

A Test for Independent Censoring Using Correlated Survival Data

Xuelin Huang,^{1,*} Robert A. Wolfe² and Chengcheng Hu³

¹Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center
Box 447, 1515 Holcombe Boulevard, Houston, TX 77030, U.S.A.

²Department of Biostatistics, The University of Michigan
Ann Arbor, MI 48109-2029, U.S.A.

³Department of Biostatistics, Harvard University
Boston, MA 02115, U.S.A.

October 4, 2002

SUMMARY

It is common in survival studies that censoring is dependent on potential failure. When subjects in the study are independent each other, the dependence between failure and censoring is not identifiable. However, this article shows that, when subjects are correlated, they can serve as replicates for each other under certain circumstances. With replicates, dependent censoring is identifiable. A simple test using martingale residuals is developed. Simulation studies evaluate the performance of the test under different situations. For illustration, the test is applied to a data set for kidney disease patients.

KEY WORDS: Clustered data; Competing risks; Correlation; Dependent censoring; Martingale residuals; Survival analysis.

* *email*: xlhuang@mdanderson.org

1. Introduction

It has long been an important topic in survival analysis to identify dependent censoring. In biomedical studies, patient survival time can be censored by many causes, for example, patient withdrawal, changing treatment, or the end of the study. While censoring by the end of the study is often independent of failure time, it is commonly true that patient voluntary withdrawal and changing treatment are correlated with failure time (Lagakos, 1979). That is to say, both independent and dependent censoring events are commonly present in survival studies. However, if all the subjects in the study are independent of each other, then neither the existence nor the degree of the dependence between failure and censoring is identifiable (Tsiatis, 1975). This is because, with right censored data collected from independent subjects, only one of each pair of censoring and survival times is observed, so the correlation between them cannot be evaluated. Consequently, all the aforementioned events are treated as independent censoring without distinction in common statistical practice. This will usually result in biased inference (Klein and Moeschberger, 1987).

Various models have been proposed to deal with dependent censoring, see Emoto and Matthews (1990), Robins and Rotnitzky (1992), Zheng and Klein (1995), Lin et al. (1996), and references therein. These models are usually complicated. It would be desirable to do a simple test to determine the presence of dependent censoring before jumping to those complicated models. Lee and Wolfe (1998) proposed a test for independent censoring. They used data collected after censoring. However, data after censoring are usually not available. This article offers another test for independent censoring. It does not need data collected after censoring. It uses correlated data, or more precisely, data collected from subjects from different clusters. Subjects in the same cluster are correlated. They are assumed to share common proneness to failure and censoring after adjustment for covariate effects. This assumption make subjects within a cluster replicates for each other, so that a test for independent censoring is possible.

This assumption may not be true, but can be tested by the methods proposed by Gray (1995) and Commenges and Andersen (1995). They used martingale residuals to test for the presence of clustering in survival data. This article is motivated by their approaches.

We use the following example to illustrate the idea of using clustered data to identify dependent censoring. In the data set for kidney disease patients analyzed in section 4 of this article, there are 10,290 patients from 152 kidney centers. In this example, each kidney center is a cluster, and patients are subjects in clusters. During the study, some patients died, some patients withdrew, and the remaining majority were still alive at the end of the study. We calculate the percentage dead and percentage withdrawn for each center. The correlation coefficient between these two groups of percentages is 0.25. This descriptive approach gives some evidence that withdrawal is positively correlated with failure across clusters. It suggests that the relationship between failure and censoring can be assessed by clustered data. However, this simple correlation ignores the lengths of survival and censoring times, and is not adjusted by covariates. To account for these factors, we replace the percentages dead and withdrawn respectively by martingale residuals from survival analysis models for death and withdrawal. This makes the proposed test more sophisticated than this naive percentage correlation approach. Nevertheless, the proposed test is still easy to implement.

The test is developed in section 2. In this section, we first introduce the basic idea and approach of the test. Then we show that an adjustment is necessary due to the estimation of the cumulative hazard functions. Simulation study results are reported in section 3. Three models are used to generate different dependence structures between failure and censoring times. The performance of the proposed test is evaluated under these different situations. Section 4 illustrates the use of the test by the above example data set for kidney patients. Section 5 discusses the advantages and limits of the proposed test. In the exploration of the testing

methods, we find a new property of the Nelson-Aalen estimator. This property not only affects our proposal, but may also have its own independent value. The models used in the simulation studies are also useful for analyzing correlated data and dependent censoring.

2. Methods

2.1 Notation and Assumptions

Suppose there are two types of censoring, namely drop-out and administrative censoring. Drop-out could be any type of potentially dependent censoring, such as initiating a non-randomized therapy in the middle of a randomized trial. Administrative censoring, such as that caused by the end of the study, is assumed to be random.

Denote the failure, drop-out and administrative censoring times for a subject by T , C and Q respectively. Let $X = \min(T, C, Q)$. Define $\Delta^{(T)} = 1$ if $X = T$ and $\Delta^{(T)} = 0$ otherwise. Similarly define $\Delta^{(C)}$ and $\Delta^{(Q)}$. They satisfy $\Delta^{(T)} + \Delta^{(C)} + \Delta^{(Q)} = 1$. The administrative censoring time Q is assumed to be independent of both T and C , and will not be discussed in detail. From now on, “censoring” will be used to mean drop-out only. The distributions of T and C are assumed to be continuous. We discuss two situations, namely, with and without covariates. When there are covariates, denote by $Z^{(T)}$ and $Z^{(C)}$ vectors of covariates associated with failure and censoring respectively. They may be completely distinct or overlapping or even identical. They are assumed to be time-independent in this article, but the method can also be applied to data sets with external time-dependent covariates (Kalbfleisch and Prentice, 1980). When there are internal time-dependent covariates, the approach proposed by Robins and Rotnitzky (1992) can be applied. They did not consider the issue of correlation due to clustering. The method in this article can be used as a supplement to their approach.

Suppose there are m clusters and n_i subjects in the i^{th} cluster. The total sample size is

then $N = \sum_{i=1}^m n_i$. We first assume equal sample size $n_i = n$ for $i = 1, \dots, m$, then discuss the situation of unequal cluster sizes at the end of § 2.4. Clusters are assumed to be independent each other. However, within each cluster, $T_{ij}, j = 1, \dots, n$, are correlated with each other. It is also assumed that the joint distribution of failure times, censoring times, and covariates (if any) are identical across clusters.

2.2 General Methods

The basic idea to detect the dependence between failure and censoring is to use the correlation between two groups of martingale residuals: one defined by failure and the other by censoring. Denote the marginal cumulative hazard functions at time u for T_{ij} and C_{ij} respectively by $\Lambda_{ij}^{(T)}(u)$ and $\Lambda_{ij}^{(C)}(u)$. Then, the martingale residuals are the following.

$$\begin{aligned} R_{ij}^{(T)} &= \Delta_{ij}^{(T)} - \Lambda_{ij}^{(T)}(X_{ij}), \\ R_{ij}^{(C)} &= \Delta_{ij}^{(C)} - \Lambda_{ij}^{(C)}(X_{ij}). \end{aligned}$$

Note that $E(R_{ij}^{(T)}) = 0$ and $E(R_{ij}^{(C)}) = 0$. To take advantage of the clustering, let $R_i^{(T)} = \frac{1}{n} \sum_{j=1}^n R_{ij}^{(T)}$ and similarly define $R_i^{(C)}$. They also have expectation zero.

We conjecture that the Pearson correlation coefficient,

$$\rho = \text{Corr}(R_i^{(T)}, R_i^{(C)}),$$

can be used to detect the correlation between failure and censoring. It will be shown later that, under the assumption of independent censoring and some other conditions, we have $\rho = 0$. Using the empirical estimator $\hat{\rho}$ for ρ , A test for $\rho = 0$ is as following.

$$t_1 = \frac{\hat{\rho}}{\sqrt{\frac{1-\hat{\rho}^2}{m-2}}} \sim t_{m-2} \quad \text{when } \rho = 0. \quad (1)$$

This t -test can be found, for example, in Lehmann (1991). It is derived under the bi-variate normal distribution assumption. Here the assumption of the test is violated. $R_i^{(T)}$ and $R_i^{(C)}$

are not independent (It is easy to see that $R_{ij}^{(T)}$ and $R_{ij}^{(C)}$ cannot be both positive), and their distributions are skewed.

Another approach is to test $\mu \equiv E(R_{i\cdot}^{(T)} R_{i\cdot}^{(C)}) = 0$. In this case, $R_{i\cdot}^{(T)} R_{i\cdot}^{(C)}$ is viewed as a single variable and a generic t -test can be used. That is to say, let

$$\begin{aligned}\hat{\mu} &= \frac{1}{m} \sum_{i=1}^m R_{i\cdot}^{(T)} R_{i\cdot}^{(C)}, \\ \hat{\sigma}^2 &= \frac{1}{m-1} \sum_{i=1}^m (R_{i\cdot}^{(T)} R_{i\cdot}^{(C)} - \hat{\mu})^2.\end{aligned}$$

Then, when $\mu = 0$, we have that

$$t_2 = \frac{\sqrt{m}\hat{\mu}}{\hat{\sigma}} \sim t_{m-1}. \quad (2)$$

The underlying assumption is that $R_{i\cdot}^{(T)} R_{i\cdot}^{(C)}$ has a normal distribution with mean μ and variance σ^2 . By this method, the dependence between $R_{ij}^{(T)}$ and $R_{ij}^{(C)}$ becomes irrelevant. The variance estimator for $\hat{\mu}$ is also more robust than that for $\hat{\rho}$ in (1). However, this is not our final proposal. A problem arises in the estimation of cumulative hazards. In §2.4, we will adjust our test statistic to take care the problem.

2.3 Estimation of Cumulative Hazard Functions

The hazard functions used in the computation of residuals are not known, they must be estimated. We will ignore the cluster structure to do the hazard estimation. If there are not covariates in the data set, we can use Nelson-Aalen estimator for hazard functions (see, for example, Fleming and Harrington (1991)). Ying and Wei (1994) showed that the estimated hazard functions still converge to their true values, even if the within-cluster correlation is ignored in the estimation. However, the following algebraic property of Nelson-Aalen estimator complicates our testing method, and makes adjustment necessary.

THEOREM 1. *When hazard functions are estimated by Nelson-Aalen estimators, we have that*

$$\sum_{i=1}^m \sum_{j=1}^n \hat{R}_{ij}^{(T)} \hat{R}_{ij}^{(C)} = 0.$$

Proof. See Appendix A.

We will show in the next subsection how this property affects our test. Here we would like to emphasize that this property is purely algebraic. It holds no matter whether T and C are actually independent or not, nor whether subjects are correlated or not. Note that the cluster structure is ignored here. Therefore, the above theorem simply means the sum of the products of failure and censoring residuals is zero if hazards are estimated by Nelson-Aalen estimators. That is to say, beside the well-known property that the sum of estimated residuals is zero (i.e., $\sum_{i=1}^m \sum_{j=1}^n \hat{R}_{ij}^{(T)} = 0$ and $\sum_{i=1}^m \sum_{j=1}^n \hat{R}_{ij}^{(C)} = 0$), we found another property of Nelson-Aalen estimator. This discovery may have its value independent of this article.

When there are covariates in the data set, we may assume the following marginal proportional hazard models for failure and censoring respectively.

$$\begin{cases} \lambda^{(T)}(u|Z^{(T)}, Z^{(C)}) &= \lambda_0^{(T)}(u) \exp(\beta^{(T)'} Z^{(T)}), \\ \lambda^{(C)}(u|Z^{(T)}, Z^{(C)}) &= \lambda_0^{(C)}(u) \exp(\beta^{(C)'} Z^{(C)}). \end{cases} \quad (3)$$

Lee et al. (1992) showed that by using the common Cox (1972) score equation for $\beta^{(T)}$ and $\beta^{(C)}$, and Breslow (1972) estimator for $\lambda^{(T)}$ and $\lambda^{(C)}$, the resulting parameter estimates are still consistent, and their covariance matrices can be gotten by using Sandwich estimators (White, 1982). When there are covariates, the sum of residual products as above may not be exactly equal to zero, but is always very close to zero (by simulations not shown). The implication of these properties to our testing procedure is described in the next subsection.

2.4 Adjustment

For each cluster, define

$$U_i = \sum_{j=1}^n \sum_{k \neq j} \hat{R}_{ij}^{(T)} \hat{R}_{ik}^{(C)}, \quad (4)$$

$$V_i = \sum_{j=1}^n \hat{R}_{ij}^{(T)} \hat{R}_{ij}^{(C)}. \quad (5)$$

Then $\hat{R}_i^{(T)} \hat{R}_i^{(C)} = \frac{1}{n^2}(U_i + V_i)$. The t -test in (2) is

$$t_2 = \frac{\bar{U} + \bar{V}}{\sqrt{\widehat{Var}(\bar{U} + \bar{V})}}, \quad (6)$$

where $\bar{U} = \frac{1}{m} \sum_{i=1}^m U_i$ and similarly for \bar{V} . By Theorem 1, we always have $\bar{V} = 0$. That is to say, those terms V_i make no contribution to the numerator of (6), but contributes to the variance estimator in the denominator of (6). Therefore, the variance is inflated and t_2 gives conservative results. Our simulation results (not shown) verified this point. To avoid the problem caused by V_i 's, the statistic t_2 is modified into

$$t_u = \frac{\bar{U}}{\sqrt{\widehat{Var}(\bar{U})}} = \frac{\sqrt{m}\bar{U}}{\sqrt{\frac{1}{m-1} \sum_{i=1}^m (U_i - \bar{U})^2}}. \quad (7)$$

The performance of t_u will be assessed by simulation studies in the next section.

If cluster sizes are all equal to one, i.e., there are no clusters, then we have V_i 's only, and do not have U_i 's. Thus we cannot test the independence between T and C by the above method. Actually the assumption of independent censoring cannot be tested in this case (Tsiatis, 1975).

When cluster sizes are not equal to each other, weighted version of (7) may be more appropriate. For example, if we assign weight $w_i = \frac{n_i}{N}$ to cluster i , then the resulting test is

following.

$$\begin{aligned}
\bar{U}_w &= \sum_{i=1}^m w_i U_i, \\
\hat{s}_w &= \sum_{i=1}^m w_i (U_i - \bar{U}_w)^2, \\
t_w &= \frac{\bar{U}_w}{\hat{s}_w \sqrt{\sum_{i=1}^m w_i^2}}.
\end{aligned} \tag{8}$$

The optimal weights have not yet been investigated. However, we suspect that the optimal weights depend on the within cluster correlation, which is not specified here.

2.5 A Model for Clustered Data with Dependent Censoring

We will use the model below to describe a dependence structure between T and C , and to evaluate the performance of the proposed test. This model is proposed by Huang and Wolfe (2002). For simplicity, we consider the case of without covariates first. The model assumes that, conditional on clustering effect A_i , the hazard functions for T_{ij} and C_{ij} at time u are $h_{ij}^{(T)}(u)$ and $h_{ij}^{(C)}(u)$ respectively as below.

$$\left\{ \begin{array}{l} A_i \sim N(0, \sigma^2), i.i.d., i = 1, \dots, m, \\ h_{ij}^{(T)}(u | Z_{ij}^{(T)}, Z_{ij}^{(C)}, A_i) = h_0^{(T)}(u) \exp(A_i), \\ h_{ij}^{(C)}(u | Z_{ij}^{(T)}, Z_{ij}^{(C)}, A_i) = h_0^{(C)}(u) \exp(\gamma A_i), \\ i = 1, \dots, m, j = 1, \dots, n_i. \end{array} \right. \tag{9}$$

In this model, the unobserved A_i is shared by all subjects in cluster i , and e^{A_i} is usually called frailty for cluster i . Conditional on A_i , failure time T_{ij} and censoring time C_{ij} are assumed to be independent. However, note that A_i affects not only failure, but also censoring risks. The effect of A_i on censoring risk is assumed to have a simple format $e^{\gamma A_i}$. Then, unconditional on A_i , the two event times T_{ij} and C_{ij} are correlated when $\gamma \neq 0$. When $\gamma = 0$, they are independent. Moreover, in the case of $\gamma = 0$, censoring times within a cluster $C_{ij}, j = 1, \dots, n_i$, are no longer correlated, even though failure times within a cluster $T_{ij}, j = 1, \dots, n_i$, are still correlated with each other. That is to say, this model assumes that, when censoring is uncorrelated with failure,

it is caused by some random mechanism.

Under this model, we have the following theorem.

THEOREM 2. $E(R_i^{(T)} R_i^{(C)}) = 0$ under the null hypothesis $H_0: \gamma = 0$ in (9).

Proof. See Appendix B.

Under the above model, to test the independence between T and C , we can test $H_0 : \gamma = 0$. One approach to do the test is to fit the model, get parameter estimate for γ and its standard error, then do a Wald test. However, it is relatively complicated to fit the above model (Huang and Wolfe (2002) provided an approach). As mentioned in the introduction section, our goal is to develop a simple test for independent censoring, so that we can apply the test before fitting complicated models for dependent censoring.

The above model describes a particular type of dependent censoring. We will present more models in the simulations section, and show that the proposed test can be applied to more general settings.

3. Simulation Study

The performance of the test by the statistic t_u in (7) is assessed under different situations. Data are generated first by the univariate log-frailty model in (9), then a bi-variate log-normal frailty model and a positive stable frailty model, both of which are to be presented later in this section. For the univariate frailty model, we report the results for cases with and without covariates, combined with different settings of censoring proportions. The results are similar under different settings. Therefore, for other models, to save space, only the results for the case with covariates and a fixed censoring proportion are reported.

3.1 *A Univariate Log-normal Frailty Model without Covariates*

The model in (9) is used to generate data. Since the distributions of T and C are skewed, the Pearson correlation between $\log(T)$ and $\log(C)$ (denoted by r) is a better index than that between T and C (Lindeboom and Van Den Berg, 1994). Thus we use r to measure the degree of dependence between T and C . By choosing $\sigma = 0.5$ and 1.0 , combining $\gamma = \pm 1$, we get approximately $r = \pm 0.14$ and ± 0.37 . When $\gamma = 0$, we have $r = 0$. In this case, we use $\sigma = 0.5$. We also generate an administrative censoring time Q for each subject by a uniform distribution on $(0, a)$, independent of both T and C . Various censoring rates are achieved by appropriately choosing a, h_1 and h_2 . The observed survival data are then the minimum of T, C , and Q , and an indicator showing which one is observed. In this setting, both independent and dependent censoring are present. We use t_u in (7) to detect the dependence between T and C . The results are summarized in the top half of Table 1.

[Table 1 about here.]

Simulation results suggest that the test is valid, even with moderate sample sizes. When H_0 is true, the probability that it is rejected is close to the nominal 5% level. When H_0 is not true, the power to reject H_0 is satisfactory. The stronger the dependence between T and C , the more power to detect the dependence. Given cluster size, increasing number of clusters gives more power. Given number of clusters, larger cluster sizes also result in greater power. Higher censoring rates within a reasonable range help detecting the correlation between failure and censoring.

3.2 *A Univariate Log-normal Frailty Model with Covariates*

To illustrate, we include two covariates in the models. Suppose each subject has probability 0.5 to receive a treatment (TR=1) and probability 0.5 to be a control (TR=0). Subject age after being centered has an uniform($-10, 10$) distribution. All subjects are independent to each

other with regard to covariate distributions. Specifically, the model is

$$\begin{cases} A_i \sim N(0, \sigma^2), i.i.d., i = 1, \dots, m, \\ h_{ij}^{(T)}(u|AGE_{ij}, TR_{ij}, A_i) = h_1 u \exp(0.1 AGE_{ij} - 1.4 TR_{ij} + A_i), \\ h_{ij}^{(C)}(u|AGE_{ij}, TR_{ij}, A_i) = h_2 \exp(0.2 AGE_{ij} + 1.2 TR_{ij} + \gamma A_i). \end{cases} \quad (10)$$

We use the same combinations of σ and γ as in §3.1. In this case, the correlation coefficient r between $\log(T)$ and $\log(C)$ is understood as conditional on covariates. Again the test specified in (7) is applied and the results are summarized in the bottom half of Table 1. The test performs also well in the cases when covariates are present. The results have the same patterns as in the case of no covariates.

3.3 A Bivariate Log-normal Frailty Model

By this model, we allow censoring times within a cluster to be correlated even in the case T and C are independent. Here we use the same covariate setting as in §3.2. The model is specified below.

$$\begin{cases} \begin{pmatrix} A_i \\ B_i \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix} \right) \\ h_{ij}^{(T)}(u|AGE_{ij}, TR_{ij}, A_i, B_i) = h_1 u \exp(0.05 AGE_{ij} - 1.0 TR_{ij} + A_i) \\ h_{ij}^{(C)}(u|AGE_{ij}, TR_{ij}, A_i, B_i) = h_2 \exp(0.06 AGE_{ij} + 0.8 TR_{ij} + B_i) \end{cases} \quad (11)$$

We choose $\sigma_a = \sigma_b = 0.9$, and let $\rho = 0, \pm 0.4, \pm 0.8$. That result in $r = 0, \pm 0.13, \pm 0.26$ (the correlation between $\log(T)$ and $\log(C)$). The testing results are reported in the top half of Table 2.

[Table 2 about here.]

We used conditional proportional hazard models as above to generate data, and use the marginal proportional model in (3) to estimate cumulative marginal hazard functions. However, when the frailty distribution is (univariate or bivariate) log-normal, the marginal model is not a proportional hazard model. The Model in (3) is mis-specified. We may view the

mis-specified model as an approximate for the true models. In practice, proportional hazard models like the one in (3) are often chosen to do the parameter estimation, even if they are very likely to be mis-specified. It is desirable to evaluate the performance of the proposed test in such a situation. However, we still would like to use a correct marginal model to estimate hazard functions, and evaluate the performance of the proposed testing method in such an ideal situation.

3.4 A Positive Stable Frailty Model

Both conditional and marginal models can be proportional hazard models when the distribution of frailty is positive stable (Hougaard, 1986). A generic random variable X is said to have a positive stable distribution with parameter $\theta \in (0, 1)$, denoted by $PS(\theta)$, if its Laplace transform is $E\{\exp(-sX)\} = \exp(-s^\theta)$, $s > 0$. Suppose $0 < \theta_1 < \theta_3 < 1$ and $0 < \theta_2 < \theta_3 < 1$. The following model is modified from a multivariate distribution proposed by Joe (1993).

$$\left\{ \begin{array}{l} U_i \sim PS\left(\frac{\theta_1}{\theta_3}\right) \\ V_i \sim PS\left(\frac{\theta_2}{\theta_3}\right) \\ W_i \sim PS(\theta_3) \\ h_{ij}^{(T)}(u|Z_{ij}^{(T)}, Z_{ij}^{(C)}, U_i, V_i, W_i) = h_0^{(T)}(u) U_i W_i^{\frac{\theta_3}{\theta_1}} \exp(\xi^{(T)} Z_{ij}^{(T)}), \quad j = 1, \dots, n_i \\ h_{ij}^{(C)}(u|Z_{ij}^{(T)}, Z_{ij}^{(C)}, U_i, V_i, W_i) = h_0^{(C)}(u) V_i W_i^{\frac{\theta_3}{\theta_2}} \exp(\xi^{(C)} Z_{ij}^{(C)}), \quad j = 1, \dots, n_i \end{array} \right. \quad (12)$$

In this model, the unobserved frailties U_i , V_i and W_i are shared by all subjects in cluster i . These three frailty variables are assumed to be independent each other. Conditional on these three frailty variables, $T_{ij}, j = 1, \dots, n_i$ are assumed to be independent each other, so are $C_{ij}, j = 1, \dots, n_i$; moreover, T_{ij} and C_{ik} are assumed to be independent for any $1 \leq j, k \leq n_i$. However, note that W_i affects not only failure, but also censoring risks. Therefore, unconditional on W_i , the two event times T_{ij} and C_{ij} are correlated when $\theta_3 < 1$. When $\theta_3 \rightarrow 1$, T_{ij} and C_{ij} becomes independent. When θ_3 gets close to zero, the correlation between T_{ij} and C_{ij} increases. We choose $\theta_3 = 1.0, 0.8, 0.6$, and 0.4 , which result in $r = 0, 0.34, 0.62$ and 0.83 . This model can produce positive correlation only. In a similar fashion, $\frac{\theta_1}{\theta_3}$ and $\frac{\theta_2}{\theta_3}$ control the

within-cluster correlation between failure times and and that between censoring times respectively. We let $\theta_1 = \theta_2 = 0.8\theta_3$.

By some computation, it is easy to see that the marginal cumulative hazard functions given by the above model are follows.

$$\begin{cases} \Lambda^{(T)}(u|Z^{(T)}, Z^{(C)}) &= \{H_0^{(T)}(u)\}^{\theta_1} \exp(\theta_1 \xi^{(T)} Z^{(T)}), \\ \Lambda^{(C)}(u|Z^{(T)}, Z^{(C)}) &= \{H_0^{(C)}(u)\}^{\theta_2} \exp(\theta_2 \xi^{(C)} Z^{(C)}). \end{cases} \quad (13)$$

Here $H_0^{(T)}(u) = \int_0^u h_0^{(T)}(v) dv$, and similarly for $H_0^{(C)}(u)$. By using positive stable frailty distribution, both conditional and marginal hazards have nice closed forms. That enables us to use the correct model to estimate the cumulative marginal hazard functions. We check the performance of the proposed test in the setting of this model. The results are reported in the bottom half of Table 2.

3.5 Summary of Simulation Results

When the correlation between $\log(T)$ and $\log(C)$ is equal to zero, the rejecting rates are almost always less than the nominal level 5%. While this shows the test is valid, it also shows that the test is conservative. This is due to the fact by Theorem 1, which says that the randomness of $\sum_{i=1}^m \sum_{j=1}^n R_{ij}^{(T)} R_{ij}^{(C)}$ is reduced to zero when it is estimated by $\sum_{i=1}^m \sum_{j=1}^n \hat{R}_{ij}^{(T)} \hat{R}_{ij}^{(C)}$. To understand why, image an extreme situation where subjects in each cluster are perfectly correlated, that is, they have the same failure and censoring times. Recall that in order to avoid the problem caused by the fact in Theorem 1, we decompose $\hat{R}_{i\cdot}^{(T)} \hat{R}_{i\cdot}^{(C)}$ into $U_i = \sum_{j=1}^n \sum_{k \neq j} \hat{R}_{ij}^{(T)} \hat{R}_{ik}^{(C)}$ and $V_i = \sum_{j=1}^n \hat{R}_{ij}^{(T)} \hat{R}_{ij}^{(C)}$, and use U_i 's only to do the test. However, in the above situation of perfect within-cluster correlation, the sum of U_i 's will also be zero. The randomness of both the sum of U_i 's and the sum of V_i 's is lost completely. When the within-cluster correlation is not that high, the sum of U_i 's is not zero, but the randomness is still reduced due to the estimation of hazard functions. That results in conservative testing results. Nevertheless, we can see that,

under a reasonable range of within-cluster correlation, we can apply the proposed method to test the assumption of independent censoring. As shown by simulation studies, the degree of being conservative is tolerable, and the power of the proposed test is still satisfactory.

If we do not consider the fact by theorem 1 and use t_2 in (2) to do the tests, the results will be much more conservative than that by (7), and the loss of power will be more severe, especially when cluster sizes are small. If we use the test by t_1 in (1), which uses normality assumption to estimate variances, the results will be a little more conservative than that using t_2 , which uses a robust variance estimator. The simulation results for t_1 and t_2 are not shown.

4. Example: Mortality at Kidney Dialysis Centers

In this section, we apply the test to a data set for kidney disease patients. End Stage Renal Disease is a chronic condition of total and irreversible kidney failure. Dialysis is the treatment for the patients before they receive kidney transplantation. Data were collected by the Dialysis Outcomes and Practice Patterns Study (DOPPS) during 1996 to 1999 in seven countries (Young et al., 2000). Here we use only the subset of the data for the United States. It contains 10,290 patients from 152 randomly sampled dialysis facilities. The number of patients in a facility ranged from 21 to 124. During the study, 3,188 (31%) patients died, 424 patients (4%) withdrew from the study (e.g., transferred to a dialysis facility not in our sample), the other 6,678 patients were alive and remaining in the study (some of them received kidney transplantation). It is suspected that the most common reason for withdrawal is worsened health status. Thus withdrawal is likely to be correlated with failure. We assume here that receiving transplantation and the end of the study are independent censoring.

Facility level mortality rates range from 0% to 48%. Withdrawal rates range from 0% to 20%. They are positively correlated (Pearson correlation coefficient = 0.25). This gives some

evidence that withdrawal and failure are correlated. However, this raw correlation coefficient ignores the lengths of survival times and is not adjusted by covariates. The method developed in section 2 is used to take these factors into account. Marginal proportional hazard models as in (3) are fitted. Four important patient level covariates in kidney studies (Wolfe, 1994) are included in each model. They are age (in years), race (black=1, other=0), gender (male=1, female=0), and diabetes (diabetes=1, no diabetes=0). In this example, patients are clustered by facilities. The survival outcomes of the patients in the same facility should have some similarities since they shared the same service. The facility level residual correlation between failure and withdrawal is 0.20 (with 95% confidence interval = (0.04, 0.36)). Using (7), $t_u = 2.63$. Using (8), $t_w = 1.71$. The results are not quite consistent. However, all suggest evidences that withdrawal is positively correlated with failure, after being adjusted by covariates. If we naively assume withdrawal as independent censoring, the inference is very likely to be biased.

5. Discussion

One of the advantages of the test developed in this article is that it is very easy to implement by standard software such as SAS. There is no complicated programming involved. The other advantage of this test is that the data it uses are very common in real practice.

Essentially we used the similarity between subjects in the same cluster to do the test. There are situations where subjects in the same cluster compete for resources (for example, litter mates), and within-cluster correlation is negative. Our test cannot be applied in such a situation. The test can be applied when a shared frailty model is appropriate and clustering effect is significant. If this is in question, the tests provided by Commenges and Andersen (1995), Gray (1995), and Andersen et al. (1999) can be used.

When some covariates affect failure time, but are not included in the regression model, then

censoring is dependent (Kalbfleisch and Prentice, 1980). If these covariates are correlated with clustering, then their effects can be modeled by frailty models as in the simulation section. The proposed test can be used to detect this type of dependent censoring. When these excluded covariates are independent of clustering, the proposed test have no power to detect the dependence between failure and censoring. This is a case of no significant clustering effect.

There are some technical points worth mentioning. First, we use independence between failure and censoring as independent censoring, even though formally speaking the latter is somewhat weaker. This is unlikely to make any difference in real practice. Second, we use the name martingale residual in this article. However, due to the within cluster correlation, the martingale residuals may or may not be martingales, depending on the respecting filtration. Third, for convenience, we considered continuous failure time distributions only in this article, but the test shall be applicable when the distribution is discrete.

If the test shows correlation between failure and censoring, then the next problem is how to adjust it and get unbiased parameter estimates. To solve this problem, Huang and Wolfe (2002) proposed a frailty model for informative censoring, which is the model specified by (9) in this article. They also provided a method to fit the model.

REFERENCES

- Andersen, P. K., Klein, J. P. and Zhang, M. (1999). Testing for centre effects in multi-centre survival studies: a monte carlo comparison of fixed and random effects tests. *Statistics in medicine* **18**, 1489–1500.
- Breslow, N. E. (1972). Discussion of the paper by d r cox cited below. *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

- Commenges, D. and Andersen, P. K. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis* **1**, 145–160.
- Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.
- Emoto, S. E. and Matthews, P. C. (1990). A weibull model for dependent censoring. *The Annals of Statistics* **18**, 1556–1577.
- Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley, New York, NY, USA.
- Gray, R. J. (1995). Tests for variation over groups in survival data. *Journal of the American Statistical Association* **90**, 198–203.
- Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika* **73**, 671–678.
- Huang, X. and Wolfe, R. A. (2002). A frailty model for informative censoring. *Biometrics* **58**, 510–520.
- Joe, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis* **46**, 262–282.
- Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.
- Klein, J. P. and Moeschberger, M. L. (1987). Independent or dependent competing risks-does it make a difference. *Communications in statistics-simulation and computation* **16**, 507–533.
- Lagakos, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics* **35**, 139–156.
- Lee, E. W., Wei, L. J. and Amato, D. A. (1992). Cox-type regression analysis for large number of small groups of correlated failure time observations. In Klein, J. P. and Goel, P. K., editors, *Survival Analysis: State of the Art*, pages 237–247, Netherlands. Kluwer Academic Publishers.
- Lee, S. and Wolfe, R. A. (1998). A simple test for independent censoring under the proportional

- hazard model. *Biometrics* **54**, 1176–1182.
- Lehmann, E. L. (1991). *Testing Statistical Hypotheses*. Wadsworth, Pacific Grove, California.
- Lin, D. Y., Robins, J. M. and Wei, L. J. (1996). Comparing two failure time distributions on the presence of dependent censoring. *Biometrika* **83**, 381–393.
- Lindeboom, M. and Van Den Berg, G. J. (1994). Heterogeneity in models for bivariate survival: The importance of the mixing distribution. *Journal of the Royal Statistical Society, Series B* **56**, 49–60.
- Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In Jewell, N. P., Dietz, K. and Farewell, V. T., editors, *AIDS Epidemiology: Methodological Issues*, pages 297–331, Boston. Birkhäuser.
- Tsiatis, A. A. (1975). A non-identifiability aspect of the problem of competing risks. *Proceedings of National Academic Science* **72**, 20–22.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
- Wolfe, R. A. (1994). The standardized mortality ratio revisited: improvements, innovations, and limitations. *American journal of Kidney Disease* **24**, 290–297.
- Ying, A. and Wei, L. J. (1994). The kaplan-meier estimate for dependent failure time observations. *Journal of Multivariate Analysis* **50**, 17–29.
- Young, E. W., Goodkin, D. A., Mapes, D. L., Port, F. K. and et al (2000). The dialysis outcomes and practice patterns study (dopps): An international hemodialysis study. *Kidney International* **57**, Suppl **74**, S-74–S-81.
- Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* **82**, 127–138.

APPENDIX A

Proof for Theorem 1

Proof. There are N subjects in the study. We replace the double subscript (i, j) in the text by a single subscript i , with i running from 1 to N . Suppose that the observed survival times X_i , $i = 1, \dots, N$, have been sorted in ascending order. We need to show $\sum_{i=1}^N \hat{R}_i^{(T)} \hat{R}_i^{(C)} = 0$. Define notation $Y_i(t) = 1(X_i \geq t)$, $\bar{Y}(t) = \sum_{i=1}^N Y_i(t)$, $a \wedge b = \min(a, b)$, and $a \vee b = \max(a, b)$. Then,

$$\begin{aligned}
& \sum_{i=1}^N \hat{R}_i^{(T)} \hat{R}_i^{(C)} \\
&= \sum_{i=1}^N (\Delta_i^{(T)} - \hat{\Lambda}^{(T)}(X_i)) (\Delta_i^{(C)} - \hat{\Lambda}^{(C)}(X_i)) \\
&= 0 - \sum_{i=1}^N \Delta_i^{(T)} \hat{\Lambda}^{(C)}(X_i) - \sum_{i=1}^N \Delta_i^{(C)} \hat{\Lambda}^{(T)}(X_i) + \sum_{i=1}^N \hat{\Lambda}^{(T)}(X_i) \hat{\Lambda}^{(C)}(X_i) \quad (\text{A.1})
\end{aligned}$$

Consider the last term.

$$\begin{aligned}
& \sum_{i=1}^N \hat{\Lambda}^{(T)}(X_i) \hat{\Lambda}^{(C)}(X_i) \\
&= \sum_{i=1}^N \left(\sum_{j \leq i} \frac{\Delta_j^{(T)}}{\bar{Y}(X_j)} \right) \left(\sum_{k \leq i} \frac{\Delta_k^{(C)}}{\bar{Y}(X_k)} \right) \\
&= \sum_{j=1}^N \sum_{k=1}^N \sum_{i=j \vee k} \frac{\Delta_j^{(T)} \Delta_k^{(C)}}{\bar{Y}(X_j) \bar{Y}(X_k)} \\
&= \sum_{j=1}^N \sum_{k=1}^N \{ \bar{Y}(X_j) \wedge \bar{Y}(X_k) \} \frac{\Delta_j^{(T)} \Delta_k^{(C)}}{\bar{Y}(X_j) \bar{Y}(X_k)} \\
&= \sum_{j=1}^N \sum_{k=j}^N \frac{\Delta_j^{(T)} \Delta_k^{(C)}}{\bar{Y}(X_j)} + \sum_{j=2}^N \sum_{k=1}^{j-1} \frac{\Delta_j^{(T)} \Delta_k^{(C)}}{\bar{Y}(X_k)} + \sum_{j=1}^N \sum_{k=j}^j \frac{\Delta_j^{(T)} \Delta_k^{(C)}}{\bar{Y}(X_k)} \quad (\text{last term is zero}) \\
&= \sum_{k=1}^N \Delta_k^{(C)} \sum_{j \leq k} \frac{\Delta_j^{(T)}}{\bar{Y}(X_j)} + \sum_{j=1}^N \Delta_j^{(T)} \sum_{k \leq j} \frac{\Delta_k^{(C)}}{\bar{Y}(X_k)} \\
&= \sum_{k=1}^N \Delta_k^{(C)} \hat{\Lambda}^{(T)}(X_k) + \sum_{j=1}^N \Delta_j^{(T)} \hat{\Lambda}^{(C)}(X_j)
\end{aligned}$$

The above cancels the first two terms in (A.1). Consequently, $\sum_{i=1}^N \hat{R}_i^{(T)} \hat{R}_i^{(C)} = 0$.

APPENDIX B

Proof for Theorem 2

Proof. Note that $R_i^{(T)} R_i^{(C)} = \frac{1}{n_i^2} \sum_{j=1}^{n_i} R_{ij}^{(T)} R_{ij}^{(C)} + \frac{1}{n_i^2} \sum_{j \neq k} R_{ij}^{(T)} R_{ik}^{(C)}$. We will show these two terms have expectation zero respectively.

1). Show $E(R_{ij}^{(T)} R_{ij}^{(C)}) = 0$.

Define two counting processes

$$\begin{aligned} N_{ij}^{(T)}(u) &= I(X_{ij} \leq u, \Delta_{ij}^{(T)} = 1), \\ N_{ij}^{(C)}(u) &= I(X_{ij} \leq u, \Delta_{ij}^{(C)} = 1). \end{aligned}$$

By the assumption that the distributions of T and C are continuous, the above two counting processes have no common jumps. Therefore, they form a bi-variate counting process. Further define

$$M_{ij}^{(T)}(u) = N_{ij}^{(T)}(u) - \Lambda_{ij}^{(T)}(u \wedge X_{ij}), \tag{B.2}$$

$$M_{ij}^{(C)}(u) = N_{ij}^{(C)}(u) - \Lambda_{ij}^{(C)}(u \wedge X_{ij}). \tag{B.3}$$

Then, under the assumption that T_{ij} and C_{ij} are independent, both $M_{ij}^{(T)}(\cdot)$ and $M_{ij}^{(C)}(\cdot)$ are martingales with respect to the filtration $\mathcal{F}_{ij} = \{\mathcal{F}_{ij,u}, 0 \leq u < \infty\}$, where $\mathcal{F}_{ij,u}$ is the σ -field generated by the information up to time u from the $(i, j)^{th}$ subject, i.e.,

$$\mathcal{F}_{ij,u} = \sigma\{N_{ij}^{(T)}(v), N_{ij}^{(C)}(v), 0 \leq v \leq u\}.$$

Then by the Theorem 2.5.2 of Fleming and Harrington (1991), the process

$$\{M_{ij}^{(T)}(u)M_{ij}^{(C)}(u), 0 \leq u < \infty\}$$

is also a martingale with respect to \mathcal{F}_{ij} and $E\{M_{ij}^{(T)}(u)M_{ij}^{(C)}(u)\} = 0$ for any $0 \leq u < \infty$.

Therefore, we have that

$$\begin{aligned}
E(R_{ij}^{(T)} R_{ij}^{(C)}) &= E\{M_{ij}^{(T)}(X_{ij})M_{ij}^{(C)}(X_{ij})\} \\
&= \lim_{u \rightarrow \infty} E\{M_{ij}^{(T)}(u \wedge X_{ij})M_{ij}^{(C)}(u \wedge X_{ij})\} \\
&= \lim_{u \rightarrow \infty} E\{M_{ij}^{(T)}(u)M_{ij}^{(C)}(u)\} \\
&= 0.
\end{aligned}$$

The last but one equality is gotten by noticing that $M_{ij}^{(T)}(u \wedge X_{ij}) = M_{ij}^{(T)}(u)$ and $M_{ij}^{(C)}(u \wedge X_{ij}) = M_{ij}^{(C)}(u)$.

2). Show $E(R_{ij}^{(T)} R_{ik}^{(C)}) = 0$, for any $j \neq k$.

In this part of the proof, we use the assumption that C_{ij} is independent of C_{ik} .

$$\begin{aligned}
&E(R_{ij}^{(T)} R_{ik}^{(C)}) \\
&= E \left[\left\{ I(T_{ij} \leq C_{ij}) - \Lambda_{ij}^{(T)}(T_{ij} \wedge C_{ij}) \right\} \left\{ I(C_{ik} \leq T_{ik}) - \Lambda_{ik}^{(C)}(T_{ik} \wedge C_{ik}) \right\} \right] \\
&= E \left(E \left[\left\{ I(T_{ij} \leq C_{ij}) - \Lambda_{ij}^{(T)}(T_{ij} \wedge C_{ij}) \right\} \left\{ I(C_{ik} \leq T_{ik}) - \Lambda_{ik}^{(C)}(T_{ik} \wedge C_{ik}) \right\} \right] | T_{ik} \right) \\
&= E \left[E \left\{ I(T_{ij} \leq C_{ij}) - \Lambda_{ij}^{(T)}(T_{ij} \wedge C_{ij}) | T_{ik} \right\} E \left\{ I(C_{ik} \leq T_{ik}) - \Lambda_{ik}^{(C)}(T_{ik} \wedge C_{ik}) | T_{ik} \right\} \right] \\
&= E \left[E \left\{ I(T_{ij} \leq C_{ij}) - \Lambda_{ij}^{(T)}(T_{ij} \wedge C_{ij}) | T_{ik} \right\} \cdot 0 \right] \\
&= 0
\end{aligned}$$

This concludes the proof.

Table 1
Size and power (in %) of 5% level tests using t_u in (7), Data Generated by Univariate Log-Normal Frailty Model in (10)

Censoring rate	Overall sample size	Number of clusters	Cluster size	Rates of Rejecting H_0 under different degree of dependence between T and C				
				$r = 0$	$r = -0.14$	$r = 0.14$	$r = -0.37$	$r = 0.37$
When there are not covariates								
30% censored by C	200	100	2	4.5	29.2	28.9	54.5	67.4
	200	40	5	4.6	54.5	48.8	83.6	90.6
	200	20	10	4.4	55.6	49.2	83.8	84.1
	400	200	2	5.0	50.0	51.5	81.5	93.5
	400	80	5	4.3	82.9	83.3	97.9	99.9
	400	40	10	3.9	88.4	85.9	99.3	99.7
15% censored by C	200	100	2	4.8	15.7	15.5	34.6	40.3
	200	40	5	5.3	35.0	33.2	65.6	77.2
	200	20	10	4.6	41.0	35.8	69.3	79.4
	400	200	2	5.0	29.0	32.3	62.1	74.9
	400	80	5	4.5	61.9	62.5	94.1	97.5
	400	40	10	5.0	70.3	74.4	97.6	99.2
When there are covariates								
30% censored by C	200	100	2	4.1	28.5	24.4	60.6	60.1
	200	40	5	4.2	55.1	45.7	87.9	85.7
	200	20	10	4.6	55.0	45.1	88.6	80.7
	400	200	2	5.0	54.4	45.3	90.7	91.3
	400	80	5	4.5	86.9	75.5	99.3	99.8
	400	40	10	4.3	86.9	81.4	99.5	98.6
15% censored by C	200	100	2	4.2	14.8	15.1	38.3	44.9
	200	40	5	4.7	36.1	31.9	73.2	75.7
	200	20	10	4.6	40.6	35.5	75.0	77.2
	400	200	2	4.9	30.8	29.6	69.5	74.2
	400	80	5	5.2	64.0	57.5	96.1	97.0
	400	40	10	4.5	71.4	71.2	97.6	98.6

Note: r is the Pearson correlation coefficient between $\log(T)$ and $\log(C)$. 10,000 replicates for size. 1,000 replicates for power.

There are two censoring settings. One has approximately 50% of failure times T observed, 30% censored by C (drop-out time, may be correlated with T), the rest 20% censored by Q (administrative censoring time, independent of T and C). This setting is labeled as “30% censored by C ” in the first column. The other setting has 35% failure times T observed, 15% censored by C , the rest 50% censored by A . This setting is labeled as “15% censored by C ”.

Table 2
Size and power (in %) of 5% level tests using t_u in (7)

Overall sample size	Number of clusters	Cluster size	Rates of Rejecting H_0 under different degree of dependence between T and C				
			$r = 0$	$r = -0.13$	$r = 0.13$	$r = -0.26$	$r = 0.26$
Data Generated by Bivariate Log-Normal Frailty Model in (11)							
N	m	n	$r = 0$	$r = -0.13$	$r = 0.13$	$r = -0.26$	$r = 0.26$
200	100	2	4.3	8.3	13.0	26.2	34.2
200	40	5	3.3	13.2	14.2	44.5	51.4
200	20	10	3.0	9.1	11.2	39.5	43.6
400	200	2	3.9	13.0	15.7	50.3	56.8
400	80	5	3.3	22.8	26.2	78.5	82.5
400	40	10	2.8	15.7	20.2	73.7	80.9
Data Generated by Positive Stable Frailty Model in (12)							
N	m	n	$r = 0$	$r = 0.34$	$r = 0.62$	$r = 0.83$	
200	100	2	4.4	12.6	50.1	89.7	
200	40	5	3.6	18.9	62.6	92.5	
200	20	10	3.1	15.1	43.3	70.3	
400	200	2	4.6	29.2	88.3	99.8	
400	80	5	3.5	51.3	95.6	100.0	
400	40	10	2.7	38.8	85.7	98.7	

Note: r is the Pearson correlation coefficient between $\log(T)$ and $\log(C)$. 10,000 replicates for size. 1,000 replicates for power.

For top half table, 60% failure times T observed, 20% censored by C (drop-out time, may be correlated with T), the rest 10% censored by Q (administrative censoring time, independent of both T and C).

For bottom half table, 35% failure times T observed, 40% censored by C , the rest 25% censored by Q .