The Korean Communications
in Statistics Vol. 6, No. 2, 1999
pp. 565-583

Application of Covariance Process to Tests for Censored Paired Data[1)]

Gyu-Jin Jeong[2)]

Abstract

The covariance process of two martingales provides a useful tool to capture the dependence structure for paired censored data. In this paper, it is applied to modify the variances of weighted logrank tests in order to take account of dependence between paired subjects. In the process of modification, a 'variance correction term' is introduced. Some variance estimators based on separate samples are considered together. Performance of the estimators are compared through simulation studies. Several independence tests for bivariate survival data are also proposed, which are naturally reduced from the weighted logrank tests accomodating dependence structure. Simulation studies are carried out to compare the independence tests. Both the weighted logrank tests and the independence tests are illustrated by an example.

1. Introduction

In a clinical trial, an experimenter might often encounter paired subjects such as twins and a pair of eyes of a patient. The paired subjects produce censored paired survival times with dependence structure. For example, the survival times (measured in days) of closely-matched and poorly-matched skin grafts on each patient are found in the well known skin-graft data (Bachelor and Hackett, 1970). In analysing such a kind of data, testing equality of two marginal survival functions as well as testing independence of survival times is, of course, included in our primary concerns.

When two survival times are independent, lots of procedures to test equality of survival functions have been developed. Especially, in the presence of censoring, the class of weighted logrank tests has been highlighted for a couple of decades because it contains many important tests both in practice and in theory. It includes the logrank test (Mantel and Haenszel, 1959), Gehan-Wilcoxon test (Gehan, 1965), Tarone-Ware test (Tarone and Ware, 1977), Prentice-Wilcoxon test (Prentice, 1978), and the class of tests with Harrington and Fleming (1982) weights. Properties and recent developments of weighted logrank tests are presented in, for example, Fleming and Harrington (1991), and Klein and Moeschberger (1997).

---

1) This paper was accomplished with Research Fund provided by Korea Research Foundation, Support for Faculty Research Aboard.

2) Professor, Department of Informational Statistics, Hannam University, Taejon 300-791, Korea

To accomodate dependence structure for correlated survival data, the usual one-sample and two-sample tests for independent samples have been generalized through modifying variances of the test statistics. Among others, Woolson and Lachenbruch (1980), and Dabrowska (1990) generalized the signed rank test. Wei (1980), O'Brien and Fleming (1987), and Dabrowska (1989) proposed to use modified Gehan-Wilcoxon test, Prentice-Wilcoxon test and logrank test, respectively. Due to the complexity caused by the dependence structure, most of the above tests were considered under the assumption of univariate censoring and/or the location model. However, Jung (1998) recently presented a simple variance estimator without these assumptions using martingale representation for the weighted logrank tests.

In this paper, various modifications are considered for the variance estimators of weighted logrank tests. As a factor producing different estimators, consider two sampling schemes, independent sample case and dependent sample case. For each sampling scheme, two simple estimation methods are used. One is replacing the cumulative hazards and covariance process with their usual consistent estimators, the other is a kind of moment estimating method based on the martingale residuals, which was used in Jung (1998). Estimators based on both separate samples and a combined sample are included in estimating the cumulative hazards and covariance process. In addition, estimators with a 'variance correction term' are introduced. All estimators are compared through simulation studies.

Comparing the estimators for a paired sample with those for independence samples, a couple of independence tests are naturally derived from the weighted logrank tests. For the independence tests of survival times, several semiparametric and nonparametric procedures are available from the literature. A method based on the frailty model was introduced by Clayton (1978) and further developed by Oakes (1982). On the other hand, Hsu and Prentice (1996) and Shih and Louis (1996) proposed the tests based on the covariance process of two martingale residuals (Fleming and Harrington, 1991, and Prentice and Cai, 1992).

Besides two independence tests reduced from the weighted logrank tests, we also mention a modification of Shih and Louis (1996) test by means of a 'variance correction term', and carry out a simulation to compare the proposed tests with the tests of Hsu and Prentice (1996) and Shih and Louis (1996). As an illustration, the skin graft data is analysed using both equality tests of marginal survival functions and independence tests.

2. Test Statistics

For $i=1,\cdots,n$, let $\{(X_{1i},\delta_{1i}),(X_{2i},\delta_{2i})\}$ be n independent observations of $\{(X_1,\delta_1),(X_2,\delta_2)\}$, where $X_k=T_k\wedge C_k$ $(k=1,2)$ for the survival time T_k and the censoring time C_k and $\delta_k=I(X_k=T_k)$ denotes the censoring indicator. We assume that T_k and C_k are independent.

Before describing the hypotheses of interest and test statistics, it seems useful to express first the covariance process and its estimator in the martingale framework. As shown in Fleming and Harrington (1991), the covariance process of paired survival times (T_1,T_2) is

defined by

$$A(dt_1, dt_2) = E\{M_1(dt_1)M_2(dt_2) \mid T_1 \geq t_1, T_2 \geq t_2\},$$

where $M_k(t) = N_k(t) - \int_0^t Y_k(s)d\Lambda_k(s)$ $(k=1,2)$ is the marginal martingale, $N_k(t) = I(X_k \leq t, \delta_k = 1)$ is the counting process, $Y_k(t) = I(X_k \geq t)$ is the size of the risk set at time t, and Λ_k denotes the cumulative hazard function of T_k.

A simple moment estimator of the covariance process can be obtained by replacing the marginal martingales with martingale residuals. Prentice and Cai (1992) presented a consistent estimator $\hat{A}(dt_1, dt_2)$ of $A(dt_1, dt_2)$ as

$$\hat{A}(dt_1, dt_2) = Y(t_1, t_2)^{-1} \sum_{i \in R(t_1, t_2)} \hat{M}_{1i}(dt_1)\, \hat{M}_{2i}(dt_2)$$

if (t_1, t_2) belongs to the grid formed by the observed failure times and 0 otherwise. In this expression, $R(t_1, t_2)$ and $Y(t_1, t_2)$ denote the risk set at time (t_1^-, t_2^-) and the size of $R(t_1, t_2)$, respectively, and the martingale residuals $\hat{M}_{ki}$ are defined by $\hat{M}_{ki}(t) = N_{ki}(t) - \int_0^t Y_{ki}(s)\, d\hat{\Lambda}_k(s)$. Here $\hat{\Lambda}_k(t) = \int_0^t Y_k^{-1}(s)dN_k(s)$ is the usual Nelson-Aalen (Nelson, 1969) estimator of $\Lambda_k(t)$ where $N_k(t) = \sum_{i=1}^{n} N_{ki}(t)$ and $Y_k(t) = \sum_{i=1}^{n} Y_{ki}(t)$. This estimator $\hat{A}(dt_1, dt_2)$ is used to construct test statistics for paired survival data in the next sections.

2.1 Logrank test

In this subsection, we are interested in testing the null hypothesis,

$$H_0 : \Lambda_1(t) = \Lambda_2(t) \text{ for all } t \geq 0,$$

and we consider the usual two sample weighted logrank statistics V with a weight function $W(t)$,

$$V = n^{1/2} \int_0^\infty W(t)\{d\hat{\Lambda}_1(t) - d\hat{\Lambda}_2(t)\}.$$

Note that $W(t) = n^{-1}Y_1(t)Y_2(t)/\{Y_1(t) + Y_2(t)\}$ for the logrank test, $n^{-2}Y_1(t)Y_2(t)$ for Gehan-Wilcoxon test, and $W(t) = n^{-1}\hat{S}(t^-)Y_1(t)Y_2(t)/\{Y_1(t) + Y_2(t)\}$ for Prentice-Wilcoxon test. Here $\hat{S}(t^-)$ is the Kaplan-Meier (1958) estimator of the common survival function under the null hypothesis.

Under H_0, the logrank statistic V can be written as a sum of independent stochastic integrals,

$$V = n^{-1/2} \sum_{i=1}^{n} (\varepsilon_{1i} - \varepsilon_{2i}),$$

where $\varepsilon_{ki} = \int_0^\infty nW(t)/Y_k(t)\, dM_{ki}(t)$ for $k = 1, 2$. Since V is also a martingale, from the standard martingale central limit theorem, it follows that V converges in distribution to a normal with mean 0 and variance σ^2,

$$\begin{aligned} \sigma^2 = \; & E[\int_0^\infty n^2 W^2(t)\{(1-\Delta\Lambda_1(t))\, d\Lambda_1(t)/Y_1(t) + (1-\Delta\Lambda_2(t)) d\Lambda_2(t)/Y_2(t)\} \\ & - 2\int_0^\infty \int_0^\infty n^2 W(t_1) W(t_2) Y(t_1, t_2)/\{Y_1(t_1) Y_2(t_2)\} A(dt_1, dt_2)]. \end{aligned} \tag{1}$$

Here $\Delta\Lambda_k(t) = \Lambda_k(t) - \Lambda_k(t^-)$ is the correction term for discontinuity of the underlying distribution. Therefore, when there exists a consistent estimator $\hat{\sigma}$ of σ, the standard normal tests can be applied to test H_0 based on the statistic $V/\hat{\sigma}$. See Fleming and Harrington (1991) for more details about martingale theory including martingale central limit theorem.

In this setting, the performance of the test procedure may depend heavily on the variance estimator $\hat{\sigma}$, and there are several factors which produce different estimators. As such a factor, first, consider two sampling schemes, independent sample case and dependent sample case. Since the second term of the right hand side of (1) is dropped if T_1 and T_2 are independent, we can use the first term of (1) as an asymptotic variance in the independent sample case. For each sampling scheme, a simple way to get a consistent estimator of σ is substituting Λ_1, Λ_2 and A in (1) with their consistent estimators, that is, the Nelson-Aalen estimator $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ for Λ_1 and Λ_2 and the Prentice and Cai (1992) estimator $\hat{A}$ for A.

To access the second method, note that $\sigma^2 = n^{-1} \sum_{i=1}^{n} E(\varepsilon_{1i} - \varepsilon_{2i})^2$ which is reduced to $\sigma^2 = n^{-1} \sum_{i=1}^{n} E(\varepsilon_{1i}^2 + \varepsilon_{2i}^2)$ in the independent sample case because V is a sum of independent variables. Hence we obtain another estimator of σ by simply calculating a moment estimator $n^{-1} \sum_{i=1}^{n} (\hat{\varepsilon}_{1i} - \hat{\varepsilon}_{2i})^2$, where $\hat{\varepsilon}_{ki} = \int_0^\infty nW(t)/Y_k(t)\, d\hat{M}_{ki}(t)$. In the independent sample case, we use $n^{-1} \sum_{i=1}^{n} (\hat{\varepsilon}_{1i}^{\,2} + \hat{\varepsilon}_{2i}^{\,2})$. A few applications of this method are found in Lam and Longnecker (1983) for constructing Wilcoxon test in the absence of censoring, in O'Brien and Fleming (1987) and Jung (1998) for the logrank tests, and in Hsu and Prentice (1996) for the independence test.

When constructing a consistent estimator of σ along these two methods, we are allowed to use both of the combined sample estimator $\hat{\Lambda}$ and separate sample estimators $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ for Λ_1 and Λ_2, since these estimators are asymptotically equivalent under the null hypothesis. This is the same as that we can use both the pooled variance and separate variances in the

paired t-test. When the combined sample estimator is used, it should be noted that $\hat{A}$ must be calculated on each grid point (T_i, T_j) formed by uncensored failure times T_i and T_j in the combined sample.

The last factor is related to discontinuity correction term $\Delta\Lambda_k$ in (1). As an estimator of discontinuity correction term $\Delta\Lambda_k$, Fleming and Harrington (1991) recommended to use $\Delta\hat{\Lambda}^* = \{\Delta N_1 + \Delta N_2 - 1\}/\{Y_1 + Y_2 - 1\}$ in the combined sample, or $\Delta\hat{\Lambda}_k^* = \{\Delta N_k - 1\}/\{Y_k - 1\}$ in separate samples, only when discontinuity exists. However, it is shown in Appendix that

$$\hat{\sigma}_k^2 = n^{-1}\sum_{i=1}^{n} \hat{\varepsilon}_{ki}^2 = \int_0^\infty n^2 W^2(t)/Y_k(t)\,(1-\Delta\hat{\Lambda}_k(t))\,d\hat{\Lambda}_k(t), \tag{2}$$

where $\Delta\hat{\Lambda}_k = \Delta N_k/Y_k$ and $d\hat{\Lambda}_k = dN_k/Y_k$. Since $\hat{\sigma}_k^2$ of (2) is a consistent estimator of corresponding population variance σ_k^2, it seems possible that we include a 'variance correction term', $1-\Delta\hat{\Lambda}_k(t)$, or $1-\Delta\hat{\Lambda}(t)$, in constructing a consistent estimator of σ^2 given by (1) regardless of continuity of the underlying distribution. The 'variance correction terms' are less than the discontinuity correction terms when there are some ties. It leads us to expect that the tests with the 'variance correction terms' have smaller variance than the others.

The variance estimators obtained from the above discussions are listed in Table 1. In the sequel, let $V_k = V/\hat{\sigma}_k$ for $k = 1, \cdots, 10$. The first five tests are for the independent samples and the others are derived to accomodate dependence of paired data. The 'variance correction terms' are included in V_2, V_4, V_7 and V_9. Replacing them with the discontinuity correction terms, it yields their counterparts, V_1, V_3, V_6, and V_8. Four tests, V_3, V_4, V_8, and V_9, are based on separate samples and their respective counterparts based on a combined sample are V_2, V_4, V_7, and V_9. The sum of squares type variance estimators are used in V_4, V_5, V_9 and V_{10}. By the equation (1), they are also affected the 'variance correction terms'. Note that, if we use the separate sample estimators $d\hat{\Lambda}_1$ and $d\hat{\Lambda}_2$, the fifth estimator is equal to the fourth estimator. Among ten tests, the last one was proposed by Jung (1998) and the first corresponds to the usual weighted logrank test. The performance of these tests will be compared through a simulation study in a later section.

2.2 Independence test

In the previous section, we discussed the variance estimators of weighted logrank tests. Since some of them are taking the dependence into account, a class of independence tests is accessible by comparing them with their counterparts derived for the independent sample case.

From the third and eighth estimators, for example, we may have a test statistic

$$S = \int_0^\infty \int_0^\infty n^2 W(t_1) W(t_2) Y(t_1, t_2)/\{Y_1(t_1) Y_2(t_2)\}\hat{A}(dt_1, dt_2).$$

Table 1. Consistent variance estimators for weighted logrank tests

$\hat{\sigma}_1^2 = \int_0^\infty n^2 W^2(t)\{1/Y_1(t)+1/Y_2(t)\}(1-\Delta\hat{\Lambda}^*(t))\,d\hat{\Lambda}(t)$

$\hat{\sigma}_2^2 = \int_0^\infty n^2 W^2(t)\{1/Y_1(t)+1/Y_2(t)\}(1-\Delta\hat{\Lambda}(t))\,d\hat{\Lambda}(t)$

$\hat{\sigma}_3^2 = \sum_{i=1}^{2}\int_0^\infty n^2 W^2(t)/Y_k(t)\,(1-\Delta\hat{\Lambda}_k^*(t))\,d\hat{\Lambda}_k(t)$

$\hat{\sigma}_4^2 = \sum_{i=1}^{2}\int_0^\infty n^2 W^2(t)/Y_k(t)\,(1-\Delta\hat{\Lambda}_k(t))\,d\hat{\Lambda}_k(t)$

$\hat{\sigma}_5^2 = n^{-1}\sum_{i=1}^{n}(\hat{\varepsilon}_{1i}^2+\hat{\varepsilon}_{2i}^2)$, calculated with the common $d\hat{\Lambda}$.

$\hat{\sigma}_6^2 = \hat{\sigma}_1^2 - 2\int_0^\infty\int_0^\infty n^2 W(t_1)W(t_2)Y(t_1,t_2)/\{Y_1(t_1)Y_2(t_2)\}\hat{A}(dt_1,dt_2)$,

$\hat{A}$ calculated with the common $d\hat{\Lambda}$.

$\hat{\sigma}_7^2 = \hat{\sigma}_2^2 - 2\int_0^\infty\int_0^\infty n^2 W(t_1)W(t_2)Y(t_1,t_2)/\{Y_1(t_1)Y_2(t_2)\}\hat{A}(dt_1,dt_2)$,

$\hat{A}$ calculated with the common $d\hat{\Lambda}$.

$\hat{\sigma}_8^2 = \hat{\sigma}_3^2 - 2\int_0^\infty\int_0^\infty n^2 W(t_1)W(t_2)Y(t_1,t_2)/\{Y_1(t_1)Y_2(t_2)\}\hat{A}(dt_1,dt_2)$,

$\hat{A}$ calculated with $d\hat{\Lambda}_1$ and $d\hat{\Lambda}_2$.

$\hat{\sigma}_9^2 = n^{-1}\sum_{i=1}^{n}(\hat{\varepsilon}_{1i}-\hat{\varepsilon}_{2i})^2$, calculated with $d\hat{\Lambda}_1$ and $d\hat{\Lambda}_2$.

$\hat{\sigma}_{10}^2 = n^{-1}\sum_{i=1}^{n}(\hat{\varepsilon}_{1i}-\hat{\varepsilon}_{2i})^2$, calculated with the common $d\hat{\Lambda}$.

Note that the Prentice and Cai estimator $\hat{A}(dt_1,dt_2)$ equals to $Y(t_1,t_2)^{-1}\sum_{i=1}^{n}\hat{M}_{1i}(dt_1)\,\hat{M}_{2i}(dt_2)$ because $\hat{M}_{1i}(dt_1)\,\hat{M}_{2i}(dt_2)=0$ for any $i\notin R(t_1,t_2)$. Therefore S can be written by

$$S_1 = n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau_1}\int_0^{\tau_2} W(t_1,t_2)\,\hat{M}_{1i}(dt_1)\,\hat{M}_{2i}(dt_2),$$

where the weight function $W(t_1,t_2)=n^2W(t_1)W(t_2)Y(t_1,t_2)/\{Y_1(t_1)Y_2(t_2)\}$ which is assumed to converge uniformly in probability to a fixed function over $[0,\tau_1]\times[0,\tau_2]$ and $\tau_k=\sup\{t:P(X_{k1}>t)>0\}$. This statistic S_1 has the same form as that proposed by Hsu and Prentice (1996) and Shih and Louis (1996).

Under the null hypothesis of independence, they showed that the statistic S_1 has an asymptotic normal distribution with mean 0 if some regularity conditions are satisfied.

Furthermore, Hsu and Prentice(1996) proposed a consistent variance estimator

$$\hat{\sigma}_{HP}^{\,2} = n^{-1}\sum_{i=1}^{n}\{\int_0^{\tau_1}\int_0^{\tau_2} W(t_1, t_2)\,\widehat{M}_{1i}(dt_1)\,\widehat{M}_{2i}(dt_2)\}^2,$$

while Shih and Louis (1996) presented a consistent estimator

$$\hat{\sigma}_{SL1}^{\,2} = n^{-1}\int_0^{\tau_1}\int_0^{\tau_2} W^2(t_1, t_2)Y(t_1, t_2)d\widehat{\Lambda}_1(t_1)d\widehat{\Lambda}_2(t_2).$$

With the same reason as mentioned for logrank test, the latter can be modified by using the variance correction terms as

$$\hat{\sigma}_{SL2}^{\,2} = n^{-1}\int_0^{\tau_1}\int_0^{\tau_2} W^2(t_1, t_2)Y(t_1, t_2)(1-\Delta\widehat{\Lambda}_1(t_1))(1-\Delta\widehat{\Lambda}_2(t_2))d\widehat{\Lambda}_1(t_1)d\widehat{\Lambda}_2(t_2).$$

Comparing ninth estimator with third and fourth estimators in the Table 1, we can get two more statistics, S_2 and S_3, applicable to the independence test,

$$S_2 = (\hat{\sigma}_3^{\,2} - \hat{\sigma}_9^{\,2})/2$$

$$= n^{-1}\sum_{i=1}^{n}[\sum_{i=1}^{2}\int_0^{\infty} n^2\{W(t)/Y_k(t)\}^2 Y_{ki}(t)(1-\Delta\widehat{\Lambda}_k^{*}(t))d\widehat{\Lambda}_k(t) - (\hat{\varepsilon}_{1i}-\hat{\varepsilon}_{2i})^2]/2.$$

The statistic S_3 is obtained by replacing $\Delta\widehat{\Lambda}_k^{*}$ with $\Delta\widehat{\Lambda}_k$. The variances of S_2 and S_3 can be estimated consistently by summing up squares of quantity next to the first summation as shown in $\hat{\sigma}_{HP}^{\,2}$. Note that S_2 is apt to be larger than S_3 because $\hat{\sigma}_3^{\,2}$ has no 'variance correction term' while $\hat{\sigma}_9^{\,2}$ has a 'variance correction term'.

So far we introduced the following five tests ;

1. test1 reduced from logrank test : $S_{LR1} = S_2/\sqrt{\widehat{Var}(S_2)}$.
2. test2 reduced from logrank test : $S_{LR2} = S_3/\sqrt{\widehat{Var}(S_3)}$.
3. Hsu-Prentice : $S_{HP} = S_1/\hat{\sigma}_{HP}$.
4. Shih-Louis : $S_{SL1} = S_1/\hat{\sigma}_{SL1}$.
5. modified Shih-Louis : $S_{SL2} = S_1/\hat{\sigma}_{SL2}$.

Among the tests, three tests, S_{LR1}, S_{LR2} and S_{SL2} include the 'variance correction terms'. These tests as well as several weight functions will be compared through simulation studies in the next section.

3. Simulation

3.1 Simulation scheme

To compare the performance of test procedures described in section 2, simulation studies were carried out. Logrank, Prentice-Wilcoxon and Gehan-Wilcoxon weights were chosen for

the weighted logrank tests. For the independence tests, we denote these three weights from the weighted logrank tests by $W_2(t_1, t_2)$, $W_3(t_1, t_2)$ and $W_4(t_1, t_2)$, and three more weights $W_1(t_1, t_2)=1$, $W_5(t_1, t_2)=\widehat{S}_1(t_1^-)\widehat{S}_2(t_2^-)$, and $W_6(t_1, t_2)=\widehat{S}_1(t_1^-)\{1-\widehat{S}_1(t_1^-)\}\widehat{S}_2(t_2^-)\{1-\widehat{S}_2(t_2^-)\}$, were added. Here, $W_1(t_1, t_2)$ and $W_5(t_1, t_2)$ can be considered simplified logrank and Prentice-Wilcoxon weights, respectively.

For both the weighted logrank tests and the independence tests, correlated random numbers were generated from three bivariate distributions. One was a bivariate exponential distribution constructed by Moran's algorithm (Moran, 1967). The other two were from the Clayton's family (Clayton, 1978). One of them has exponential margins while the other has logistic margins. Random variates for the Clayton's family were obtained by Oakes (1982) method.

Three parameters were considered for each distribution, marginal failure rates λ_1, λ_2 and correlation coefficient ρ for Moran's exponential, marginal failure rates λ_1, λ_2 and Kendall's tau τ for Clayton's exponential, and marginal means μ_1, μ_2 and Kendall's tau τ for Clayton's logistic. We chose λ_1=1, λ_2=1, 0.75, and μ_1=100, μ_2=100, 100.5 and ρ(or τ)=0.0, 0.2. We used about 30% censoring together with no censoring. As in Jung (1998), censoring variates were generated from $U(0, 3/\lambda_k)$, $k=1, 2$, for the exponential distributions. In the case of the logistic distributions, 30% censoring could be achieved by using $U(-3, 7)+\mu_k$, $k=1, 2$, as the censoring distribution. All empirical type 1 error probabilities and powers were calculated based on 1000 samples of size n=50. All tests were done at level 0.05.

3.2 Simulation results: logrank tests

Some typical results for empirical type I errors are given in Table 2 and 3. From the tables, we observe that the tests with a 'variance correction term' (V_2, V_4, V_7, V_9) are less conservative than their counterparts (V_1, V_3, V_6, V_8). Another anti-conservativeness is shown in V_4, V_5, V_9 and V_{10} for independent samples, and in V_9 for a paired sample. Note that these tests have their sum of squares type variance estimators so that they have the 'variance correction term' effects. So we guess that the anti-conservativeness can be explained by the 'variance correction term'.

In fact, the type I errors of the tests, affected by a 'variance correction term' effect, appear to exceed the nominal level for some situations. Especially, it seems severe when they are combined with the logrank weights (LR). As known, the logrank weights have relatively larger values on late time than Prentice - Wilcoxon and Gehan - Wilcoxon weights. Furthermore, this property of the logrank weights is combined with the 'variance correction term' effect which also is larger on late time. Note that the 'variance correction term' decreases in time t. It might be an explanation for the reason that the type I errors for the logrank weights are larger than the other weights.

We don't observe any special differences between tests based on the combined sample and tests on separated samples so that the separated sample estimators may be used instead of the combined sample estimators for computational convenience. In most cases, censoring gives larger type I errors. If we take into account that the considered tests are designed for censored data, it does not seem to be surprised. As expected, it is also observed that the tests for independent samples are quite conservative when dependency exists. In the independent sample cases, Table 2 shows there seem to be no serious difference between the tests for independent samples and the tests for a paired sample.

In the case of τ=0.2, no censoring and Clayton family with equal marginal means, it should be mentioned that the type I errors for exponential and logistic distributions are found to be the same. It seems due to the algorithm of random number generation which makes the same ranks for both distributions. It is also observed that the same situation occurs in the Table 7 of the independence tests.

Table 4 and 5 represents the empirical powers for independent samples and paired sample, respectively. They show very similar results as in comparison of type I errors. So, we will describe them briefly. The tests with a 'variance correction term' and with the sum of squares type estimators produce slightly higher powers than their counterparts. The tests based on the separated samples are comparable to those based on the combined sample. The tests for independent samples have, of course, smaller power than the tests for dependent samples, while both are similar in the case of independent samples. Although a few exceptions are found, the logrank weights and Prentice-Wilcoxon weights seem to be optimal for the exponential distributions and the logistic distributions, respectively. However, the theoretical consideration of optimality was difficult, so we could not access it. It is also observed that Prentice - Wilcoxon weights seem to be less affected by censoring than the other weights.

3.3 Simulation results: independence tests

The empirical type I errors are shown in Table 6. First of all, the empirical type I errors of S_{LR1} are found to exceed nominal level for the logrank weights W_1 and W_2, while it is rather conservative for the other weights. It seems due to the combined effect of the 'variance correction term' and the logrank weights. This effect is also found a little in the rest two tests affected by the 'variance correction term', S_{LR2} and S_{SL2}. It is observed for S_{LR2} and S_{SL2} that the former is quite conservative whereas the latter has slightly larger type I errors than its counterpart S_{SL1}. The type I errors of S_{HP} and S_{SL2} are closer to the nominal level. On the contrary, S_{LR1} is more conservative than S_{HP} and S_{SL2}. As for censoring and the marginal distributions, there seem to be nothing special to remark.

The empirical powers are summarized in Table 7. S_{LR1}, in spite of its conservativeness for W_3- W_6, are superior to the others except for W_4. Especially, it has pretty high power for W_1

and W_2, which seems to be due to its anti-conservativeness. Except for S_{LR1}, S_{SL2} has slightly higher powers than others in most circumstances. According to these observations, we guess that the 'variance correction term' affects a little to S_{LR1}, S_{LR2} and S_{SL2} even though it is not clear. It is also observed that W_2 and W_3, the weights from logrank tests, give powers similar to those of corresponding weights W_1 and W_5. Though the powers of logrank weights are relatively high for most cases, it is interesting that any single weight does not dominate the others even for the Clayton family, for which Shih and Louis (1996) showed that W_1 is optimal. The powers appear to be heavily affected by censoring for all tests and all weights. There are nothing special in the marginal distributions.

4. Example

The well known skin graft data are used for illustration. The data are taken from Woolsen and Lachenbruch (1980). There are 11 pairs of survival times, recorded in days, of closely and poorly matched skin grafts on each patient. Two observations for closely matched skin grafts are censored.

The two-sided p-values for the weighted logrank tests are shown in Table 8. For each test, logrank, Prentice-Wilcoxon and Gehan-Wilcoxon weights are used. However, the results for Prentice-Wilcoxon and Gehan-Wilcoxon test coincide accidentally, so the latter test is omitted in Table 8.

Table 8. Two-sided p-values of V_k for logrank and Prentice-Wilcoxon weights.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
logrank	.057	.047	.055	.034	.021	.044	.035	.011	.002	.012
P-W	.085	.075	.084	.063	.064	.052	.042	.024	.010	.032

Before starting discussion, it should be mentioned that Table 8 was obtained from a small data while the tests are based on asymptotic theory. Though it may lead to wrong conclusions, but the results here are quite similar to those from the simulations.

As shown in the independence tests below, this data seem to have a little dependence. It makes the tests for independent samples less significant than those for a paired sample. It is also observed that tests with a 'variance correction term' (V_2, V_4, V_7, V_9) give smaller p-values than their counterparts without the correction term (V_1, V_3, V_6, V_8). As comparing

Table 2. Empirical type I error of weighted logrank tests with LR: logrank, PW: Prentice-Wilcoxon and GW: Gehan-Wilcoxon weights for independent samples from I : exponential($\lambda_1=\lambda_2=1$) and II : logistic($\mu_1=\mu_2=100$) distributions.

		V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
						no censoring					
I	LR	0.048	0.050	0.040	0.051	0.056	0.047	0.052	0.043	0.057	0.061
	PW=GW	0.045	0.045	0.044	0.045	0.046	0.045	0.046	0.045	0.051	0.045
II	LR	0.055	0.060	0.043	0.062	0.063	0.046	0.052	0.045	0.057	0.057
	PW=GW	0.045	0.045	0.044	0.045	0.049	0.045	0.049	0.047	0.051	0.049
						30% censoring					
I	LR	0.063	0.066	0.055	0.067	0.068	0.061	0.063	0.056	0.067	0.066
	PW	0.061	0.066	0.056	0.063	0.063	0.063	0.063	0.064	0.074	0.064
	GW	0.060	0.062	0.058	0.063	0.060	0.060	0.061	0.061	0.066	0.060
II	LR	0.065	0.068	0.058	0.076	0.080	0.067	0.069	0.060	0.078	0.074
	PW	0.057	0.061	0.051	0.060	0.059	0.056	0.058	0.057	0.062	0.057
	GW	0.065	0.066	0.059	0.065	0.065	0.057	0.059	0.059	0.062	0.059

Table 3. Empirical type I error of weighted logrank tests with LR: logrank, PW: Prentice-Wilcoxon and GW: Gehan-Wilcoxon weights for a paired sample (ρ or $\tau=0.2$) from I : bivariate exponential($\lambda_1=\lambda_2=1$) by Moran's algorithm, II : Clayton family with exponential margins($\lambda_1=\lambda_2=1$) and III : Clayton family with logistic margins($\mu_1=\mu_2=100$).

		V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
						no censoring					
I	LR	0.022	0.028	0.018	0.029	0.032	0.038	0.043	0.039	0.052	0.052
	PW=GW	0.029	0.030	0.028	0.030	0.032	0.034	0.036	0.035	0.041	0.036
II	LR	0.023	0.024	0.019	0.028	0.030	0.048	0.056	0.052	0.066	0.058
	PW=GW	0.026	0.027	0.025	0.027	0.026	0.052	0.052	0.052	0.055	0.052
III						same as II					
						30% censoring					
I	LR	0.042	0.043	0.037	0.043	0.044	0.054	0.055	0.054	0.062	0.057
	PW	0.038	0.039	0.036	0.039	0.041	0.057	0.057	0.057	0.062	0.056
	GW	0.041	0.043	0.038	0.041	0.041	0.053	0.054	0.055	0.058	0.053
II	LR	0.028	0.036	0.024	0.036	0.036	0.056	0.060	0.050	0.065	0.061
	PW	0.022	0.024	0.022	0.024	0.025	0.044	0.045	0.045	0.048	0.045
	GW	0.021	0.023	0.020	0.022	0.022	0.034	0.039	0.038	0.044	0.037
III	LR	0.031	0.032	0.026	0.039	0.037	0.057	0.062	0.053	0.068	0.062
	PW	0.032	0.033	0.029	0.033	0.030	0.047	0.048	0.047	0.054	0.048
	GW	0.032	0.033	0.028	0.034	0.034	0.045	0.049	0.047	0.052	0.046

Table 4. Empirical power of weighted logrank tests with LR: logrank, PW: Prentice-Wilcoxon and GW: Gehan-Wilcoxon weights for independent samples from I : exponential(λ_1=1, λ_2=0.75) and II: logistic(μ_1=100, μ_2=100.5) distributions.

		V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
						no censoring					
I	LR	0.331	0.344	0.304	0.349	0.359	0.314	0.329	0.306	0.352	0.348
	PW=GW	0.245	0.251	0.233	0.249	0.253	0.228	0.235	0.234	0.243	0.240
II	LR	0.246	0.255	0.235	0.264	0.263	0.232	0.239	0.228	0.258	0.251
	PW=GW	0.291	0.295	0.280	0.293	0.293	0.270	0.277	0.281	0.293	0.279
						30% censoring					
I	LR	0.229	0.234	0.210	0.237	0.245	0.222	0.228	0.216	0.238	0.239
	PW	0.212	0.217	0.202	0.213	0.223	0.191	0.200	0.194	0.207	0.201
	GW	0.194	0.198	0.187	0.194	0.198	0.180	0.185	0.183	0.189	0.183
II	LR	0.238	0.249	0.217	0.253	0.251	0.237	0.247	0.230	0.259	0.254
	PW	0.266	0.268	0.256	0.266	0.268	0.250	0.256	0.250	0.260	0.255
	GW	0.262	0.265	0.249	0.260	0.264	0.254	0.258	0.256	0.264	0.253

Table 5. Empirical power of weighted logrank tests with LR: logrank, PW: Prentice-Wilcoxon and GW: Gehan-Wilcoxon weights for a paired sample (ρ or τ=0.2) from I : bivariate exponential(λ_1=1, λ_2=0.75) by Moran's algorithm, II: Clayton family with exponential margins(λ_1=1, λ_2=0.75) and III: Clayton family with logistic margins(μ_1=100, μ_2=100.5).

		V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
						no censoring					
I	LR	0.294	0.305	0.277	0.312	0.330	0.354	0.365	0.354	0.405	0.395
	PW=GW	0.230	0.236	0.220	0.234	0.241	0.278	0.284	0.283	0.301	0.288
II	LR	0.240	0.258	0.219	0.263	0.276	0.407	0.437	0.407	0.478	0.457
	PW=GW	0.185	0.190	0.177	0.189	0.192	0.295	0.300	0.299	0.310	0.301
III	LR	0.188	0.198	0.172	0.205	0.203	0.351	0.371	0.345	0.396	0.371
	PW=GW	0.244	0.251	0.233	0.245	0.246	0.363	0.376	0.373	0.395	0.370
						30% censoring					
I	LR	0.215	0.224	0.196	0.219	0.232	0.246	0.257	0.240	0.282	0.270
	PW	0.193	0.198	0.184	0.193	0.198	0.229	0.238	0.231	0.244	0.236
	GW	0.177	0.182	0.174	0.177	0.183	0.210	0.217	0.215	0.231	0.218
II	LR	0.201	0.208	0.179	0.207	0.219	0.264	0.277	0.255	0.304	0.289
	PW	0.173	0.179	0.166	0.172	0.179	0.242	0.253	0.243	0.263	0.253
	GW	0.171	0.173	0.160	0.170	0.173	0.218	0.229	0.223	0.241	0.228
III	LR	0.213	0.223	0.191	0.231	0.230	0.285	0.299	0.276	0.328	0.301
	PW	0.257	0.262	0.245	0.253	0.254	0.319	0.326	0.317	0.337	0.325
	GW	0.249	0.254	0.237	0.245	0.248	0.308	0.314	0.308	0.322	0.305

Table 6. Empirical type I error of independence tests S_{LR1}, S_{LR2}, S_{HP}, S_{SL1} and S_{SL2} with weights $W_1 - W_6$ for
I : exponential(λ_1=1, λ_2=1, 0.75) and
II : logistic(μ_1=100, μ_2=100, 100.5) distributions.

		λ_2=1(μ_2=100)					λ_2=0.75(μ_2=100.5)				
		S_{LR1}	S_{LR2}	S_{HP}	S_{SL1}	S_{SL2}	S_{LR1}	S_{LR2}	S_{HP}	S_{SL1}	S_{SL2}
						no censoring					
I	W_1	0.057	0.028	0.043	0.023	0.046	0.081	0.032	0.041	0.034	0.056
	W_2	0.053	0.024	0.044	0.036	0.056	0.081	0.028	0.041	0.034	0.058
	W_3	0.024	0.015	0.047	0.041	0.047	0.033	0.024	0.046	0.041	0.050
	W_4	0.024	0.016	0.046	0.041	0.047	0.030	0.021	0.041	0.041	0.048
	W_5	0.023	0.017	0.043	0.042	0.044	0.033	0.027	0.049	0.045	0.047
	W_6	0.037	0.018	0.060	0.039	0.049	0.034	0.022	0.040	0.038	0.048
II	W_1	0.073	0.028	0.040	0.027	0.053	0.071	0.029	0.042	0.022	0.052
	W_2	0.063	0.027	0.036	0.032	0.051	0.064	0.022	0.041	0.025	0.049
	W_3	0.022	0.017	0.045	0.039	0.042	0.033	0.024	0.054	0.041	0.045
	W_4	0.026	0.016	0.044	0.041	0.047	0.033	0.022	0.051	0.041	0.050
	W_5	0.022	0.017	0.043	0.042	0.044	0.034	0.028	0.049	0.042	0.046
	W_6	0.034	0.012	0.059	0.038	0.050	0.041	0.031	0.059	0.044	0.052
						30% censoring					
I	W_1	0.063	0.018	0.052	0.034	0.051	0.049	0.022	0.042	0.037	0.057
	W_2	0.055	0.020	0.053	0.038	0.049	0.050	0.017	0.046	0.036	0.052
	W_3	0.029	0.016	0.052	0.038	0.047	0.027	0.021	0.052	0.042	0.051
	W_4	0.026	0.015	0.047	0.045	0.052	0.022	0.016	0.055	0.041	0.051
	W_5	0.029	0.013	0.045	0.038	0.044	0.028	0.015	0.048	0.043	0.052
	W_6	0.043	0.020	0.055	0.042	0.059	0.038	0.019	0.041	0.038	0.052
II	W_1	0.082	0.024	0.049	0.032	0.053	0.070	0.019	0.041	0.030	0.046
	W_2	0.078	0.024	0.057	0.035	0.055	0.056	0.013	0.043	0.029	0.045
	W_3	0.030	0.019	0.061	0.046	0.057	0.020	0.015	0.045	0.040	0.043
	W_4	0.019	0.015	0.057	0.047	0.053	0.015	0.012	0.045	0.038	0.041
	W_5	0.031	0.019	0.059	0.045	0.056	0.027	0.017	0.045	0.041	0.048
	W_6	0.044	0.015	0.059	0.047	0.072	0.042	0.015	0.045	0.025	0.041

Table 7. Empirical power of independence tests S_{LR1}, S_{LR2}, S_{HP}, S_{SL1} and S_{SL2} with weights $W_1 - W_6$ for a paired sample (ρ or τ=0.2) from
I : bivariate exponential(λ_1=1, λ_2=1, 0.75) by Moran's algorithm,
II: Clayton family with exponential margins(λ_1=1, λ_2=1, 0.75) and
III: Clayton family with logistic margins(μ_1=100, μ_2=100, 100.5).

		λ_2=1(μ_2=100)					λ_2=0.75(μ_2=100.5)				
		S_{LR1}	S_{LR2}	S_{HP}	S_{SL1}	S_{SL2}	S_{LR1}	S_{LR2}	S_{HP}	S_{SL1}	S_{SL2}
						no censoring					
I	W_1	0.378	0.233	0.126	0.174	0.253	0.383	0.239	0.128	0.176	0.246
	W_2	0.374	0.238	0.128	0.180	0.251	0.376	0.214	0.142	0.182	0.241
	W_3	0.204	0.170	0.189	0.167	0.179	0.210	0.173	0.188	0.167	0.177
	W_4	0.207	0.164	0.190	0.163	0.178	0.215	0.165	0.189	0.160	0.181
	W_5	0.208	0.173	0.195	0.168	0.185	0.217	0.177	0.185	0.165	0.185
	W_6	0.252	0.178	0.118	0.158	0.186	0.258	0.181	0.118	0.161	0.192
II	W_1	0.789	0.672	0.391	0.623	0.695	0.796	0.635	0.423	0.571	0.676
	W_2	0.805	0.673	0.426	0.633	0.704	0.802	0.616	0.464	0.595	0.686
	W_3	0.557	0.503	0.522	0.514	0.537	0.589	0.512	0.534	0.527	0.545
	W_4	0.551	0.496	0.520	0.517	0.531	0.590	0.514	0.539	0.522	0.545
	W_5	0.569	0.503	0.521	0.505	0.538	0.597	0.519	0.536	0.542	0.563
	W_6	0.681	0.582	0.433	0.557	0.603	0.650	0.560	0.406	0.528	0.580
III	W_1						0.796	0.635	0.422	0.571	0.676
	W_2						0.790	0.630	0.452	0.609	0.676
	W_3			same as II			0.595	0.511	0.536	0.530	0.548
	W_4						0.586	0.508	0.545	0.520	0.546
	W_5						0.597	0.519	0.536	0.542	0.563
	W_6						0.650	0.560	0.405	0.528	0.580
						30% censoring					
I	W_1	0.249	0.133	0.124	0.104	0.145	0.255	0.136	0.126	0.104	0.136
	W_2	0.245	0.138	0.135	0.106	0.145	0.241	0.127	0.141	0.096	0.145
	W_3	0.139	0.106	0.128	0.105	0.123	0.134	0.101	0.133	0.114	0.126
	W_4	0.113	0.082	0.112	0.104	0.120	0.115	0.085	0.122	0.109	0.118
	W_5	0.141	0.108	0.130	0.111	0.128	0.143	0.108	0.136	0.118	0.139
	W_6	0.152	0.102	0.092	0.088	0.109	0.174	0.094	0.086	0.080	0.109
II	W_1	0.549	0.397	0.366	0.329	0.400	0.553	0.399	0.360	0.331	0.427
	W_2	0.550	0.390	0.369	0.338	0.404	0.536	0.353	0.372	0.327	0.392
	W_3	0.357	0.283	0.343	0.314	0.337	0.350	0.291	0.336	0.300	0.335
	W_4	0.281	0.226	0.304	0.276	0.303	0.289	0.231	0.294	0.274	0.294
	W_5	0.365	0.293	0.346	0.323	0.351	0.368	0.306	0.347	0.318	0.348
	W_6	0.426	0.275	0.250	0.238	0.303	0.423	0.296	0.267	0.252	0.309
III	W_1	0.593	0.370	0.303	0.282	0.387	0.563	0.364	0.271	0.275	0.398
	W_2	0.585	0.362	0.308	0.286	0.392	0.540	0.318	0.291	0.287	0.373
	W_3	0.345	0.276	0.329	0.311	0.334	0.324	0.257	0.307	0.279	0.304
	W_4	0.242	0.199	0.270	0.258	0.284	0.236	0.173	0.259	0.250	0.272
	W_5	0.363	0.283	0.329	0.307	0.337	0.342	0.272	0.316	0.291	0.316
	W_6	0.408	0.261	0.197	0.217	0.277	0.421	0.271	0.208	0.229	0.292

V_1, V_2, V_6 and V_7 with V_3, V_4, V_8, and V_9, we can find out that the separated samples estimators produce more significant p-values for this data than the combined sample estimators. In the previous sections, we mentioned test statistics with sum of squares type estimators, V_4, V_5, V_9 and V_{10}. From Table 8, these statistics are also found to have quite small p-values. For the weights functions, as noted in Jung(1998), the logrank weights shows smaller p-values than the Prentice-Wilcoxon weights because the latter puts more weights on early times while larger differences of two survival times are detected for late times.

In summary, this data show some dependence and more differences on late times. Besides, if we recall the type I error problem in V_9, two tests V_8 and V_{10} with the logrank weights seem to be proper for this data.

Next, the data are analysed with the independence tests. All six weights are used for each of 5 statistics. The next table represents the one-sided p-values.

Table 9. One-sided p-values
of independence tests for 6 weights.

	W_1	W_2	W_3	W_4	W_5	W_6
S_{LR1}	0.007	0.009	0.005	0.007	0.005	0.014
S_{LR2}	0.027	0.026	0.009	0.018	0.009	0.078
S_{HP}	0.030	0.041	0.039	0.042	0.039	0.069
S_{SL1}	0.170	0.096	0.053	0.064	0.062	0.250
S_{SL2}	0.082	0.033	0.028	0.034	0.033	0.167

Table 9 shows several interesting things. First of all, the Fleming-Harrington's weight W_6 seems to be irrelevant to this data. Secondly, for all tests except for S_{HP}, the Prentice-Wilcoxon type weights(W_3, W_5) are detecting dependence more efficiently than the logrank type ones(W_1, W_2). This results coincides with that found by Shih and Louis (1996), that this data have stronger association on early times than on late times. As expected from the simulation results, the weights W_2 and W_3 from logrank tests does not appear to have apparent advantage compared to W_1 and W_5. As for the test statistics, the reduced tests, S_{LR1} and S_{LR2}, have much smaller p-values.

In summary, this data have stronger association on early times than on late times, so the Prentice-Wilcoxon weight W_5 seem to be proper. As for the test statistics, though S_{LR1} and S_{LR2} have much smaller p-values, S_{HP} or S_{SL2} may be good if the unstability of the reduced tests in the type 1 errors is concerned.

5. Discussion

For the weighted logrank tests, 'the variance correction terms' play a role to reduce the variance. As a result, the variance estimators with the correction terms as well as the sum of squares type estimators give higher powers. In the previous sections, we also found that the methods based on the separate samples are comparable to the methods on the combined sample. Therefore, the former may be preferred for convenience. In addition, the tests for a paired sample are comparable to those for independent samples with respect to both type I error and power. It means that we need not to concern whether two survival times are independent or not.

With this point of view, the statistic V_9 with all the above factors may be regarded as the best. However, as noted in the simulation results, it shows quite large type I errors in some cases. As a matter of fact, it is also found in some small sample situations that the type I errors of V_4, V_5 and V_{10} also exceed the nominal level. Among them, V_{10} seems to be more stable than V_9. If we consider this point, V_7, V_8 and V_{10} are recommendable. According to the results from simulations and example, we also guess that the properties of the weight functions, known for the independent sample tests, are preserved for paired data. That is, the logrank weights are optimal for the exponential distributions and more sensitive to detect late differences, Prentice - Wilcoxon weights are optimal for the logistic distributions, and Prentice - Wilcoxon and Gehan - Wilcoxon weights are more sensitive to detect early differences.

For the independence tests, the 'variance correction terms', together with the logrank weights, seem to affect the performances of S_{LR1}, S_{LR2} and S_{SL2}. Particularly it is rather severe for S_{LR1}, so the type I errors of S_{LR1} with the logrank weights exceed the nominal level. As shown in the simulation and example, the weights from logrank tests does not appear to surpass their simpler versions. Therefore, the simple weights W_1, W_5 and W_6 will be fine as the weight functions. In summary, the Hsu-Prentice test and modified Shih-Louis test with the simple weights seem to be stable and preferable for the independence tests.

Appendix
proof of equation (2)

To show that

$$\sum_{i=1}^{n}\{\int_0^\infty W(t)/Y_k(t)\ d\hat{M}_{ki}(t)\}^2=\int_0^\infty W^2(t)/Y_k(t)\,(1-\Delta\hat{\Lambda}_k(t))\,d\hat{\Lambda}_k(t), \tag{A.1}$$

we first drop the subscript k for notational simplicity and let $w(t)=W(t)/Y(t)$. Further assume the ordering among observations such that $X_1\le\cdots\le X_n$.

Note that with $d\hat{M}_i(t) = dN_i(t) - Y_i(t)\,dN(t)/Y(t)$, the left hand side of (A.1), say A, can be expressed by

$$\begin{aligned} A &= \sum_i [\int w(t)\{dN_i(t) - Y_i(t)\,dN(t)/Y(t)\}]^2 \\ &= \sum_i [w(X_i)\delta_i - \sum_j w(X_j)Y_i(X_j)\delta_j/Y(X_j)]^2 \\ &= B - 2C + D, \end{aligned}$$

where

$$B = \sum_i w^2(X_i)\delta_i,$$

$$C = \sum_i \sum_j w(X_i)w(X_j)Y_i(X_j)\delta_i\delta_j/Y(X_j)$$

and

$$D = \sum_i \sum_j \sum_k w(X_j)w(X_k)Y_i(X_j)Y_i(X_k)\delta_j\delta_k/\{Y(X_j)Y(X_k)\}.$$

Since $Y_i(X_j) = I(X_i \geq X_j)$, under the ascending ordering of the observations, C can be written by $C = C_1 + C_2 + C_3$ where

$$C_1 = \sum_i w^2(X_i)\delta_i/Y(X_i),$$

$$C_2 = \sum\sum_{\{(i,j):\, i>j\}} w(X_i)w(X_j)\delta_i\delta_j/Y(X_j),$$

$$C_3 = \sum\sum_{\{(i,j):\, i<j,\, X_i = X_j\}} w^2(X_i)\delta_i\delta_j/Y(X_i).$$

Note that when there are no tie in observations, C_3 becomes zero.

Similarly, D is also split into three terms, $D = D_1 + D_2 + D_3$, and these terms are simplified from the relation $Y_i(X_j)Y_i(X_k) = Y_i(X_j)$ $(j \geq k)$ as follows;

$$D_1 = \sum_i \sum_j w^2(X_j)Y_i(X_j)\delta_j/Y^2(X_j) = \sum_j w^2(X_j)\delta_j/Y(X_j),$$

$$\begin{aligned} D_2 &= \sum_i \sum\sum_{\{(j,k):\, j>k\}} w(X_j)w(X_k)Y_i(X_j)\delta_j\delta_k/\{Y(X_j)Y(X_k)\} \\ &= \sum\sum_{\{(j,k):\, j>k\}} w(X_j)w(X_k)\delta_j\delta_k/Y(X_k), \\ D_3 &= \sum_i \sum\sum_{\{(j,k):\, j<k\}} w(X_j)w(X_k)Y_i(X_k)\delta_j\delta_k/\{Y(X_j)Y(X_k)\} \\ &= \sum\sum_{\{(j,k):\, j>k\}} w(X_j)w(X_k)\delta_j\delta_k/Y(X_j). \end{aligned}$$

Since $C_1 = D_1$ and $C_2 = D_2 = D_3$, we have $A = B - (C_1 + 2C_3)$.

On the other hand, the right hand side of (A.1), A', can be written by

$$\begin{aligned} A' &= \int w^2(t)dN_i(t) - \int w^2(t)\Delta N(t)\,dN(t)/Y(t) \\ &= \sum_i w^2(X_i)\delta_i - \sum_i \sum_j w^2(X_i)\Delta N_j(X_i)\delta_i/Y(X_i) \\ &= B - C' \end{aligned}$$

From the definition of the counting process, $\Delta N_j(X_i)$ takes the values of δ_i if $i = j$, δ_j if

$i \neq j$ and $X_i = X_j$, and 0 otherwise. Hence, C is equal to $C_1 + 2C_3$, resulting in (A.1).

References

[1] Batchelor, J. R. and Hackett, M. (1970). HL-A Matching in Treatment of Burned Patients with Skin Allografts. *Lancet*, 2, 581-583.

[2] Clayton, D. G. (1978). A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika,* 65, 141-151.

[3] Dabrowska, D. M. (1989). Rank Tests for Matched Pair Experiments with Censored Data. *Journal of Multivaiate Analysis*, 28, 88-114.

[4] Dabrowska, D. M. (1990). Signed-Rank Tests for Censored Matched Pairs. *Journal of the American Statistical Association,* 85, 478-485.

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

[6] Gehan, E. A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrarily Singly Censored Samples, *Biometrika,* 52, 203-223.

[7] Harrington, D. P. and Fleming, F. A. (1982). A Class of Rank Test Procedures for Censored Survival Data. *Biometrika*, 69, 133-43.

[8] Hsu, L. and Prentice, R. L. (1996). A Generalisation of the Mantel-Haenszel Test to Bivariate Failure Time Data. *Biometrika*, 83, 4, 905-911.

[9] Jung, S. (1998). Rank Test for Matched Survival Data. *Lifetime Data Analysis.* (to appear)

[10] Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimator from Incomplete Observations. *Journal of the American Statistical Association*, 53, 457-481.

[11] Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, New York.

[12] Lam, F. C. and Longnecker, M. T. (1983). A Modified Wilcoxon Rank Sum Test for Paired Data. *Biometrika*, 70, 510-513.

[13] Mantel, N. and Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of National Cancer Institute,* 22, 719-748.

[14] Moran, P. A. P. (1967). Testing for Correlation between Non-negative Variates. *Biometrika,* 54, 385-394.

[15] Nelson, W. (1969). Hazard Plotting for Incomplete Failure Data. *Journal of Quality Technology*, 1, 27-52.

[16] Oakes, D. (1982). A Model for Association in Bivariate Survival Data. *Journal of the Royal Statistical Society, B,* 44, 414-422.

[17] O'brien, P. C. and Fleming, T. R. (1987). A Paired Prentice-Wilcoxon Test for Censored Paired Data. *Biometrika,* 43, 169-180.

[18] Prentice, R. L. (1978). Linear Rank Tests with Right Censored Data. *Biometrika*, 65, 167-179.

[19] Prentice, R. L. and Cai, J. (1992). Covariance and Survivor Function Estimation Using Censored Multivariate Failure Time Data. *Biometrika*, 79, 495-512.

[20] Shih, J. H. and Louis, T. A. (1996). Tests of Independence for Bivariate Survival Data. *Biometrics*, 52, 1440-1449.

[21] Tarone, R. E. and Ware, J. (1977). On Distribution-free Tests for Equality of Survival Distributions. *Biometrika*, 64, 156-60.

[22] Wei, L. J. (1980). A Generalized Gehan and Gilbert Test for Paired Observations that are Subject to Arbitrary Right Censoring. *Journal of the American Statistical Association*, 75, 634-637.

[23] Woolson, R. F., and Lachenbruch, P. A. (1980). Rank Tests for Censored Matched Pairs. *Biometrika*, 67, 597-606.