontrial groups, respectively. Because of the low number of events and the unbalanced group sizes, an asymptotic test may not be appropriate. However, an exact complete permutation test may be inappropriate as well, due to substantial differences in follow-up. In this situation, exact conditional tests would be required.

Such tests will be presented in the following section. In Section 3, the empirical performance of these procedures is explored and compared with the asymptotic log-rank test and the exact complete permutation test by simulation. We revisit the breast cancer study in Section 4 and close with a brief discussion.

2. Methods

The presentation of methods is confined to linear rank tests that can accommodate unequal follow-up of groups of individuals. First, we review the most popular of these, the log-rank and related tests, and then introduce their most natural exact analogues. These approaches condition on the observed sets of individuals at risk of death at observed death times. Second, we introduce another, exact approach that directly conditions on the realized follow-up of groups.

We assume a sample of n independent individuals allocated to $r + 1$ groups. Let $j(i) \in \{0, \dots, r\}$, x_i and y_i , $i = 1, \dots, n$, denote group indicator, death time and potential follow-up time of individual i , respectively. Assume that x_i and y_i are independent, and that their distribution functions are $T_{j(i)}$ and $F_{j(i)}$, respectively. We are interested in testing the null hypothesis $H_0 : T_j = T_{j'}$ without requiring that $F_j = F_{j'}$ for all $j, j' \in \{0, \dots, r\}$.

2.1 *Permutation Tests Conditioning on Observed Risk Sets*

2.1.1 *Asymptotic log-rank and related tests.* Following the exposition by Tarone and Ware (1977), we observe survival time t_i and censoring indicator δ_i , where $t_i = \min(x_i, y_i)$ and $\delta_i = 1$ if $x_i \leq y_i$, and 0 otherwise. With each time point $t_{(h)}$ at which a death was observed, we can associate data as in Table 1. Here, N_{jh} is the number of individuals at risk in the j th group at the h th time point, M_{jh} is the number of events in the j th group at the h th time point, R_h is the total number at risk at $t_{(h)}$, and $A_{jh} = N_{jh}/R_h$ ($j = 0, 1, \dots, r; h = 1, \dots, q$).

Table 1

Summary of observations at the h th death time point $t_{(h)}$

Group	0	1	...	r	Total
Number of events	M_{0h}	M_{1h}	...	M_{rh}	M_h
Number at risk	N_{0h}	N_{1h}	...	N_{rh}	R_h

Let

$$S_j = \sum_{h=1}^q w_h M_{jh}, \quad \text{and} \quad V_{jk} = \sum_{h=1}^q w_h^2 \alpha_i A_{jh} (d_{jk} - A_{kh}),$$

where w_h denotes some weight explained below, $\alpha_h = M_h(R_h - M_h)/(R_h - 1)$ and d_{jk} is defined to be 1 if $j = k$ and 0 otherwise. The vector S is defined to be $(S_1, \dots, S_r)'$ and V is defined to be the $r \times r$ matrix with (j, k) th element V_{jk} . The expected value of S_j under the null hypothesis of equal survival distributions is

$$E_j = \mathbf{E}(S_j) = \sum_{h=1}^q w_h M_h A_{jh}$$

Define $E = (E_1, \dots, E_r)'$. Then, under the null hypothesis of no group effect, the statistic

$$(S - E)'V^{-1}(S - E)$$

has been shown to asymptotically follow a χ^2 -distribution with r degrees of freedom. If $w_h = 1$, $h = 1, \dots, q$, we obtain the log rank test (Mantel, 1966) which has optimal power if the hazard functions of the groups are proportional (Peto and Peto, 1972). The choice of $w_h = R_h$, $h = 1, \dots, q$, leads to Breslow's (1970) generalized Wilcoxon test, while the choice of the survivor-function estimator as weight factor w_h is needed for Prentice's (1978) generalized Wilcoxon test.

2.1.2 *Exact analogues of the log-rank and related tests.* Instead of relying on asymptotics, one could compute the exact distribution of the score S over the risk set tables associated with the q death times. In the following, we outline this approach in notation similar to Mehta and Patel (2000).

Define Ω as the set of all possible values s that S can assume, conditional on N_{jh} and M_h ($j = 0, \dots, r, h = 1, \dots, q$) observed in a sample:

$$\Omega = \{s \mid M_h, N_{jh}\}$$

Define M to be the matrix with the (j, h) th element being M_{jh} . Each element $s = (s_1, \dots, s_r)'$ of Ω can be the result of $c(s)$ different values m that M can assume without violating the condition $\sum_{h=1}^q w_h m_{jh} = s_j, j = 1, \dots, r$. Let $\Psi(s) = \{m : \sum_{h=1}^q w_h m_{jh} = s_j\}$ denote the set of those values. The probability of observing a particular value s under the null hypothesis is equal to the number of ways s can be achieved over the number of all possible outcomes of M :

$$\Pr(S = s) = \frac{c(s)}{\sum_{u \in \Omega} c(u)}$$

with

$$c(s) = \sum_{m \in \Psi(s)} \prod_{h=1}^q \left\{ \prod_{j=0}^r \binom{N_{jh}}{m_{jh}} \right\}$$

An efficient network algorithm, such as that of Mehta, Patel, and Gray (1985) can be employed to find the sets $\Psi(s)$ and to compute the counts $c(s)$.

Finally, one-sided P -values can be obtained as the proportion of values of $S \in \Omega$, which are more extreme than or equal to the statistic's observed value in the original sample.

2.2 Permutation Tests Conditioning on Observed Follow-Up

The tests presented in Section 2.1 condition on the risk sets observed in a sample. Alternatively, different follow-up can be accounted for in a more direct way. In this latter approach, all survival times are permuted, but original follow-up times (defined below in step 1) are retained. Permuted survival times and original follow-up times are then compared. If in a permuted data set, follow-up is shorter than survival, then the survival time is censored at the follow-up time. Otherwise, this survival time remains unchanged. For each of the data sets generated in this way, the statistic $S_E = S - E$ is computed. P -values are obtained by comparing the value of S_E observed in a sample to the distribution of S_E values of the generated data sets.

More formally, the procedure is carried out in the following six steps:

1. From survival time t_i and survival status δ_i , derive a follow-up time f_i and follow-up status ϵ_i ($i = 1, \dots, n$):

$$f_i = t_i, \quad \epsilon_i = 1 - \delta_i$$

2. From (t_i, δ_i) and (f_i, ϵ_i) , estimate empirical cumulative distribution functions for death time x and potential follow-up time y , $\hat{T}(x)$ and $\hat{F}_j(y)$, $j = 0, \dots, r$, respectively, using the Kaplan-Meier (Kaplan and Meier, 1958) method. Let $\hat{T}^{-1}(p)$ and $\hat{F}_j^{-1}(p)$ denote the respective inverse functions, i.e., the p -quantile of the estimated death time or potential follow-up time distribution, respectively. Note that distributions of potential follow-up time $\hat{F}_j(y)$ are estimated separately for each group of individuals.
3. Generate a random permutation (t_i^*, δ_i^*) of the data pairs (t_i, δ_i) by permuting the index i .
4. The comparison of survival times and follow-up times (step 5) requires replacement of censored survival times t_i^* with imputations from $\hat{T}(x | x > t_i^*)$, and replacement of censored follow-up times f_i with imputations from $\hat{F}_{j(i)}(y | y > f_i)$. Denote the random numbers drawn from the uniform distributions $U(\hat{T}(t_i^*), 1)$ and $U(\hat{F}_{j(i)}(f_i), 1)$ by u_i and v_i , respectively, where $j(i)$ is the group indicator for observation i . Furthermore, let t_{\max} denote the overall largest survival time and $f_{j_{\max}}$ the largest follow-up time in group G_j , with $j = 0, \dots, r$. Then obtain estimates of the partly unobservable death times x_i^* as follows:

$$\text{If } \delta_i^* = 0, \text{ let } (\hat{x}_i^*, c_i^*) = \begin{cases} (\hat{T}^{-1}(u_i), 1) & \text{if } u_i \leq \hat{T}(t_{\max}) \\ (t_{\max}, 0) & \text{otherwise} \end{cases}$$

$$\text{otherwise, let } (\hat{x}_i^*, c_i^*) = (t_i^*, 1)$$

Furthermore, impute follow-up times:

$$\text{if } \epsilon_i = 0, \text{ let } \hat{y}_i = \begin{cases} \hat{F}_{j(i)}^{-1}(v_i) & \text{if } v_i \leq \hat{F}_{j(i)}(f_{j_{\max}}) \\ t_{\max} & \text{otherwise} \end{cases}$$

$$\text{otherwise, let } \hat{y}_i = f_i$$

5. Obtain new survival times (t_i^P, δ_i^P) by comparing \hat{x}_i^* to \hat{y}_i :

$$(t_i^P, \delta_i^P) = \begin{cases} (\hat{x}_i^*, 1) & \text{if } \hat{x}_i^* < \hat{y}_i \vee (\hat{x}_i^* = \hat{y}_i \wedge c_i^* = 1) \\ (\hat{y}_i, 0) & \text{if } \hat{y}_i < \hat{x}_i^* \vee (\hat{x}_i^* = \hat{y}_i \wedge c_i^* = 0) \end{cases}$$

6. From the permuted data set (t_i^P, δ_i^P) , compute and store the centered log-rank statistic $S_E = S - E$.

Repeat steps 3–6, say, 1000 times, to obtain an empirical permutational null distribution. Finally, one-sided P -values are obtained by determining the proportion of values of S_E from the empirical permutational null distribution that are more extreme than, or equal to, the value of S_E observed in the original data set.

If, in step 4, the longest follow-up time in group $G_{j(i)}$ is censored, imputation of an even longer follow-up time may be necessary. This is the case if the random number v_i exceeds $\hat{F}_{j(i)}(f_{j(i)\max})$. However, since $f_{j(i)\max}$ is censored, there is no empirical basis for the potential follow-up distribution beyond this point of time. In such a case, the maximum follow-up time of the combined sample $\max_j f_{j_{\max}} = t_{\max}$ is used as imputation. Although seeming arbitrary at first sight, this choice provides equivalence of the presented approach and the exact complete permutation test in the case of no censoring of survival times. Other choices for the imputed value in this case would require *a priori* information about the F_j 's. With decreasing proportions of censored survival times, the impact of any assumption of follow-up decreases as well.

Our approach requires computation of empirical distributions of the unobservable death times x_i and the unobservable potential follow-up times y_i . Because of censoring, a Kaplan-Meier (Kaplan and Meier, 1958) estimate must be employed, but no assumptions beyond independence of individuals, plus independence of survival and potential follow-up are needed. With heavy censoring of either x_i or y_i , their empirical distributions may only be poorly reproduced with the Kaplan-Meier estimates, which occasionally could result in violations of the nominal size or in loss of power. Therefore, our test was subjected to an extensive empirical study presented in the next section.

The Kaplan-Meier estimate is a step function and imputations from it are discrete. Consequently, the distribution of the scores S_E under the null hypothesis could be completely enumerated by generating all permutations of survival times and, within each permutation, obtaining the distribution of S_E for all possible imputations. Even with both fast computers and efficient algorithms, this exact procedure would be infeasible, and is abandoned in favor of its proposed Monte Carlo approximation. Because this approximation can be made virtually indistinguishable from the exact results, we shall designate this approach as “exact test conditioning on observed follow-up” in the sequence.

The randomness that is introduced by imputation could be made arbitrarily small by increasing the number of imputations per permutation. However, the imputation will only take effect if the smaller of the two times compared is censored, which will only be the case for a part of the data set. Therefore, it seems more efficient, in terms of reducing Monte Carlo sampling error, to increase the number of permutations than to increase the number of imputations per permuted data set.

3. An Empirical Study

To explore the empirical size and power of the new exact conditional tests under various conditions, a Monte Carlo study

was performed. In this study, we compared size and power of the asymptotic and exact log rank tests, conditioning on observed risk set tables as described in Section 2.1 (abbreviated by “ACR” and “ECR,” respectively), the exact log rank test conditioning on observed follow-up (“ECF,” Section 2.2), the exact complete permutation test of log rank scores (“ECP”) as implemented in StatXact (Mehta and Patel, 2000), and the exact complete permutation test after applying artificial censoring (“EAC,” proposed by Jennrich (1984)), to equalize follow-up distributions. P -values obtained by ECR may be conservative because of the discreteness of the resulting exact distribution of the log rank score S . For comparative purposes, we removed this conservatism, following Lancaster (1961), by randomizing the declaration of significance if $\Pr(S > s) < \alpha < \Pr(S \geq s)$, where s denotes the value of the log rank statistic in a data set and α is the nominal significance level. Exact P -values of ECP, EAC, and ECF were approximated by 1000 random permutations for each simulated data set.

The simulation study was primarily designed to explore the effects of the sample sizes of groups G_0 and G_1 to be compared, (n_0, n_1) , and of their follow-up distributions, F_0 and F_1 , on size and power of one-sided tests. Investigations for each cell of the resulting factorial design were based on 50,000 generated samples.

We shall present results for sample sizes (6,6), (6,30), (30,30), and (3,120). For the generation of follow-up times, a fixed accrual period of 48 time units and constant accrual rate were assumed, plus an additional observation period of 12 time units. The resulting administrative follow-up times are uniformly distributed over the interval [12,60]. To investigate the effect of different loss on follow-up in the groups, we assumed that time until loss to follow-up follows an exponential distribution with hazard rates γ_0 and γ_1 . Results are presented for the following combinations of loss hazard

rates for groups G_0 and G_1 , $(\gamma_0, \gamma_1): (0, 0), (0, 0.04), (0.04, 0), (0.04, 0.04)$.

Survival times are sampled from an exponential distribution, with death hazard rates (λ_0, λ_1) of (0.04, 0.04) for assessing size, and of (0.08, 0.04) or (0.04, 0.08) for assessing power, to detect the one-sided alternatives of shorter or longer survival of group G_0 , respectively. A survival time was censored if it exceeded either the corresponding administrative follow-up time or the time till loss to follow-up.

The performance of the investigated procedures can be understood from the results presented in Tables 2 and 3. The high precision of the entries in these tables (for a true size of 0.05, the simulation standard error is 0.00097) permits detection of very small departures from nominal significance levels.

First we learn that ECR cannot remedy the known anti-conservative behavior of ACR in small samples, and second, that the empirical size of the ECP can substantially depart from nominal size under unequal follow-up, i.e., with unequal loss hazard rates.

The unsatisfactory behavior of ECR in very small, unbalanced data sets results from conditioning the permutational null distribution on observed risk tables. In a particular sample, the composition of a risk table at $t_{(h)}$ (Table 1) depends to a high degree on the group membership of the individuals dying up to $t_{(h-1)}$. Under the null hypothesis, all realizations of the statistic are equiprobable, but not all sequences of risk tables possible. Thus, it can be shown that the (randomized) P -values by ECR are not uniformly distributed (Heinze, 2002, pp. 38–39), and hence ECR does not keep the nominal size at all significance levels. Censoring between death times reduces the dependency of risk tables on the sequence of previously deceased individuals, which is also reflected in our simulation results.

A systematic censoring mechanism is used in the EAC approach of Jennrich (1984). Here, each time an individual

Table 2

Results on the size ($\times 1000$) of log-rank tests of equal survival against shorter/longer survival of group G_0 versus G_1 ($\alpha = 0.05$, one-sided)

n_0	n_1	γ_0	γ_1	% c_0	% c_1	ACR	ECR	ECP	EAC	ECF
Equal follow-up ($F_0 = F_1$)										
6	6	0.00	0.00	27.5	27.6	56/57	51/52	49/52	71/75	48/50
6	6	0.04	0.04	54.5	54.7	57/55	51/50	51/49	51/50	53/53
30	30	0.00	0.00	27.5	27.5	50/53	49/51	49/50	41/41	46/48
30	30	0.04	0.04	54.9	54.8	49/52	47/50	47/50	40/44	47/50
6	30	0.00	0.00	27.5	27.5	78/35	63/43	49/48	2/0	45/47
6	30	0.04	0.04	54.9	54.8	77/34	60/45	49/50	12/0	48/50
3	120	0.00	0.00	27.3	27.5	113/25	84/39	50/49	0/0	50/43
3	120	0.04	0.04	54.6	54.9	109/13	76/48	50/50	0/0	50/31
Unequal follow-up ($F_0 \neq F_1$)										
6	6	0.00	0.04	27.4	54.9	47/63	47/54	48/39	41/57	48/51
30	30	0.00	0.04	27.5	54.8	45/55	46/52	45/38	31/32	46/47
6	30	0.00	0.04	27.5	54.8	71/40	58/46	75/68	18/0	47/51
6	30	0.04	0.00	54.9	27.5	81/30	63/43	23/25	1/0	47/43
3	120	0.00	0.04	27.3	54.9	110/27	82/41	100/94	0/0	50/51
3	120	0.04	0.00	54.6	27.5	110/12	77/48	28/16	0/0	46/19

Note: ACR, asymptotic conditional on risk sets; ECR, exact conditional on risk sets; ECP, exact complete permutation; EAC, exact artificial censoring; ECF, exact conditional on follow-up; n_j , group size of group G_j ; γ_j , loss hazard for G_j ; F_j , distribution of potential follow-up times for G_j ; % c_j , percentage censored for G_j ; results based on 50,000 replications.

Table 3
 Results on the power ($\times 1000$) of log-rank tests of equal survival against shorter/longer survival of group G_0 versus G_1 ($\alpha = 0.05$, one-sided)

n_0	n_1	γ_0	γ_1	ACR	ECR	ECP	EAC	ECF
Equal follow-up ($F_0 = F_1$)								
6	6	0.00	0.00	280/277	257/255	254/252	299/304	248/247
6	6	0.04	0.04	213/210	190/188	197/196	172/170	198/198
30	30	0.00	0.00	772/773	765/767	762/765	477/483	758/759
30	30	0.04	0.04	615/616	605/607	605/608	398/403	603/605
6	30	0.00	0.00	459/348	405/369	355/397	17/0	336/396
6	30	0.04	0.04	366/242	311/267	282/291	57/0	278/292
3	120	0.00	0.00	407/221	324/262	223/309	0/0	222/306
3	120	0.04	0.04	342/137	260/188	196/223	2/0	193/204
Unequal follow-up ($F_0 \neq F_1$)								
6	6	0.00	0.04	227/244	214/215	233/181	170/221	223/211
30	30	0.00	0.04	674/682	672/669	675/621	360/392	676/654
6	30	0.00	0.04	418/317	371/333	426/392	76/0	329/351
6	30	0.04	0.00	386/253	327/284	188/251	14/0	277/313
3	120	0.00	0.04	399/213	319/253	370/377	2/0	223/298
3	120	0.04	0.00	346/137	263/191	128/150	0/0	177/198

Note: ACR, asymptotic conditional on risk sets; ECR, exact conditional on risk sets; ECP, exact complete permutation; EAC, exact artificial censoring; ECF, exact conditional on follow-up; n_j , group size of group G_j ; F_j , distribution of potential follow-up times for G_j ; γ_j , loss hazard for G_j ; results based on 50,000 replications.

of group j dies, an individual randomly chosen from group $j' \neq j$ is artificially censored. As a consequence, EAC is very conservative when group sizes are not balanced, as shown by our simulation results, which agree with the asymptotic results of Jennrich (1984).

In case of equal follow-up, i.e., with equal loss hazard rates, ECF and ECP hold the nominal size, and the relative power achieved by ECF compared to ECP is about 95%–100%, independent of sample size; it drops to 90% only in the very extreme investigated situation of group sizes of 3 and 120. ECP’s empirical size exceeds nominal size by a factor of 1.3–2 in situations of unbalanced group sizes, with the larger group having shorter follow-up. In such situations, as well as all the other situations of unequal follow-up investigated, ECF holds the nominal size. For unequal follow-up and balanced group sizes, or if the smaller of the groups has shorter follow-up, ECP is mostly outperformed by ECF, reflected by a relative power of ECP compared to ECF ranging from 50% to 100%. Substantial violations of the size by ACR and ECR render their occasional advantages in terms of power (up to 1.8-fold compared to ECP and ECF) meaningless. All these conclusions follow from Tables 2 and 3, and from a technical report (Heinze, 2002), which extends the presented results to significance levels of 0.01 and 0.1, two-sided comparisons, and to situations of unequal accrual and observation periods.

Summarizing our empirical results, both ECP and ECF are considered reasonable choices for equal follow-up, but only ECF can be recommended under the condition of unequal follow-up.

4. Breast Cancer Study Revisited

We now return to the example introduced in Section 1. We applied the asymptotic log rank test conditioning on observed risk sets (ACR), as implemented in SAS/PROC LIFETEST (SAS Institute, 1999), and its exact analogue (ECR), as pre-

sented in Section 2.1.2, the exact complete permutation test (ECP) as implemented in StatXact (Mehta and Patel, 2000), the exact complete permutation test after artificial censoring (EAC), as suggested by Jennrich (1984), and the permutation test conditioning on observed follow-up (ECF), as introduced in Section 2.2. While the ACR, the ECR, and the ECF take the differing follow-up distributions of the two groups into account, the ECP does not. The EAC artificially equalizes the follow-up distributions before permuting the group labels among all individuals. Application of an asymptotic test may be questionable, with the low number of events and the unbalanced group sizes.

Results in terms of one-sided P -values are presented in Table 4. We learn that, while ACR is marginally significant at the level of 0.05, ECP and EAC fail to detect a significant difference. Because of the highly discrete distribution of the log rank score S when conditioning on the observed risk sets,

Table 4
 Comparison of breast cancer patients enrolled versus not enrolled in clinical trials

Type of log-rank test	Standard software used	P -value
Asymptotic, conditional on risk sets	SAS/PROC LIFETEST	0.049
Exact, conditional on risk sets		0.045 ^a
Exact, complete permutation	StatXact	0.098 ^b
Exact, artificial censoring		0.333 ^b
Exact, conditional on follow-up		0.031 ^b

^a mid- P -value; conservative P -value is 0.090.

^b P -value based on 10,000 permutations.

ECR produces a range of P -values between $\Pr(S > s) = 0$ and $\Pr(S \geq s) = 0.09$, with the mid- P -value (Lancaster, 1961) at $\Pr(S > s) + \Pr(S = s)/2 = 0.045$. The ECF arrives at a P -value of 0.031, which lets us conclude that breast cancer patients enrolled in a clinical study experience longer survival.

5. Concluding Remarks

Previous developments in the field of nonparametric tests for censored data focused on asymptotic and exact complete permutation tests and on asymptotic conditional permutation tests. Failure of the latter tests in highly unbalanced samples, in particular, of the widely used log-rank test, has already been diagnosed in the early 1980s, without, however, arriving at a solution. We have shown that the failure of the asymptotic log-rank test is not only due to an inappropriate asymptotic approximation, which could be replaced by an exact evaluation, but is also due to ignoring the interdependency of the observed risk sets.

In contrast to asymptotic conditional tests, exact conditional tests have been a neglected area of research. We have developed and presented such a test and have demonstrated that, under unequal follow-up, it currently is the best choice. However, in situations of equal follow-up, it is slightly outperformed by the exact complete permutation test. An analogous result is known for the comparison of two samples of normally distributed outcomes with unknown variances. While Student's t -test is optimal for equal variances, use of Satterthwaite's (1946) approximation is preferable under heteroscedasticity.

By way of conclusion, we therefore recommend use of the newly presented permutation test conditional on follow-up whenever the assumption of equal follow-up is in doubt. An SAS (SAS Institute, 1999) macro implementing all presented exact tests is available from the authors.

ACKNOWLEDGEMENTS

We acknowledge helpful comments by Terry Smith, M.D. Anderson Cancer Center, University of Texas, Houston, and Alexandra Kaider, Vienna University, on an earlier version of our article. We also thank the editor, an associate editor, and a referee for their suggestions which considerably improved this article.

RÉSUMÉ

Les tests asymptotiques du log-rank et le test de Wilcoxon généralisé sont des procédures standards pour comparer des distributions de survie en présence de censures. Pour la comparaison d'échantillons de tailles très différentes, un test exact basé sur l'ensemble des permutations du log-rank ou des scores de Wilcoxon est disponible. Alors que les tests asymptotiques ne conservent pas leur risque nominal lorsque les échantillons sont de tailles très différentes, le test exact basé sur les permutations exige cependant un suivi identique dans les échantillons. C'est pourquoi nous avons développé deux nouveaux tests exacts adaptés à des suivis inégaux. Le premier est exactement équivalent au log-rank test asymptotique,

conditionnellement à l'ensemble des risques observés, alors que la deuxième approche permute les temps de survie conditionnellement au suivi observé dans chaque groupe. Dans une étude empirique, nous comparons ces nouvelles procédures au log-rank test asymptotique, au test exact de permutation complète, ainsi qu'à une approche proposée antérieurement qui égalise les distributions de suivi en les censurant artificiellement. Les résultats confirment les performances très satisfaisantes du test exact conditionnellement au suivi observé, particulièrement dans le cas de suivis inégaux. L'avantage du nouveau test sur les autres possibilités d'analyse est finalement illustré par l'analyse d'une étude portant sur le cancer du sein.

REFERENCES

- Breslow, N. E. (1970). A generalized Kruskal-Wallis test for comparing K -samples subject to unequal patterns of censoring. *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.
- Heinze, G. (2002). *Exact linear rank tests for possibly heterogeneous follow-up*. Technical Report 08/2002, University of Vienna, Department of Medical Computer Sciences, Section of Clinical Biometrics.
- Jennrich, R. I. (1984). Some exact tests for comparing survival curves in the presence of unequal right censoring. *Biometrika* **71**, 57–64.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.
- Kellerer, A. M. and Chmelevsky, D. (1983). Small-sample properties of censored-data rank tests. *Biometrics* **39**, 675–682.
- Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association* **56**, 223–234.
- 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.
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**, 163–170.
- Mantel, N. (1985). Propriety of the Mantel-Haenszel variance for the log rank test. *Biometrika* **72**, 471–472.
- Mehta, C. and Patel, N. (2000). *StatXact4 for Windows User Manual*. Cambridge, Massachusetts: Cytel Software.
- Mehta, C., Patel, N., and Gray, R. (1985). Computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables. *Journal of the American Statistical Association* **80**, 969–973.
- Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant procedures (with discussion). *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.

- SAS Institute (1999). *SAS/STAT User's Guide, Version 8*. Cary, North Carolina: SAS Institute.
- Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* **2**, 110–114.
- Savage, I. R. (1956). Contributions to the theory of rank order statistics—The two sample case. *Annals of Mathematical Statistics* **27**, 590–615.
- Schemper, M. (1984). A survey of permutation tests for censored survival data. *Communications in Statistics—Theory and Methods* **13**, 1655–1665.
- Tarone, R. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156–160.

Received September 2002. Revised March 2003.

Accepted May 2003.