BIOMETRICS 43, 169–180
March 1987

A Paired Prentice–Wilcoxon Test for Censored Paired Data

Peter C. O'Brien

Medical Research Statistics, Mayo Clinic, Rochester, Minnesota 55905, U.S.A.

and

Thomas R. Fleming

Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.

SUMMARY

Prentice (1978, *Biometrika* **65,** 167–179) has provided a censored-data generalization of classical linear rank statistics. This paper investigates a method to generalize the Prentice statistics to the analysis of censored paired data. Particular emphasis is given to the paired-data version of the Prentice–Wilcoxon statistic, with discussion of the reasons for preferring Prentice over the Gehan scores. The paired Prentice–Wilcoxon (PPW) statistic makes use of interblock information and can be viewed as a censored-data generalization of the Conover–Iman (1981, *The American Statistician* **35,** 124–129) "paired *t* test on the ranks" or of the Lam–Longnecker (1983, *Biometrika* **70,** 510–513) modified Wilcoxon rank sum statistic. In simulations comparing the PPW, sign, and generalized signed rank (GSR) statistics, the PPW is most powerful against all but the exponential scale alternative. Furthermore, the small advantage of the GSR over the PPW in that alternative is lost if outlier pairs are introduced. When compared to the two-sample Prentice–Wilcoxon statistic, the PPW is nearly as powerful in uncorrelated data and its power becomes superior as correlation between paired members increases.

1. Introduction

The paired *t* test is the commonly used procedure for testing shift in location with matched-pair (x, y) data in which the measurement of interest is continuous and uncensored. The test statistic may be represented as

$$t = \sqrt{n}(\bar{x} - \bar{y})/(s_x^2 + s_y^2 - 2s_{xy})^{1/2}. \qquad (1.1)$$

Thus, the paired *t* test may be viewed as a generalization of the two-sample *t* test. To the extent that x and y are uncorrelated (so that $s_{xy} \approx 0$), the test statistic is the same as the two-sample *t* test making full use of interblock information. An alternative method for using interblock information is to replace the observed data with their ranks in the pooled sample and perform a paired *t* test on the ranks (Conover and Iman, 1981). When ranks are employed, the numerator of (1.1) is identical to that of the two-sample Wilcoxon statistic for independent samples, and hence to the numerator of the paired-data test of Lam and Longnecker (1983). The latter authors then use a variance estimator that is consistent whether or not the paired observations are correlated. Lam and Longnecker have shown that their paired Wilcoxon procedure has power comparable to a conventional paired *t* test when sampling from a bivariate normal distribution. It was also found to be more powerful than the sign and signed rank procedures when testing for shift in location, particularly when the underlying distribution is asymmetric.

Key words: Censored data; Paired data; Paired *t* test; Sign test; Signed rank test; Wilcoxon test.

An example can illustrate the different manners in which these paired-data tests use interblock information. The data in Table 1 represent survival times for skin grafts on 11 patients, each of whom received both a closely and a poorly matched graft. The data are taken from Woolson and Lachenbruch (1980) and are slightly modified from those originally reported by Batchelor and Hackett (1970). The data have been analyzed by Holt and Prentice (1974) and by Kalbfleisch and Prentice (1980, pp. 190–194) using the sign test, derived using a logistic model. For present purposes, we ignore the censoring in subjects 3 and 11 and treat all observations as observed survival times.

Clearly, a conventional paired t test would be inappropriate for analysis of such data. With the sign test, signed rank test, and paired t test on the ranks, we obtain $P = .033$, $P = .007$, and $P = .010$, respectively (all tests one-sided). These results suggest that when the largest observations are all in the same direction (thereby supporting the hypothesis of a treatment effect), the paired t test on the ranks and the signed rank test provide essentially the same answer, and the sign test is relatively insensitive. However, the situation may be quite different when one or more of the largest observations are in the opposite direction.

Suppose that the poorly matched graft on subject 4 had survived 126 days. Notice that the corresponding pair of values on this subject are both outliers, and that the difference is large and in the opposite direction of the other largest differences. This is the type of data that may occur in practice if the subject (rather than the individual graft) is the cause of the outlier, with the causal mechanism unrelated to the treatment effect. Now, with the sign test and signed rank test, $P = .113$ and $P = .074$, respectively. However, the effect on the paired t test on the ranks is less pronounced: $P = .023$.

In contrast to the paired t and signed rank tests, the paired t test on the ranks uses interblock information solely through a rank metric, i.e., the distance between x and y for any given pair is determined by the number of other values falling between them. As a result, the paired t test on the ranks will be less sensitive to the type of outlier pair seen in the previous example. As a further consequence of the use of the rank metric, this test procedure will approach the sign test when the correlation between x and y and the ratio of interpair variability to intrapair variability increase, a situation in which there is absence of useful interblock information. Conversely, in uncorrelated data, the test criterion appropriately will make full use of interblock information.

Although nonparametric alternatives to the paired t test usually arise within the context of studying asymmetric or heavy-tailed distributions, the paired t test on the ranks with its rank metric might be advantageous even in the absence of such departures from normality.

Table 1
Observed survival times to skin grafts on burn patients

Patient number	Close match (x_j)	Poor match (y_j)	$x_j - y_j$	Ranked data		Δ
				$R(x_j)$	$R(y_j)$	
1	37	29	+8	16	14.5	+1.5
2	19	13	+6	8	2	+6
3	57[a]	15	+42	19	3.5	+15.5
4	93	26	+67	22	12.5	9.5
5	16	11	+5	5	1	4
6	22	17	+5	11	6	5
7	20	26	−6	9	12.5	−3.5
8	18	21	−3	7	10	−3
9	63	43	+20	21	18	3
10	29	15	+14	14.5	3.5	11
11	60[a]	40	+20	20	17	3

[a] Indicates the survival time was censored at the time indicated. Computations in this table ignore this censoring.

Specifically, the conventional paired t and signed rank tests treat pair differences the same, regardless of where the individual x and y values occurred in their respective distributions. This would be appropriate if, for example, the variance of the conditional distribution of y given x is the same at all levels of x. In practice, however, this variance depends on x in ways that are difficult to anticipate, particularly when x is very large or very small. For example, changes in blood pressure may be more highly variable among persons with very high or very low pressures initially. The paired t test on the ranks would avoid overemphasizing these variable changes occurring in people having extreme initial pressures.

The previous discussion suggests that a generalization of the Conover–Iman "paired-t-on-the-ranks" procedure would be very useful in the analysis of survival data subject to arbitrary right censoring. Such a procedure was proposed by Wei (1980), and was shown to belong to a larger class of procedures by Cheng (1984). In the next section, we will propose a class of censored paired-data statistics by indicating how classical linear rank statistics, generalized to censored data by Prentice (1978), in turn can be generalized to paired censored data. One important member of this new class will be the paired Prentice–Wilcoxon (PPW) statistic. In Section 3, we will illustrate why the PPW statistic provides an important alternative to Wei's statistic. The latter is shown to be essentially a paired version of the Gehan–Wilcoxon (Gehan, 1965) statistic. In Section 4, we provide Monte Carlo simulations to compare the PPW statistic with the sign test and the censored-data signed rank test of Woolson and Lachenbruch (1980).

2. A Class of Rank Statistics for Censored Paired Data

Prentice (1978) showed how the efficient score function, used to generate locally most powerful and fully efficient rank tests, could be adapted to formulate censored-data versions of these tests. Recent work of Cuzick (1982) and Struthers (unpublished Ph.D. dissertation, University of Waterloo, Waterloo, Ontario, 1984) has revealed that the Prentice censored-data rank statistics do maintain the efficiency properties in the presence of censoring, as long as certain regularity conditions are satisfied. In this section, we will discuss one approach to obtaining censored paired-data versions of these Prentice linear rank statistics. These versions also will be computationally simple.

Assume we have n pairs of censored survival times. Denote the survival and censoring times associated with each pair by $\{(S_x, C_x), (S_y, C_y)\}$, where the resulting n four-tuples are assumed to be independent and identically distributed, with the survival times independent of the censoring times. Define $C = \min(C_x, C_y)$ to be a common censoring time for both pair members, ignoring any additional follow-up occurring beyond that point. This convention is adopted to eliminate bias that may arise when censoring and survival times are not independent. The types of bias that may arise in practice, and the extent to which this convention is helpful, are discussed in Section 5.

Using these data and Prentice's (1978) adaption of the efficient score function, one can compute scores (k_{xl}, k_{yl}) for the lth pair, $l = 1, \ldots, n$. Then define $\Delta_l = k_{xl} - k_{yl}$, and let $\delta_l = +1$ if $\Delta_l > 0$ and $\delta_l = -1$ if $\Delta_l < 0$. From a probabilistic standpoint, condition on the values of $|\Delta_l|$ and consider the n^* pairs $\{l_1, \ldots, l_{n^*}\}$ such that $|\Delta_{l_j}| > 0, j = 1, \ldots, n^*$. Under this conditioning and under the null hypothesis H_0: $S_x = S_y$, $\{\delta_{l_j}: j = 1, \ldots, n^*\}$ are independent and identically distributed such that $\Pr(\delta_{l_j} = 1) = \frac{1}{2} = \Pr(\delta_{l_j} = -1)$; thus, $\sum_{l=1}^{n} \Delta_l$ has mean zero and variance $\sum_{l=1}^{n} (\Delta_l^2)$. In turn, under our conditioning, it follows from the Lindberg–Feller theorem that if

$$\max_{1 \leq l \leq n} \Delta_l^2 \bigg/ \left(\sum_{l=1}^{n} \Delta_l^2\right) \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty, \tag{2.1}$$

then

$$Z_n \equiv \sum_{l=1}^{n} \Delta_l \Big/ \left(\sum_{l=1}^{n} \Delta_l^2\right)^{1/2} \xrightarrow{D} Z, \tag{2.2}$$

where $\xrightarrow{D}$ denotes convergence in distribution and Z has a standard normal distribution. The proposed test procedure compares the relative deviate Z_n in (2.2) to tables of the standard normal distribution. In a given application, one might compute $\max_{1 \leq l \leq n} \Delta_l^2/(\sum_{l=1}^{n} \Delta_l^2)$ to assess the adequacy of this normal approximation. We note that a permutation test also could be obtained by considering the 2^n equally likely (under H_0) values of the test statistic that are obtained by permuting treatment within pairs for the n pairs.

Consider now the specific form of Z_n for the paired Prentice–Wilcoxon (PPW) statistic. Rank all the observation times from smallest to largest, and let $n(j)$ denote the number of persons in the pooled sample with observation times greater than or equal to the jth distinct ordered observed death time, t_j, $j = 1, \ldots, D$. Assume momentarily that there are no ties in times of observed deaths. To compute the Prentice–Wilcoxon scores [see Kalbfleisch and Prentice (1980, p. 147)], define

$$s_i = \prod_{j=1}^{i} n(j)/[n(j) + 1], \quad i = 1, \ldots, D.$$

The score assigned to the individual having the ith observed death is given by $1 - 2s_i$. Individuals who are censored at the time of the ith death up to the $(i + 1)$st death are assigned a score of $1 - s_i$.

Suppose now that ties exist in times of observed deaths. If m_j deaths are observed at t_j, arbitrarily order these times by assigning them distinct values infinitesimally to the left of t_j. When this is completed for all j, compute the scores as indicated above. Then, each individual with an observed death at t_j is assigned the average of the scores for the corresponding m_j individuals.

To complete the computation of the PPW statistic, one uses these scores to compute Δ_l $(l = 1, \ldots, n)$ and, in turn, the statistic in (2.2). When the data in Table 1 are analyzed, this time taking censoring into account, the paired Prentice–Wilcoxon procedure yields PPW $= 4.524/(4.568)^{1/2} = 2.117$, and $P = .017$.

Suppose one uses log-rank scores to obtain Z_n in (2.2). The resulting "paired log-rank" (PLR) procedure would be similar to that proposed by Mantel and Ciminera (1979). However, by defining $C = \min(C_x, C_y)$ to be the common censoring time for both pair members, undesirable properties inherent in the Mantel–Ciminera procedure [see Michalek and Mihalko (1983)] are avoided.

3. Censored Paired-Data Statistics: Prentice vs Gehan Scores

In this section, we will explore the relative properties of the paired Prentice–Wilcoxon (PPW), paired log-rank (PLR), and paired Gehan–Wilcoxon (PGW) statistics. For the latter statistic, the score for each individual is simply the proportion of the $2n$ participants whose survival times are *known* to be less than that of the individual minus the proportion whose survival times are *known* to be greater. This investigation will help to clarify the relationship between PPW and the statistic proposed by Wei (1980), and will be based on the following important example.

Suppose a study is conducted to compare a "standard" with an "experimental" therapy. A matched-pair design is used with, say, 5000 pairs. Suppose the data are very heavily censored. [Although this example is hypothetical, the heavy censoring described sometimes occurs; e.g., see Massey et al. (1984).] In this hypothetical example, suppose we have one

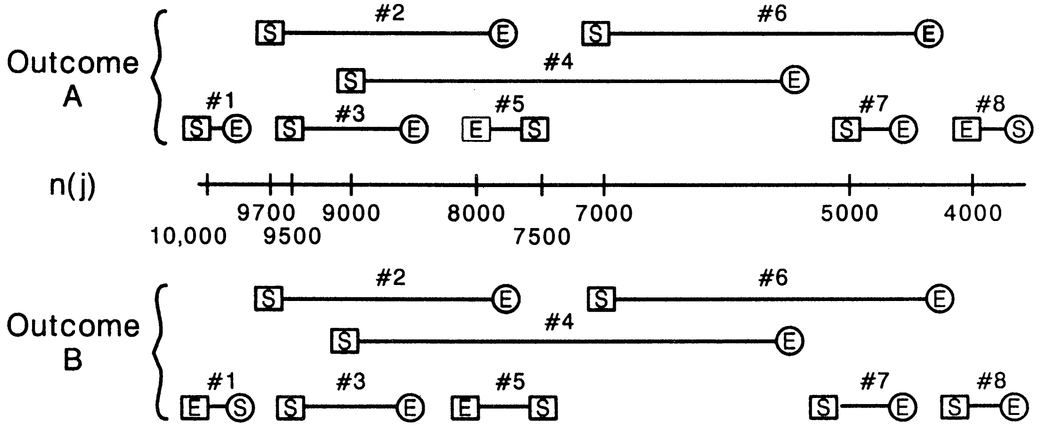

Figure 1. Hypothetical survivorship data from 5000 matched pairs, with one uncensored pair (S_5, E_5) and seven singly-censored pairs $\{(S_i, E_i): i = 1, 2, 3, 4, 6, 7, 8)\}$. Uncensored and censored observations are denoted by squares and circles, respectively.

uncensored pair (S_5, E_5) and seven singly-censored pairs $\{(S_i, E_i): i = 1, 2, 3, 4, 6, 7, 8\}$. Note that, under our convention of using $C = \min(C_x, C_y)$ as the common censoring time for a pair, $\Delta_l = 0$ for each of the 4992 doubly-censored pairs. Figure 1 presents two closely related outcomes (labeled A and B) for the eight non-doubly-censored pairs. In both outcomes, $\hat{S} \approx 1$ at all nine observed "death" times, where $\hat{S}$ is a pooled sample estimator of the survival probability. When such a situation exists in unpaired data, it is well known that the Prentice–Wilcoxon and log-rank statistics are essentially equivalent. We then expect, and Table 2 confirms, that the paired versions of these statistics (PPW and PLR) are also essentially equivalent. In addition, when $\hat{S} \approx 1$, the PPW and PLR statistics essentially reduce to the Mantel–Haenszel statistic for combining 2 × 2 tables for matched pairs [see Fleiss (1981, §10.4)] where for each pair the 2 × 2 table has categories "dead vs censored" and "standard vs experimental." Of course, only pairs "discordant" in "dead vs censored" (i.e., only singly-censored pairs) contribute to the Mantel–Haenszel, which

Table 2
Computation of test statistics for outcomes A and B

	Outcome A				Outcome B			
Pair i	Paired Prentice–Wilcoxon scores Δ_i	Paired log-rank scores Δ_i	"Mantel–Haenszel scores" Δ_i	Paired Gehan–Wilcoxon scores Δ_i	Paired Prentice–Wilcoxon scores Δ_i	Paired log-rank scores Δ_i	"Mantel–Haenszel scores" Δ_i	Paired Gehan–Wilcoxon scores Δ_i
#1	.9999	1.0000	1	1.0000	−.9999	−1.0000	−1	−1.0000
#2	1.0001	1.0003	1	.9703	1.0001	1.0003	1	.9703
#3	.9998	1.0001	1	.9501	.9998	1.0001	1	.9501
#4	1.0000	1.0004	1	.9003	1.0000	1.0004	1	.9003
#5	−.0003	−.0002	0	−.0501	−.0003	−.0002	0	−.0501
#6	.9994	1.0002	1	.7001	.9994	1.0002	1	.7001
#7	.9990	1.0000	1	.5000	.9990	1.0000	1	.5000
#8	−.9988	−1.0000	−1	−.4000	.9988	1.0000	1	.4000
$\Sigma\Delta_i$	4.9991	5.0008	5	4.5710	4.9969	5.0008	5	3.3707
$\Sigma\Delta_i^2$	6.9940	7.0020	7	4.5736	6.9940	7.0020	7	4.5736
Z_n	1.89	1.89	1.89	2.14	1.89	1.89	1.89	1.58
1-tailed P	.029	.029	.029	.016	.029	.029	.029	.057

is exactly what we see happening in Table 2 for the PPW, PLR, and Mantel–Haenszel scores.

Contrasting outcomes A and B displays a fundamental unappealing property of the Gehan (i.e., the PGW) scores. In both outcomes, both pairs #1 and #8 have one individual who dies with the paired individual being censored immediately thereafter. In outcome A, the individual on standard therapy dies in pair #1 and on experimental therapy dies in pair #8. The reverse happens in outcome B. Other than this difference, the two outcomes are identical. Because $\hat{S} \approx 1$ throughout, there is no reason to weight the information from pairs #1 and #8 any differently; thus, the *P*-values should be the same in the two examples. This is true for the PPW, PLR, and Mantel–Haenszel statistics, but certainly fails to be the case for the PGW statistic.

This advantage for Prentice scores over Gehan scores has been discussed frequently in the unpaired situation. Prentice and Marek (1979) present a heuristic discussion. Gill (1980) proves that it is the Prentice–Wilcoxon scores that give rise to a fully efficient censored-data test for detecting time-transformed location alternatives to the logistic distribution.

Wei (1980) proposed a censored paired-data statistic based on Gehan scores. Although the Wei statistic is computationally more complex than the PGW statistic, these two statistics using Gehan scores yield the same results in the previous example. To be specific, using outcome A, the Wei statistic is

$$[(5000)^{1/2}(4.57)/5000]/(4.57/5000)^{1/2};$$

in outcome B, it has value

$$[(5000)^{1/2}(3.37)/5000]/(4.57/5000)^{1/2}.$$

Thus, the Wei statistic shares the same fundamental unappealing property held by the PGW statistic.

In the next section, we compare the small-sample operating characteristics of the newly proposed paired Prentice–Wilcoxon statistic to those of the sign and Woolson and Lachenbruch signed rank procedures.

4. Simulations: PPW vs Sign and Signed Rank Statistics

In the uncensored null hypothesis setting, the paired Prentice–Wilcoxon (PPW) test statistic is asymptotically identical to Lam and Longnecker's (1983) method, since it also uses a consistent estimator of the null variance in the denominator. Thus, the asymptotic Pitman efficiency calculations by Lam and Longnecker may be referred to for an uncensored or lightly censored comparison of the PPW, generalized signed rank, and sign tests. The previously discussed results indicate marked superiority for the PPW test for alternatives involving shift in location.

In terms of the data in Table 1, we observed previously that using the PPW procedure, one obtains $P = .017$. For Woolson and Lachenbruch's (1980) generalized signed rank and sign tests, $P = .008$ and .033, respectively. Using the modified data for this example yields $P = .032$, .058, and .113, respectively, for the PPW, generalized signed rank, and sign tests. It is apparent that the relative efficiencies of the three methods may vary depending on the types of distributions under study. It is therefore of interest to evaluate the operating characteristics in a wide variety of situations.

We performed Monte Carlo simulations to compare the operating characteristics of the paired Prentice–Wilcoxon (PPW), Woolson and Lachenbruch's generalized signed rank (GSR), and the sign (S) tests. The two-sample censored-data Prentice–Wilcoxon test (W) was also simulated to provide an additional basis for comparison. All simulations included 1000 samples of 30 pairs of observation times. Censoring times were generated to achieve moderate (approximately 30%) censoring. The minimum of the realized censoring times in each pair was used to censor both survival times (except for the W procedure), since this is the practice that we would advocate in applications.

Three distributions were used to evaluate the ability of the procedures to control the size of the test: exponential, mixed exponential, and log-logistic. The exponential variates all had unit expectation. In the mixed exponential samples, the first three pairs of observation times were replaced with uncensored mean 10 exponential variates shifted 5 units to the right (thereby introducing outliers). Log-logistic variates were obtained by generating uniformly distributed variables, U, and computing $(U^{-1} - 1)$. Censoring times were independent exponential variates with expectation 4.

To evaluate the procedures with correlated data, the simulations were repeated after adding an exponential variable with unit expectation to each pair member. To maintain moderate censoring, the expectation of the censoring times was increased to 10.

For uncorrelated data, the GSR, W, and PPW procedures provided accurate control over the size of the test, although the Wilcoxon was conservative in the presence of outliers. The sign test was conservative due to the marked discreteness of the small-sample binomial distribution (Table 3). The results with correlated data were similar except that the W test was very conservative, reflecting its tendency to overstate the variance of paired differences. For the situations simulated, it can be concluded that the conditional asymptotic result in (2.2) can be used to obtain accurate approximations to the small-sample unconditional distribution of PPW.

To compare power, we considered three alternatives to equal exponential distributions. In the first, the survival times were exponentially distributed such that S_x and S_y had expectations 1.5 and .5, respectively. Since this shift of scale has the effect of increasing skewness in the consistent direction of treatment effect, this condition should be quite favorable to the signed rank test.

The second alternative consisted of generating unit exponential survival times and then adding .5 to each value of S_x. In this case, the skewness will be the same for both S_x and S_y, so spuriously large observations are nearly as likely in each direction—a hazardous situation for the signed rank test.

Although these two alternatives provide useful insights into the operating characteristics of the procedures, we inspected a third alternative to equal exponential distributions because of its frequency of occurrence in practice. Often, survival curves on semi-log paper fall linearly at first (reflecting early large hazards associated with the disease under study) and then flatten out (implying the disease has become less predominant among competing causes of death). These considerations suggest that an alternative in a middle range between shift in scale and shift in location may be more realistic. Therefore, we generated S_x to have an exponential distribution with unit expectation, but S_y to be piecewise exponential. We first obtained S_y' and S_y'' to be exponential with expectations .15 and 1.0, respectively. If S_y' was less than .1, S_y was set equal to S_y' (representing death due to disease). Otherwise, S_y was set equal to $.1 + S_y''$ (representing death due to other causes, relying on the memoryless property of the exponential distribution).

To compare power in the presence of spurious outliers, a mixed exponential distribution was obtained by replacing the first three pairs of observation times in the exponential shift of scale simulation with uncensored mean 10 exponential variates shifted 5 units to the right.

We also considered the logistic distribution with $S_x = (U_x^{-1} - 1)$ and $S_y = (U_y^{-1} - 1)/3.5$, where U_x and U_y are independent and uniformly distributed. The censoring variables were exponential with expectation 4 in all power configurations having uncorrelated data.

To introduce correlations within pairs, an exponential variable was added to each pair member. This variable had mean .5 for the exponential (location) and piecewise exponential

Table 3
Observed rejection rates: H_0 true

	Uncorrelated data					Correlated data				
		Test procedure					Test procedure			
Configuration	Nominal α	Sign (S)	Generalized signed rank (GSR)	Two-sample Wilcoxon (W)	Paired Prentice–Wilcoxon (PPW)	Nominal α	Sign (S)	Generalized signed rank (GSR)	Two-sample Wilcoxon (W)	Paired Prentice–Wilcoxon (PPW)
Exponential	.10	.072	.094	.089	.098	.10	.058	.093	.032	.102
	.05	.035	.049	.045	.048	.05	.031	.044	.007	.048
	.01	.006	.012	.011	.007	.01	.004	.006	.001	.007
Exponential (outliers)	.10	.070	.096	.047	.092	.10	.063	.090	.010	.096
	.05	.032	.049	.015	.043	.05	.033	.036	.002	.043
	.01	.008	.013	.003	.008	.01	.005	.010	.001	.005
Log-logistic	.10	.068	.100	.097	.093	.10	.062	.090	.057	.095
	.05	.027	.051	.047	.053	.05	.025	.038	.029	.041
	.01	.007	.009	.010	.010	.01	.005	.006	.001	.005

Table 4

Observed rejection rates: H_A true

	Uncorrelated data					Correlated data				
		Test procedure					Test procedure			
Configuration	Nominal α	Sign (S)	Generalized signed rank (GSR)	Two-sample Wilcoxon (W)	Paired Prentice–Wilcoxon (PPW)	Nominal α	Sign (S)	Generalized signed rank (GSR)	Two-sample Wilcoxon (W)	Paired Prentice–Wilcoxon (PPW)
Exponential (scale)	.10	.793	.942	.952	.915	.10	.734	.921	.821	.893
	.05	.678	.899	.910	.863	.05	.601	.847	.687	.813
	.01	.447	.719	.784	.620	.01	.367	.633	.387	.529
Exponential (scale/outliers)	.10	.683	.847	.887	.894	.10	.608	.764	.646	.861
	.05	.571	.748	.810	.817	.05	.492	.639	.452	.772
	.01	.333	.455	.570	.575	.01	.250	.355	.155	.463
Exponential (location)	.10	.707	.768	.853	.885	.10	.597	.678	.685	.767
	.05	.557	.669	.780	.796	.05	.447	.551	.562	.649
	.01	.300	.378	.569	.547	.01	.228	.285	.288	.368
Piecewise exponential	.10	.691	.712	.865	.893	.10	.629	.648	.577	.715
	.05	.560	.584	.799	.811	.05	.488	.509	.436	.581
	.01	.345	.314	.563	.536	.01	.274	.257	.200	.300
Log-logistic	.10	.633	.781	.839	.822	.10	.544	.688	.614	.679
	.05	.500	.675	.761	.721	.05	.407	.564	.477	.544
	.01	.285	.406	.551	.446	.01	.184	.301	.205	.282

alternatives, and mean unity for other configurations. The effect on power of introducing this type of correlation should be most favorable to the GSR test, since it affects pair values on an additive scale in accordance with the assumption implicit in the GSR test. In correlated data, censoring variates were again exponential with expectation 10.

The results (Table 4) for uncorrelated data indicate that the PPW procedure has broad sensitivity. As expected, the PPW is especially powerful against the exponential shift in location and piecewise exponential alternative. The GSR test's particular sensitivity to outlier pairs, as illustrated in earlier examples, is clearly confirmed by the simulations. Introducing such outliers eliminates the small power advantage of GSR in the exponential scale alternatives. It is interesting to observe that little efficiency is lost in using the PPW test instead of the W test with uncorrelated data. This is apparent even for the log-logistic scale alternative against which W is known to be fully efficient.

Introduction of correlation through the addition of an exponential variate, of course, has no effect on the power of the S and GSR tests, while having a mild to severe effect on the power of W, depending on the degree of that correlation. The results for correlated data in Table 4 are consistent with this fact. They also reveal that the GSR and PPW have comparable power in the presence of moderate correlation. Again it is apparent that the GSR procedure is very sensitive to outlier pairs.

5. Discussion

A class of computationally simple and intuitively appealing rank statistics for censored paired data has been proposed through a modification of the efficient Prentice two-sample linear rank statistics. Particular attention has been given to the paired Prentice–Wilcoxon (PPW) statistic, an important member of this new class. This statistic has more desirable properties in heavily censored data than alternative procedures based on Gehan rather than Prentice scores.

The PPW statistic has several important properties resulting from the use of the rank metric to assess the distance between pair members. First, the use of interblock information by PPW depends on the degree of correlation. In uncorrelated data, the statistic makes full use of interblock information and, as a result, has power nearly equal to that of the two-sample Wilcoxon (see Table 4, Uncorrelated data). On the other hand, as correlation becomes high and the ratio of interpair variability to intrapair variability increases, the value of interblock information becomes suspect and, indeed, the PPW appropriately approaches the sign test. In contrast, the use of interblock information by the signed rank test is independent of the degree of correlation and, of course, the sign test never uses such information. A second important property of the PPW test, resulting from the rank metric, is its relative insensitivity to outlier pairs. Simulations in Table 4 and examples discussed earlier verify that outlier pairs will cause considerably greater power loss with the generalized signed test. Third, the PPW is quite powerful against those alternatives for which spuriously large observations are nearly as likely in each sample. These include the exponential shift in location and the piecewise exponential defined earlier. These spuriously large observations typically are given less weight by the PPW, with its rank metric, than by other tests such as the GSR or the conventional paired t test.

In either paired or unpaired survival data, considerable bias can arise in evaluating survivorship if censoring times and survival times are not independent. With paired data, the problem can be eliminated in some instances by using the common censoring-time convention. To illustrate, assume two interventions are to be compared, and individuals are paired based on commonality of survival prognosis. If censoring and survival remain dependent even when one conditions on level of prognosis for any given pair, statistical procedures can have substantial bias whether or not interblock information is used.

Conversely, suppose conditional independence holds. Procedures such as the sign test that do not employ interblock information then would be free of bias. However, even with this conditional independence, bias can still arise for procedures using interblock information, if heaviness of censorship differs between "good prognosis" pairs vs "poor prognosis" pairs and also differs between the two intervention groups. Fortunately, the convention for common censoring does eliminate this potential for bias arising when interblock information is used in such situations.

It is desirable to make regular use of the common censoring-time convention. In the simulations, this convention was followed for all procedures except W, this exception being made to reflect the way the Wilcoxon procedure normally is used. However, to eliminate the possibility of bias, one also should use common censoring when applying W to paired data, which in turn will result in some reduction in its power.

In summary, the PPW statistic provides a versatile procedure for censored paired data. Results of Lam and Longnecker (1983) indicate the procedure in uncensored data has power comparable to a conventional paired *t* test when sampling from a bivariate normal distribution. In uncorrelated censored data, simulations reveal that PPW provides noticeably better overall power than either the sign or GSR procedure, while it is nearly as powerful as W, even against log-logistic scale alternatives. In moderately correlated data where the operating characteristics of W are inferior, PPW provides similar power to that of the GSR procedure, which in turn is an improvement over the sensitivity of the sign test. Finally, PPW suffers a smaller power loss than the GSR test in the presence of outlier pairs.

Acknowledgement

The authors would like to thank Mr Jon Kosanke for programming the Monte Carlo simulations.

Résumé

Prentice (1978, *Biometrika* **65,** 167–179) a donné une généralisation pour les données censurées des statistiques de rangs linéaires classiques. Cet article explore une méthode qui généralise les statistiques de Prentice à l'analyse de données appariées censurées. Il donne un éclairage particulier sur la version des données appariées de la statistiques de Prentice–Wilcoxon, en discutant les raisons de préférer les scores de Prentice plutôt que ceux de Gehan. La statistique pour des données appariées de Prentice–Wilcoxon (PPW) utilise l'information interblocs et peut être vue comme une généralisation du "test *t* des paires sur les rangs" de Conover–Iman (1981, *The American Statistician* **35,** 124–129) ou de la statistique sur les sommes de rangs de Wilcoxon modifiée par Lam–Longnecker (1983, *Biometrika* **70,** 510–513). Sur des simulations comparant la PPW, les statistiques de signes et des rangs signées (GSR), la PPW est bien plus puissante que toutes les autres sauf pour l'alternative d'échelle exponentielle. De plus, le petit avantage de la GSR sur la PPW pour cette alternative est perdu si on introduit des paires de données aberrantes. Quand on la compare à la statistique de Prentice–Wilcoxon de deux échantillons, la PPW est pratiquement aussi puissante pour des données non corrélées et sa puissance devient d'autant plus grande que la corrélation entre les unités appariées augmente.

References

Batchelor, J. R. and Hackett, M. (1970). HL-A matching in treatment of burned patients with skin allografts. *Lancet* **2,** 581–583.

Cheng, K. F. (1984). Asymptotically nonparametric tests with censored paired data. *Communications in Statistics—Theory and Methods* **13(12),** 1453–1470.

Conover, W. J. and Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician* **35,** 124–129.

Cuzick, J. (1982). Rank tests for association with right-censored data. *Biometrika* **69,** 351–364.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*, 2nd edition. New York: Wiley.

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

Gill, R. D. (1980). *Censoring and Stochastic Integrals.* Mathematical Centre Tracts 124. Amsterdam: Mathematisch Centrum.

Holt, J. D. and Prentice, R. L. (1974). Survival analysis in twin studies and matched-pair experiments. *Biometrika* **61,** 17–30.

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

Lam, F. C. and Longnecker, M. T. (1983). A modified Wilcoxon rank sum test for paired data. *Biometrika* **70,** 510–513.

Mantel, N. and Ciminera, J. (1979). Use of log-rank scores in the analysis of litter-matched data on time to tumor appearance. *Cancer Research* **39,** 4308–4315.

Massey, F. J., Bernstein, G. S., O'Fallon, W. M., et al. (1984). Vasectomy and health: Results from a large cohort study. *Journal of the American Medical Association* **252,** 1023–1029.

Michalek, J. E. and Mihalko, D. (1983). On the use of log-rank scores in the analysis of litter-matched data on time to tumor appearance. *Statistics in Medicine* **2,** 315–321.

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,** 861–869.

Wei, L. J. (1980). A generalized Gehan and Gilbert test for paired observations that are subject to arbitrary right censorship. *Journal of the American Statistical Association* **75,** 634–637.

Woolson, R. F. and Lachenbruch, P. A. (1980). Rank tests for censored matched pairs. *Biometrika* **67,** 597–606.

Received July 1985; *revised May* 1986.