Published in final edited form as:

Comput Stat Data Anal. 2014 June 1; 74: 39–51. doi:10.1016/j.csda.2013.12.006.

Sample size determination for paired right-censored data based on the difference of Kaplan-Meier estimates

Pei-Fang Su^a, Chung-I Li^b, and Yu Shyr^{c,*}

^aDepartment of Statistics, National Cheng Kung University, Tainan 70101, Taiwan

^bDepartment of Applied Mathematics, National Chiayi University, Chiayi 60004, Taiwan

^cCenter for Quantitative Sciences, Vanderbilt University, Nashville, TN 37232, USA

Abstract

Sample size determination is essential to planning clinical trials. Jung (2008) established a sample size calculation formula for paired right-censored data based on the logrank test, which has been well-studied for comparing independent survival outcomes. An alternative to rank-based methods for independent right-censored data, advocated by Pepe and Fleming (1989), tests for differences between integrated weighted Kaplan-Meier estimates and is more sensitive to the magnitude of difference in survival times between groups. In this paper, we employ the concept of the Pepe-Fleming method to determine an adequate sample size by calculating differences between Kaplan-Meier estimators considering pair-wise correlation. We specify a positive stable frailty model for the joint distribution of paired survival times. We evaluate the performance of the proposed method by simulation studies and investigate the impacts of the accrual times, follow-up times, loss to follow-up rate, and sensitivity of power under misspecification of the model. The results show that ignoring the pair-wise correlation results in overestimating the required sample size. Furthermore, the proposed method is applied to two real-world studies, and the R code for sample size calculation is made available to users.

Keywords

Logrank test; Kaplan-Meier statistic; hazard function; paired observation; positive stable frailty model

1. Introduction

Determination of an adequate sample size is critical to the design of research ventures. When correlation exists among observations, this association should be considered when estimating sample size, to provide a more reasonable and feasible estimation. For example, nearly 12 million Americans are affected by diabetes, the leading cause of blindness in working-age Americans, accounting for approximately 12% of the new cases of blindness annually (Patz and Smith, 1991). To determine the benefits of early photocoagulation in

© 2014 Elsevier B.V. All rights reserved.

Corresponding author: yu.shyr@vanderbilt.edu (Yu Shyr^{c,}).

URL: <https://medschool.vanderbilt.edu/cqs> (Yu Shyr^{c,*})

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

patients with type I versus type II diabetes, the Diabetic Retinopathy Study (DRS) examined the effectiveness of this technology in delaying onset of blindness in patients with diabetic retinopathy. The DRS accrued patients with diabetes and nonproliferative or early proliferative retinopathy in both eyes. Each patient had one eye randomized to argon laser treatment and the other eye to xenon arc photocoagulation. Patients then were followed for five to nine years to detect vision loss, with "survival time" defined as duration from initiation of treatment to diagnosis of blindness (i.e., visual acuity below 5/200 for at least two consecutive visit). Because time to vision loss is positively correlated within individuals, paired right-censored data arise; studies utilizing such "naturally" paired systems (e.g., eyes, ears, twins) to compare two treatments benefit from a reduction in between-subject variability. In designing such cases, however, required sample size may be overestimated if outcomes are treated as if independent. Consequently, dealing with correlation is important for accurate sample size estimation to achieve sufficient statistical power.

Throughout this paper, sample size refers to number of pairs. The problem of determining an adequate sample size can be addressed through the traditional hypothesis testing framework. To compare the lifetime distributions of two independent groups, the logrank test (Mantel, 1966), a well-known two-sample test, is commonly used. This rank-based method and various related methods such as the Gehan-Wilcoxon test (Gehan, 1965) and Prentice-Wilcoxon test (Prentice, 1978) have been unified by Gill (1980). For comparing the equality of two survival functions for paired right-censored data, Jung (1999) extended the logrank test to adjust for possible dependence between pairs. Similarly, Murray (2000) discusses an adjusted group sequential method for paired rank tests. For sample size calculation based on paired survival outcomes, Gangnon and Kosorok (2004) propose a sample size formula that specifies a correlation coefficient between the two parts of the rank statistic, but does not address correlation between pairs. By contrast, Jung (2008) has proposed a formula that calculates sample size through direct specification of the correlation coefficient between pairs, as well as specification of joint survival distribution, accrual period, and additional follow-up period.

The logrank test is based on integrated weighted differences between two estimated hazard functions. An alternative to rank-based methods, advocated by Pepe and Fleming (1989) for independent right-censored data, looks at differences between integrated weighted survival curves. For paired right-censored data, Murray (2001) extended weighted Kaplan-Meier statistics to compare two survival functions and found that rank-based methods might not be sensitive to the magnitude of difference in survival times between groups. For sample size calculation, this non-rank-based method has not been investigated. Thus, in this paper, we propose an alternative to Jung (2008)'s sample size calculation based on the logrank test: sample size estimation for paired right-censored data based on the difference between Kaplan-Meier statistics. Our method can easily be applied to independent survival outcomes, as well.

The paper is organized as follows: Section 2 introduces notation and paired integrated weighted differences in Kaplan-Meier estimates (Kaplan and Meier, 1958). Section 3 describes calculation of required sample size, step by step. Because Jung (2008) has demonstrated that the correlation coefficient depends on the censoring distribution as well as the joint survival distributions, we also separate the contribution of the censoring distribution from that of the dependency between pairs. For computational ease, the positive stable frailty model (Hougaard, 1986) is used for modeling the joint distribution of paired survival times; and the exponential marginal hazard rates, accrual period (or accrual rate), and additional follow-up period are specified as parameters in the formula. In Section 4, We evaluate the performance of the proposed method by simulation studies and investigate the

impacts of the accrual times, follow-up times, loss to follow-up rate, and sensitivity of power under misspecification of the joint survival distribution. In Section 5, two pilot studies, including the diabetic retinopathy study and skin graft study, are used to illustrate sample size calculation. Finally, we offer some concluding remarks.

2. Paired Pepe-Fleming statistics

In this section, we describe integrated weighted differences in Kaplan-Meier estimates for paired right-censored data. Let (T_{1i}, T_{2i}) be the bivariate survival times, and (C_{1i}, C_{2i}) be the i.i.d. censoring times independent of (T_{1i}, T_{2i}) for the i th ($i = 1, \dots, n$) pair of subjects. The common marginal survival functions for T_{ki} are denoted by S_k ; the marginal cumulative hazard functions are denoted by Λ_k ; and the hazard functions are denoted as λ_k for group k ($k = 1, 2$). When a dataset includes some right-censored observations, one can only observe the times $X_{ki} = \min(T_{ki}, C_{ki})$ and $\Delta_{ki} = I(T_{ki} < C_{ki})$, where $I(A)$ is an indicator function of event A , taking value 1 if the event A occurs and value 0 otherwise.

Let $N_k(t) = \sum_{i=1}^n N_{ki}(t)$ count the number of individuals from group k who fail at time t , where $N_{ki}(t) = \Delta_{ki}I(X_{ki} \leq t)$ and $Y_k(t) = \sum_{i=1}^n Y_{ki}(t)$ count the number of individuals still at risk at time t , where $Y_{ki}(t) = I(X_{ki} \geq t)$. For testing the equality of two survival functions (null hypothesis, $H_0 : S_1(t) = S_2(t), \forall t$), the logrank test (LR) is based on the integrated weighted differences between two estimated hazard functions. The numerator of the logrank test is

$$LR = \int_0^\infty H(t) \{d\hat{\Lambda}_1(t) - d\hat{\Lambda}_2(t)\},$$

where $\hat{\Lambda}_k(t) = \int_0^t Y_k^{-1}(s) dN_k(s)$ is the Nelson-Aalen estimator (Nelson, 1972), and $H(t) = n^{-1} Y_1(t) Y_2(t) / (Y_1(t) + Y_2(t))$. For independent survival outcomes, readers may refer to Gill (1980) or Fleming and Harrington (1991) for a variance estimate of the test statistics. For paired right-censored outcomes, the standard error of LR should be modified to accommodate the paired correlation; Jung (1999) has completed an excellent work to derive the asymptotic variance of the rank test statistics.

Because the logrank test is rank-based, this method might not be sensitive to the magnitude of difference in survival times against a specific alternative. Therefore, Pepe and Fleming (1989) proposed the weighted integrated survival difference (KM) for comparing two independent samples of right-censored data. The statistic is

$$KM = \int_0^\infty w(t) \{\hat{S}_1(t) - \hat{S}_2(t)\} dt, \quad (1)$$

where $\hat{S}_k(t)$ is the Kaplan-Meier estimator (Kaplan and Meier, 1958) for the k th group and $w(t)$ is a selected weight function. This form suggests a natural measure of the difference in survival function between two groups, which can be interpreted as the difference in mean survival times when $w(t) = 1$. More attractively, this method is sensitive to the magnitude of difference in survival times. The power of the Pepe-Fleming test is comparable to that of rank-based tests. When two survival curves are expected to cross the study period, the simulations showed that KM has higher power than rank-based methods to detect survival differences (Pepe and Fleming, 1989).

For comparison in the paired right-censored survival setting, Murray (2001) extends KM by taking into account the positive paired correlation between estimated survival curves. That

is, let the joint and conditional hazards $\lambda_{12}(t_1, t_2) = \lim_{\delta_1, \delta_2 \rightarrow 0} P(t_1 - X_{1i} < t_1 + \delta_1, t_2 - X_{2i} < t_2 + \delta_2, \Delta_{1i} = 1, \Delta_{2i} = 1 | X_{1i} = t_1, X_{2i} = t_2) / \delta_1 \delta_2$, $\lambda_{1|2}(t_1 | t_2) = \lim_{\delta_1 \rightarrow 0} P(t_1 - X_{1i} < t_1 + \delta_1, \Delta_{1i} = 1 | X_{1i} = t_1, X_{2i} = t_2) / \delta_1$, and $\lambda_{2|1}(t_2 | t_1) = \lim_{\delta_2 \rightarrow 0} P(t_2 - X_{2i} < t_2 + \delta_2, \Delta_{2i} = 1 | X_{1i} = t_1, X_{2i} = t_2) / \delta_2$. Under the alternative hypothesis, $\sqrt{n}KM$ is approximately normal with mean

$$\mu = \int_0^\infty w(t) \{S_1(t) - S_2(t)\} dt,$$

and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$, where

$$\sigma_k^2 = \int_0^\infty \frac{A_k^2(u) \lambda_k(u)}{P(X_{ki} \geq u)} du$$

($k = 1, 2$) are of group k with $A_k(t) = \int_t^\infty w(u) S_k(u) du$, and

$$\sigma_{12} = \int_0^\infty \int_0^\infty A_1(u) A_2(v) G_{12}(u, v) du dv$$

gives the covariance, characterizing the dependence between two Kaplan-Meier estimators, where

$$G_{12}(u, v) = \frac{P(X_{1i} \geq u, X_{2i} \geq v)}{P(X_{1i} \geq u)P(X_{2i} \geq v)} \lambda_{12} - \lambda_{1|2} \lambda_2 - \lambda_{2|1} \lambda_1 + \lambda_1 \lambda_2.$$

In practice, σ^2 can be estimated by $\hat{\sigma}^2$ as

$$\sum_{k=1}^2 \left[\int_0^\infty \frac{n}{Y_k^2(u)} \hat{A}_k^2(u) dN_k(u) \right] - 2 \int_0^\infty \int_0^\infty \hat{A}_1(u) \hat{A}_2(v) \hat{G}_{12}(u, v) du dv$$

where $\hat{A}_k(t) = \int_t^\infty w(u) \hat{S}_k(u) du$ and

$$\hat{G}_{12}(u, v) du dv = \frac{n Y_{12}(u, v)}{Y_1(u) Y_2(v)} \left[\frac{dN_{12}(u, v)}{Y_{12}(u, v)} - \frac{dN_{1|2}(u|v) dN_2(v)}{Y_{12}(u, v) Y_2(v)} - \frac{dN_{2|1}(v|u) dN_1(u)}{Y_{12}(u, v) Y_1(u)} + \frac{dN_1(u) dN_2(v)}{Y_1(u) Y_2(v)} \right]$$

with $Y_{12}(u, v) = \sum_{i=1}^n I(X_{1i} \geq u, X_{2i} \geq v)$, counting the number of pairs still at risk for failure at time u and v in group one and two;

$$dN_{1|2}(u|v) = \sum_{i=1}^n I(X_{1i} = u, X_{2i} \geq v, \Delta_{1i} = 1), \quad dN_{2|1}(v|u) = \sum_{i=1}^n I(X_{1i} \geq u, X_{2i} = v, \Delta_{2i} = 1)$$

, and $dN_{12}(u, v) = \sum_{i=1}^n I(X_{1i} = u, X_{2i} = v, \Delta_{1i} = 1, \Delta_{2i} = 1)$, respectively. Note that we assume common censoring times for a pair. The choice of the weight function $w(t) = P(\hat{C}_i \leq t)$, the survival function estimator for censoring random variables, was recommended to ensure the stability of the statistic (Pepe and Fleming, 1989). When the absolute value of

$\sqrt{n}KM/\hat{\sigma}$ is larger than $z_{1-\alpha/2}$, the $100(1 - \alpha/2)$ percentile of the standard normal distribution, we reject the null hypothesis and conclude the two groups have different survival times.

3. Sample size calculation

In this section, we focus on how to calculate sample size based on the difference in Kaplan-Meier estimators for paired right-censored data. For designing a clinical trial, sample size calculation would proceed as follows: (1) specify the joint survival function $S(t_1, t_2) = P(T_{1i} > t_1, T_{2i} > t_2)$ and the joint survival function for censoring variables $G(t_1, t_2) = P(C_{1i} > t_1, C_{2i} > t_2)$; (2) calculate mean μ and variance σ^2 based on KM ; and (3) specify type I error rate α , power $1 - \beta$, and calculate the required sample size n .

3.1. Specify the joint survival function and censoring function

Because the positive stable frailty model (Hougaard, 1986) facilitates calculation of joint and marginal survival functions, we consider this model as used by Jung (2008) throughout this paper. The positive stable frailty model is

$$S(t_1, t_2) = \exp\{-[(-\log S_1(t_1))^{1/\theta} + (-\log S_2(t_2))^{1/\theta}]^\theta\},$$

where $0 < \theta \leq 1$ represents the level of dependency of survival times within each pair, with $\theta = 1$ as independence. Because exponential distributions have been widely used to model independent survival distributions in the design stage of a clinical trial, we assume the marginal exponential survival distributions $S_k(t)$ equals $\exp(-\lambda_k t)$. Then,

$$S(t_1, t_2) = \exp\{-[(\lambda_1 t_1)^{1/\theta} + (\lambda_2 t_2)^{1/\theta}]^\theta\}.$$

In addition, we assume pairs are uniformly accrued and the common censoring times $C_1, C_2, \dots, C_n$ are drawn from an i.i.d. uniform distribution, $U(b, a + b)$, where a is the accrual period and b is the follow-up period. If a proportion of subjects are expected to be lost to follow-up, we consider adopting the following term. Assume censoring time and follow-up time are independent. Censoring time has a uniform distribution, $U(b, a+b)$, and loss to follow-up time has an exponential distribution with hazard rate ν before the scheduled end of study. Then, the survival function of censoring time becomes

$$G(t) = \begin{cases} \exp(-\nu t) & \text{if } t < b \\ \{1 - (t-b)/a\} \exp(-\nu t) & \text{if } b \leq t < a+b \\ 0 & \text{otherwise} \end{cases}.$$

$G(t)$ incorporates both parameters: loss to follow-up and censoring time. We will investigate the impacts of the accrual period, follow-up period, loss to follow-up rate on sample size calculation later.

3.2. Calculate mean, variance and covariance

In this step, we calculate mean μ and variance σ^2 based on Kaplan-Meier estimates. The positive stable frailty model (Hougaard, 1986) is used for modeling the joint distribution of the paired survival times. Then, μ can be derived as

$$\mu = \int_0^{a+b} (S_1(t) - S_2(t)) G(t) dt$$

Appendix A further details this derivation. Moreover, $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$, where variance can be derived as

$$\sigma_k^2 = \lambda_k \int_0^{a+b} \frac{A_k^2(t)}{G(t)S_k(t)} dt, k=1, 2.$$

Covariance is given as

$$\sigma_{12} = \int_0^{a+b} \int_0^{a+b} A_1(t_1) A_2(t_2) \frac{G(t_1 \vee t_2) S(t_1, t_2)}{G(t_1) G(t_2) S_1(t_1) S_2(t_2)} dA(t_1, t_2),$$

with $t_1 \vee t_2$ defined as $\max(t_1, t_2)$, and $A_k(t)$ and $dA(t_1, t_2)$ are defined in Appendix A.

3.3. Sample size calculation

We now utilize the asymptotic results of *KM* to derive sample size formulas. The theoretical power for a given n is

$$1 - \beta = \Phi \left(z_{1-\alpha/2} - \frac{\sqrt{n}\mu}{\sigma} \right).$$

For a two-sided α and a given power $1 - \beta$, the required sample size is

$$n = \frac{\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{\mu^2},$$

where μ and σ are described in Section 3.2. In addition, all clinical trials should be planned and designed such that study objectives may be met as quickly as feasible. To minimize exposure of subjects to ineffective treatment, accrual rate is a critical factor in trial progress. Thus, the potential accrual rate should be carefully considered during the trial planning phase, to ensure timely completion of scientific objectives. Suppose accrual rate r is also a given parameter. Under uniform accrual, we have $n = a \times r$. Using numerical methods, sample size can be solved through

$$a \times r = \frac{\sigma^2(a) (z_{1-\alpha/2} + z_{1-\beta})^2}{\mu^2(a)},$$

where $\sigma^2 = \sigma^2(a)$ and $\mu = \mu(a)$ are functions of a .

4. Numerical studies

To evaluate the performance of the proposed formula for paired right-censored data, we conducted some numerical studies. In the first part, we calculate required sample size (denote this as n_K) under various design settings without considering loss to follow up rate. We also show sample sizes n_L as calculated using Jung (2008)'s formula based on the log-rank test. The impact of accrual period and follow-up period is investigated as well. In the second part, we consider loss to follow-up rate v and create the power curves under different scenarios. In the third part, we investigate the sensitivity of the power under misspecification of the joint survival distribution.

4.1. Sample size calculation without considering loss to follow-up rate

In this subsection, we fix the hazard rate $\lambda_1 = 0.5$ for the first group. We vary the hazard rate as $\lambda_2 = 0.35, 0.3$, or 0.25 for the second group. Power is set as $1 - \beta = 0.8$ or 0.9 , and the type I error rate as $\alpha = 0.05$. Accrual period is $a = 3$, and follow-up period is set as $b = 0, 1$, or 2 . We specify θ in the positive stable frailty model. A large θ yields smaller intra-cluster correlations, with $\theta = 1$ yielding independent cases. We set the values of $\theta = 0.3, 0.6, 0.9$, representing high correlation, moderate correlation, and low correlation within each pair; the corresponding $\rho = 0.803, 0.449, 0.103$, respectively.

Under the positive stable frailty model, paired survival times were generated as follows. First, T_1 is generated from $S_1(t) = \exp(-\lambda_1 t)$. Given t_1 , T_2 then is generated by solving

$$\begin{aligned} S_{2|1}(t_2|t_1) &= -\frac{\partial S(t_1, t_2)/\partial t_1}{f_1(t_1)} \\ &= -\frac{\lambda_1(\lambda_1 t_1)^{1/\theta-1} \{(\lambda_1 t_1)^{1/\theta} + (\lambda_2 t_2)^{1/\theta}\}^{\theta-1} S(t_1, t_2)}{\lambda_1 \exp(-\lambda_1 t)} = u, \end{aligned}$$

where u is generated from $U(0, 1)$ distribution. For the group with the hazard rate $0.35, 0.3, 0.25$, censoring probabilities are approximately 62%, 66%, 70%, respectively, when $b = 0$; approximately 44%, 49%, 55%, when $b = 1$; and 31%, 36%, 43%, when $b = 2$. Table B.1 shows the required sample sizes n_L and n_K under different scenarios, along with the empirical power for *KM* (or *LR*) tests based on 2000 simulation samples of size n_K (or n_L) generated from the same setting.

As seen in Table B.1, as λ_2 decreases, required sample size decreases, as expected; with the larger hazard ratio between the two groups, a smaller sample size provides enough power to detect the difference. Based on the simulation, we also observed that the required sample size decreased as follow-up time b increased. On the other hand, required sample size increases with an increase in θ . The results show that ignoring the pair-wise correlation results in overestimating the required sample size. Overall, the performance of n_K shows the same pattern as n_L . Through simulation, we found that the empirical powers close to $1 - \beta$. However, it is slightly smaller than the nominal $1 - \beta$ level when the estimated sample size is relatively small, approximately smaller than 30. That's because the *KM* test is based on the large sample theory, and the approximation does not work well when sample sizes are

small. To investigate the impact of the accrual and follow-up times under different scenarios, we also plot the required sample size for the Kaplan-Meier tests to achieve a power of 0.8 under the setting $\lambda_1 = 0.5$ and $\lambda_2 = 0.35$ for $\theta = 0.3, 0.6, 0.9$ and 1.0 in Figure B.1. The accrual times and follow-up times vary from 1 to 3.5 years. From Figure B.1, we find that the sample sizes for the four scenarios all monotonously decrease as the accrual times or follow-up times increase. On the other hand, those plots also shows that the accrual time has a larger impact on sample size than follow-up time, because the required sample size decreases fast for the accrual times, especially for the independent case ($\theta = 1.0$). In Figure B.2, we also investigate the relationship between the required sample size, hazard ratio λ_2/λ_1 and follow-up time b for different pairwise correlation θ . We fix $\lambda_1 = 0.5$, $a = 3$, and vary the hazard ratio and follow-up period b from 1.5 to 3, and 0 to 5, respectively. We observe that the required sample size decreased for increasing hazard ratio and higher paired correlation.

4.2. The impact of loss to follow-up rate v

In the second numerical study, we further assume that losses to follow-up are exponentially distributed and set the loss hazard rate at 0.1 and 0.2. The other settings are similar to those described in the first numerical study. The results are shown in Table B.2. As seen in Table B.2, we observe the same pattern: for both n_K and n_L , the required sample size decreases as λ_2 decreases. Longer follow-up time b decreases the sample size as well. Similar to the case that does not consider loss to follow-up rate v , the required sample size increases with an increase in θ . With loss to follow-up rate increasing ($v = 0.1$ to 0.2), we need a larger sample size to achieve the required power, as compared to Table B.1. Therefore, for higher pairwise correlation datasets with lower loss to follow-up rates, the required sample size is expected smaller. In Figure B.3, we fixed the hazard difference as 0.2 (e.g. $\lambda_1 = 0.5$ v.s. $\lambda_2 = 0.3$ and $\lambda_1 = 0.3$ v.s. $\lambda_2 = 0.1$) under $v = 0$ and $v = 0.2$, respectively, to show the theoretical power curves under different sample sizes and correlations. The simulation results show that the loss to follow-up rate has a slight impact on power for a fixed sample size, while pairwise correlation has a significant impact.

4.3. Sensitivity of the misspecified model

While the proposed method starts out purely nonparametrically, the major assumption (positive stable frailty model) renders the approach parametric. Therefore, in this subsection, simulation studies are presented to assess the sensitivity of power to misspecification of the joint survival distribution. For each simulated sample under the calculated sample size, we evaluate the performance of the paired log-rank test of Jung (1999) (*LR*) and Pepe and Fleming (1989)'s statistic (*KM*). For the simulation, we assume no additional follow-up time ($b = 0$) and no loss to follow-up ($v = 0$); in addition, joint survival times are generated through the positive stable frailty model (denoted as PSF) as well as Moran's algorithm (Moran, 1967) (denoted as Moran). Moran's algorithm was generated as follows: for joint survival times, let (V_1, V_3) and (V_2, V_4) be mutually independent, but each pair has a bivariate normal distribution with a marginally zero mean, unit variance, and correlation coefficient $\sqrt{\rho}$ ($\rho > 0$). We constructed the joint distributions $T_1 = 0.5(V_1^2 + V_2^2)/\lambda_1$ and $T_2 = 0.5(V_3^2 + V_4^2)/\lambda_2$. Accordingly, T_1 and T_2 have a marginal exponential distribution with failure rates λ_1 and λ_2 , respectively, and correlation coefficient ρ . In Table B.3, we show the empirical power under each set of parameters, with type I error rate given in parentheses; α was set as 0.05 based on 2000 simulation runs.

As seen in Table B.3, reasonable type I error rates are preserved; because both *LR* and *KM* tests are non-parametric, specification of the joint survival distribution is independent of type I error. With regard to power, as expected, both *LR* and *KM* increase in power under all

settings with an increase in λ_1/λ_2 . For independent cases ($\theta = 1$), both tests reaches the expected power under both the positive stable frailty model and Moran's model. For $\theta < 1$, however, when the joint survival distribution is misspecified (i.e., under Moran's model), *KM* appears to be underpowered, especially under the case of higher dependency. Actually, Jung (2008) observed the same pattern for *LR* and compared the conditional correlation coefficient between Moran and PSF model. They found that if the true survival distribution follows Moran's model with a short follow-up period, the estimated conditional correlation between pairs will be much lower than the true correlation. Therefore, the sample size calculated under the positive stable frailty model will be underpowered. Although the choice of a joint survival distribution has some impact on power, if a study has a reasonable longer follow-up period, misspecification of the frailty distribution can be expected to be of minor importance. Note that nothing is known about θ in practice. The choice of association measures for PSF model may depend on the statistical method. In the planning stage of a clinical trial, therefore, a sensitivity analysis using different values of θ should also be performed to see how a misspecification of θ in the sample size determination will affect the power.

5. Examples

5.1. The diabetic retinopathy study

Suppose we are designing the diabetic retinopathy study, as described in the introduction section. Because time to vision loss is positively correlated within individuals, paired right-censored data arise. We calculate sample size n_K . In this case, the maximum likelihood estimates are $\hat{\lambda}_1 = 0.021$ and $\hat{\lambda}_2 = 0.012$, which are assumed to be the true parameters. We set $\alpha = 0.05$, $1 - \beta = 0.9$, annual accrual rate $r = 700$ patients, and additional follow-up time $b = 2$ years. We calculate sample size based on various correlation ρ and provide the results in Figure B.4(a).

In this plot, the solid line provides required sample size n_L , and the dashed line, n_K . With high correlation between pairs ($\rho = 0.8$), $n_L = 594$, and $n_K = 474$. With no correlation ($\theta = 1$), $n_L = 1450$, and $n_K = 1692$. Finally, as seen in the plot, when θ is less than 0.6 ($\rho > 0.45$), n_K is smaller than n_L . Actually, the original study lasted from April 1980 to July 1985 and enrolled 3711 patients. Given the parameters as specified above, the study as conducted would seem substantially over-powered.

5.2. Skin grafts data

As a second example, we consider skin graft data from Holt and Prentice (1974). This clinical trial studied the survival of closely matched and poorly matched skin grafts on each of 11 burned patients. Although the close-match and poor-match come from different donors, the two grafts within each patient share the same physical conditions. Thus, paired right-censored data arise. Given the Holt and Prentice (1974) results as pilot data, suppose we wish to design a paired randomization study of the influence of HL-A antigens on the survival of allogeneic grafts in burned patients. The maximum likelihood estimates based on the positive stable frailty model are $\hat{\lambda}_1 = 0.043$, $\hat{\lambda}_2 = 0.025$ per day for the two groups, and $\hat{\theta} = 0.33$. We set $\alpha = 0.05$, $1 - \beta = 0.9$, accrual rate $r = 10$ per month, and additional follow-up $b = 2$ months. Then, the required total sample size is $n_L = 115$, or $n_K = 94$. If assuming loss to follow-up rate $v = 0.1$, the required total sample size increases to $n_L = 152$, or $n_K = 143$. In Figure B.4(b), we calculate required sample size based on various correlations ρ under $v = 0$.

6. Concluding remarks

Many published studies have evaluated the logrank test against the Kaplan-Meier statistic for comparing survival in two independent samples of right-censored data (Pepe and

Fleming, 1989; Chi and Tsai, 2001) or correlated right-censored data (Andrei and Murray, 2005; Chi and Su, 2010). In this paper, we focus on sample size determination based on Kaplan-Meier estimators for paired right-censored data. As opposed to the *LR*, the proposed method might be a useful alternative in practice, especially for highly paired correlation data. As discussed in the literature (Pepe and Fleming, 1989; Chi and Tsai, 2001; Murray, 2000, 2001), because *LR* is not sensitive to the magnitude of the difference in survival times against a specific alternative, we suggested applying the *KM* test and hence the corresponding sample size calculation. On the contrary, *KM* performs worse under late hazard difference alternatives in some studies. Therefore, *LR* might be a better choice. Also, based on our simulation, *KM* performs better than *LR* for higher paired correlation, with even the marginal survival curve proportional. The question of whether *KM* is actually more powerful than *LR* for paired right-censored data is worth further investigation in the future. It should be noted that once one method is selected (*KM* or *LR*) for sample size determination, the same method should be used for analysis.

Because exponential distributions have been widely used to model independent survival distributions in the design stage of a clinical trial, we consider using the positive stable frailty model with marginally exponential distributions to model joint survival distributions. In our simulation, we saw that under cases of higher positive association, small sample size is enough to attain the specified power. Similar results were shown by Jung (2008) based on the logrank test. Even misspecifying the model for joint survival distribution does not influence the type I error rates using the *KM* test based on the calculated sample size. If we are designing a trial for a pilot study, we may choose a model that to best fit the data and the measurement of the paired correlation seems more important. In Appendix B, the main body of R code for obtaining sample size based on our proposed formula is available to users.

Useful extensions to this work would include investigation of different marginal survival distributions for calculating sample size. For example, one could consider a Weibull distribution or log normal distribution to construct joint survival distributions. Use of models other than the positive stable frailty model for survival data, such as Moran's model (Moran, 1967) or Clayton's model (Clayton, 1998), also are suggested to extend this work. Model checking is also useful to explore in extensions to this work. For example, Shih (1998) proposes a goodness-of-fit test for a family of general joint survival models, which may be of use in testing the survival data model utilized in our approach. Although the proposed method starts out purely non-parametrically, the assumptions of the positive stable frailty model and exponential marginal survival, rendering the approach purely parametric, might be violated. Estimating the joint distribution non-parametrically (Lin et al., 1999) might be another way to think how to derive a fully non-parametric sample size calculation formula, although it is not easy. It would be a challenge and an interesting topic for future research. In addition, rather than comparing entire survival curves, researchers may wish to focus comparisons at clinically meaningful fixed time points (Su et al. (2011)). Therefore, sample size determination based on a fixed point in time would also be worthy of further investigation.

Acknowledgments

The authors wish to thank Lynne Berry and Margot Björing for editorial work on this manuscript. This work was supported by National Cancer Institute grants P30 CA068485, P50 CA090949, P50 CA095103, and P50 CA098131.

References

- Andrei AC, Murray S. Simultaneous group sequential analysis of rank-based and weighted Kaplan-Meier tests for paired censored survival data. *Biometrics*. 2005; 61:715–720. [PubMed: 16135022]

- Calyton DG. A model for association in bivariate life tables and its application to epidemiologic studies of familial tendency in chronic disease incidence. *Biometrika*. 1998; 65:141–151.
- Chi YC, Su PF. The simultaneous use of weighted logrank and weighted Kaplan-Meier statistics with clustered right-censored data. *Statistics in Medicine*. 2010; 29:130–141. [PubMed: 19827012]
- Chi YC, Tsai MH. Some versatile tests based on the simultaneous use of weighted logrank and weighted Kaplan-Meier statistics. *Communications in Statistics, Part B-Simulation and Computation*. 2001; 30:743–759.
- Fleming, TR.; Harrington, DP. *Counting Processes and Survival Analysis*. Wiley; New York: 1991.
- Gangnon RE, Kosorok MR. Sample size formula for clustered survival data using weighted log-rank statistics. *Biometrika*. 2004; 91:263–275.
- Gehan EA. A generalized Wilcoxon test for comparing arbitrarily single censored samples. *Biometrika*. 1965; 52:203–223. [PubMed: 14341275]
- Gill, RD. *Tract. Vol. 124. The Mathematical Centre*; Amsterdam: 1980. *Censoring and Stochastic Integrals*.
- Harrington DP, Fleming TR. A Class of Rank Test Procedures for Censored Survival Data. *Biometrika*. 1982; 69:553–566.
- Holt JD, Prentice RL. Survival analyses in twin studies and matched pair experiments. *Biometrika*. 1974; 61:17–30.
- Hougaard P. A class of multivariate failure time distributions. *Biometrika*. 1986; 73:671–678.
- Jung SH. Rank tests for matched survival data. *Lifetime Data Analysis*. 1999; 5:67–79. [PubMed: 10214003]
- Jung SH. Sample size calculation for the weighted rank statistics with paired survival data. *Statistics in Medicine*. 2008; 27:3350–3365. [PubMed: 18205148]
- Kaplan EL, Meier P. Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association*. 1958; 53:457–481.
- Lin DY, Sun W, Ying Z. Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika*. 1999; 86:59–70.
- Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*. 1966; 50:163–170.
- Moran PAP. Testing for correlation between non-negative variates. *Biometrika*. 1967; 54:385–394. [PubMed: 6064001]
- Murray S. Nonparametric rank-based methods for group sequential monitoring of paired censored survival data. *Biometrics*. 2000; 56:984–990. [PubMed: 11129495]
- Murray S. Using weighted Kaplan-Meier statistics in nonparametric comparisons of paired censored survival outcomes. *Biometrics*. 2001; 57:361–368. [PubMed: 11414557]
- Nelson W. Theory and applications of hazard plotting for censored failure data. *Technometrics*. 1972; 14:945–965.
- Patz A, Smith R. The ETDRS and Diabetes. *Ophthalmology*. 1991; 98:739–740. [PubMed: 2062509]
- Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*. 1989; 45:497–507. [PubMed: 2765634]
- Prentice RL. Linear rank tests with right censored data. *Biometrika*. 1978; 65:167–179.
- Shih JH. A goodness-of-fit test for association in a bivariate survival model. *Biometrika*. 1998; 85:189–200.
- Su PF, Chi YC, Li CI, Shyr Y, Liao YD. Analyzing survival curves at a fixed point in time for paired right-censored data. *Computational Statistics and Data Analysis*. 2011; 55:1617–1628.

Appendix A. Calculation of μ and σ^2

Based on Kaplan-Meier estimates,

$$KM = \int_0^\infty w(t) \{ \hat{S}_1(t) - \hat{S}_2(t) \} dt.$$

Assuming the paired observations have the common censored distribution $G(t)$, the survival function of the observed variable can be written as

$$P(X_{ki} \geq t) = P(\min(T_{ki}, C_i) \geq t) = P(T_{ki} \geq t, C_i \geq t) = S_k(t)G(t).$$

Let marginally exponential survival distributions $S_k(t) = \exp(-\lambda_k t)$; using the recommend weighting $w(t) = G(t)$, μ can be derived as

$$\mu = \int_0^{a+b} [e^{-(\lambda_1+v)t} - e^{-(\lambda_2+v)t}] dt - \frac{1}{a} \int_b^{a+b} (t-b) [e^{-(\lambda_1+v)t} - e^{-(\lambda_2+v)t}] dt.$$

Further, to derive $\sigma = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$, we need to calculate $A_k(t) = \int_t^\infty w(u)S_k(u)du$ first. Plugging $w(u)$ and $S_k(u)$ into $A_k(t)$, we have

$$A_k(t) = \begin{cases} 0 & \text{if } t > a+b \\ \int_t^{a+b} e^{-(\lambda_k+v)u} du - \frac{1}{a} \int_t^{a+b} (u-b) e^{-(\lambda_k+v)u} du & \text{if } b < t \leq a+b \\ \int_t^{a+b} e^{-(\lambda_k+v)u} du - \frac{1}{a} \int_b^{a+b} (u-b) e^{-(\lambda_k+v)u} du & \text{if } t \leq b \end{cases}.$$

Then, we can directly calculate σ_k^2 as

$$\sigma_k^2 = \lambda_k \int_0^{a+b} \frac{A_k^2(t)}{G(t)\exp(-\lambda_k t)} dt, k=1, 2.$$

Meanwhile, the covariance is

$$\sigma_{12} = \int_0^{a+b} \int_0^{a+b} A_1(t_1)A_2(t_2) \frac{G(t_1 \vee t_2)S(t_1, t_2)}{G(t_1)G(t_2)e^{-(\lambda_1 t_1 + \lambda_2 t_2)}} dA(t_1, t_2),$$

where

$$dA(t_1, t_2) = \{\lambda(t_1, t_2) - \lambda_2 \lambda_{1|2}(t_1|t_2) - \lambda_1 \lambda_{2|1}(t_2|t_1) + \lambda_1 \lambda_2\} dt_1 dt_2.$$

To derive $\lambda(t_1, t_2)$, $\lambda_{1|2}(t_1|t_2)$, and $\lambda_{2|1}(t_2|t_1)$, because

$$S(t_1, t_2) = \exp\{-(\lambda_1 t_1)^{1/\theta} + (\lambda_2 t_2)^{1/\theta}\}^\theta,$$

we can get the joint probability density function $f(t_1, t_2)$. Then, we can derive the joint hazard function and conditional hazard function of (T_{1i}, T_{2i}) as

$$\begin{aligned}\lambda(t_1, t_2) &= f(t_1, t_2) / S(t_1, t_2) \\ &= \lambda_1 \lambda_2 (\lambda_1 \lambda_2 t_1 t_2)^{1/\theta-1} \{(\lambda_1 t_1)^{1/\theta} + (\lambda_2 t_2)^{1/\theta}\}^{\theta-2} \left[\{(\lambda_1 t_1)^{1/\theta} + (\lambda_2 t_2)^{1/\theta}\}^{\theta} + \frac{1-\theta}{\theta} \right], \\ \lambda_{k|k'}(t_k | t_{k'}) &= -(\partial S(t_1, t_2) / \partial t_k) / S(t_1, t_2) = \lambda_k (\lambda_k t_k)^{1/\theta-1} \{(\lambda_1 t_1)^{1/\theta} + (\lambda_2 t_2)^{1/\theta}\}^{\theta-1},\end{aligned}$$

for $k \quad k' \in \{1, 2\}$, respectively.

Appendix B. R code for sample size calculation

In this Appendix, we provide basic R code for sample size calculation based on our proposed method. There are 8 parameters (a, b, h1, h2, theta, alpha, beta, v) in the “n.KM” function. They are: a, accrual period; b, follow-up period; h1, hazard rate λ_1 for the first group; h2, hazard rate λ_2 for the second group; theta, correlation parameter θ ; alpha, type I error rate; 1-beta, power; and v, loss to follow-up rate.

```
n.KM <- function(a, b, h1, h2, theta, alpha, beta, v){
# calculate mean
mu.1 <- function(t) exp(-h1*t)-exp(-h2*t)
mu.2 <- function(t) (t-b)*(exp(-h1*t)-exp(-h2*t))
mu <- integrate(mu.1,0,a+b)$value - (integrate(mu.2,b,a+b)$value)/a
# calculate vairance
A<-function(t,h){
if (t>(a+b)) GSt<-0
if (b<t & t<=(a+b)) {
GSt <- (exp(-h*t)-exp(-h*(a+b)))/h-(exp(-h*(a+b))*(-(a+b)/h-1/h^2)-
exp(-h*t)*(-(t/h-1/h^2))/a-b/(h*a)*exp(-h*(a+b))+ b/(h*a)*exp(-h*t)
}else{
GSt <- (exp(-h*t)-exp(-h*(a+b)))/h-(exp(-h*(a+b))*(-(a+b)/h-1/h^2)-
exp(-h*b)*(-(b/h-1/h^2))/a-b/(h*a)*exp(-h*(a+b))+ b/(h*a)*exp(-h*b)
}
}
sigma.k <- function(a, b, h0, v){
GS <- function(t, h=h0){
G <- ((t<(a+b))-((t-b)/a)*(b<=t & t<(a+b)))*exp(-v*t)
S <- exp(-h*t)
h*A(t,h)^2/(G*S)
}
integrate(Vectorize(GS), 0, a+b)$value
}
# calculate covariance
sigma12<-integrate(function(y){
sapply(y, function(t2) {integrate(function(t1) {
sapply(t1, function(t1) {
St1.t2 <- exp(-((h1*t1)^(1/theta)+ (h2*t2)^(1/theta))^(theta))
max.t <- max(t1,t2)
Gt <- ((max.t<(a+b))-((max.t-b)/a)*(b<= max.t & max.t<(a+b)))*exp(-v*max.t)
Gt1 <- ((t1<(a+b))-((t1-b)/a)*(b<= t1&t1<(a+b)))*exp(-v*t1)
Gt2 <- ((t2<(a+b))-((t2-b)/a)*(b<= t2&t2<(a+b)))*exp(-v*t2)
```

```

h.t1.t2<-h1*h2*(h1*h2*t1*t2)^(1/theta-1)*((h1*t1)^(1/theta)
+(h2*t2)^(1/theta))^(theta-2)*((h1*t1)^(1/theta)
+(h2*t2)^(1/theta))^theta+(1-theta)/theta)
h1.2<-h1*(h1*t1)^(1/theta-1)*((h1*t1)^(1/theta)+(h2*t2)^(1/theta))^(theta-1)
h2.1<-h2*(h2*t2)^(1/theta-1)*((h1*t1)^(1/theta)+(h2*t2)^(1/theta))^(theta-1)
dAt1.t2<-h.t1.t2-h2*h1.2-h1*h2.1+h1*h2
dAt1.t2*((St1.t2*Gt*A(t1,h1)*A(t2,h2))/(Gt1*Gt2*exp(-(h1*t1+h2*t2))))
}}, 0, a+b)$value
}}, 0, a+b)$value
# sample size determination
sigma <- sigma.k(a, b, h1, v)+sigma.k(a, b, h2, v)-2*sigma12
n <- (sigma*(qnorm(1-alpha/2)+qnorm(1-beta))^2)/mu^2
return(ceiling(n))
}

```

Figure B.1.

The required sample size under various accrual times and follow-up times.

Figure B.2.

The required sample size under various hazard ratio and follow-up times.

Figure B.3.
Theoretical power curves under various sample size and correlation.

Figure B.4. Sample size calculation for the (a) diabetic retinopathy study and (b) Skin grafts data under various correlation.

Table B.1

Sample size calculation (Empirical power) based on $\lambda_1 = 0.5$, $\nu = 0$, $a = 3$.

λ_2	$1 - \beta$	b	$\theta = 0.3 (\rho = 0.803)$		$\theta = 0.6 (\rho = 0.449)$		$\theta = 0.9 (\rho = 0.103)$		$\theta = 1 (\rho = 0)$	
			n_L	n_K	n_L	n_K	n_L	n_K	n_L	n_K
0.35	0.8	0	78 (.850)	58 (.822)	161 (.807)	146 (.816)	259 (.809)	260 (.811)	293 (.800)	301 (.816)
		1	54 (.861)	36 (.801)	112 (.812)	99 (.807)	180 (.825)	181 (.805)	203 (.809)	211 (.802)
		2	45 (.859)	30 (.804)	95 (.826)	84 (.802)	151 (.815)	152 (.800)	170 (.809)	175 (.804)
	0.9	0	105 (.933)	77 (.912)	215 (.904)	196 (.901)	347 (.905)	348 (.904)	393 (.907)	403 (.906)
		1	72 (.946)	48 (.903)	151 (.910)	133 (.900)	242 (.911)	242 (.903)	272 (.917)	282 (.918)
		2	60 (.941)	40 (.909)	127 (.916)	112 (.907)	203 (.911)	203 (.900)	228 (.905)	235 (.904)
0.3	0.8	0	48 (.885)	33 (.831)	87 (.836)	76 (.807)	136 (.812)	133 (.803)	154 (.805)	154 (.802)
		1	33 (.880)	20 (.796)	61 (.854)	51 (.800)	95 (.808)	92 (.803)	106 (.816)	107 (.807)
		2	28 (.903)	16 (.783)	51 (.852)	43 (.801)	79 (.817)	77 (.801)	89 (.830)	89 (.805)
	0.9	0	65 (.965)	44 (.916)	117 (.921)	101 (.908)	183 (.907)	178 (.901)	206 (.912)	207 (.906)
		1	45 (.957)	27 (.905)	82 (.923)	68 (.901)	127 (.909)	123 (.903)	142 (.906)	143 (.903)
		2	38 (.977)	22 (.899)	69 (.928)	57 (.900)	106 (.908)	102 (.896)	119 (.917)	118 (.901)
0.25	0.8	0	35 (.911)	22 (.835)	55 (.847)	45 (.819)	82 (.820)	77 (.804)	92 (.825)	89 (.811)
		1	24 (.923)	13 (.793)	38 (.845)	30 (.798)	57 (.820)	53 (.802)	63 (.825)	61 (.817)
		2	20 (.932)	11 (.812)	32 (.856)	25 (.816)	47 (.831)	43 (.800)	53 (.843)	50 (.800)
	0.9	0	47 (.975)	29 (.930)	73 (.940)	60 (.908)	110 (.920)	103 (.901)	123 (.916)	119 (.904)
		1	32 (.978)	17 (.883)	51 (.935)	39 (.899)	76 (.923)	70 (.903)	85 (.919)	82 (.903)
		2	27 (.979)	14 (.909)	43 (.944)	33 (.905)	64 (.924)	58 (.900)	71 (.924)	67 (.901)

Table B.2

Sample size calculation for loss to follow-up rate 0.1 and 0.2.

λ_2	$1 - \beta$	b	$\theta = 0.3 (\rho = 0.803)$						$\theta = 0.6 (\rho = 0.449)$						$\theta = 0.9 (\rho = 0.103)$						$\theta = 1 (\rho = 0)$																			
			n_L	n_K	n_L	n_K	n_L	n_K	n_L	n_K	n_L	n_K	n_L	n_K	n_L	n_K	n_L	n_K	n_L	n_K																				
Loss to follow up rate $\nu = 0.1$																																								
0.35	0.8	0	85	66	173	159	280	277	317	319																														
		1	60	43	124	111	200	196	225	226																														
		2	51	37	107	96	172	167	193	192																														
		0.9	0	114	88	232	213	375	370	425	427																													
	0.9	1	80	57	167	148	267	262	302	303																														
		2	69	49	144	128	230	225	258	257																														
		0.8	0	52	37	94	83	147	142	166	163																													
		1	37	24	67	57	105	100	118	115																														
	0.3	2	32	20	58	49	90	85	101	97																														
		0.9	0	70	50	126	110	197	190	223	219																													
		1	50	32	90	76	140	133	158	154																														
		2	43	27	78	65	121	113	135	130																														
0.25	0.8	0	38	25	59	49	88	82	99	94																														
		1	27	16	42	33	63	57	70	66																														
		2	23	13	36	28	54	48	60	55																														
		0.9	0	50	33	79	65	118	109	133	126																													
	0.9	1	36	21	57	44	84	76	94	88																														
		2	31	17	49	37	72	64	80	73																														
		Loss to follow up rate $\nu = 0.2$																																						
		0.35	0.8	0	92	74	188	173	302	295	342	339																												
	1		67	51	138	124	221	213	249	244																														
	2		59	45	121	110	194	186	218	211																														
	0.9		0	124	99	251	232	404	395	458	453																													
	0.3	1	90	68	184	165	295	285	333	327																														
2		78	61	162	148	259	249	291	283																															
0.8		0	57	42	102	90	160	151	180	173																														
1		42	28	75	63	116	108	131	124																															

λ_2	$1 - \beta$	b	$\theta = 0.3 (\rho = 0.803)$		$\theta = 0.6 (\rho = 0.449)$		$\theta = 0.9 (\rho = 0.103)$		$\theta = 1 (\rho = 0)$	
			n_L	n_K	n_L	n_K	n_L	n_K	n_L	n_K
	0.9	2	37	25	66	56	102	94	114	107
		0	76	56	137	120	214	202	241	232
		1	56	38	100	85	156	145	175	166
		2	49	33	88	75	136	125	153	143
0.25	0.8	0	41	28	64	53	96	87	108	100
		1	30	18	47	37	70	62	78	71
		2	27	16	42	32	62	53	68	60
		0.9	55	37	86	71	129	117	144	134
	1	1	41	25	63	49	94	83	105	95
		2	36	21	56	43	82	71	92	81

Table B.3

Empirical power (type I error rates) based on $\lambda_1 = 0.5$, $\nu = 0$, $b=0$, $a=3$.

λ_2	$1 - \beta$	Model	$\theta = 0.3 (\rho = 0.803)$			$\theta = 0.6 (\rho = 0.449)$			$\theta = 0.9 (\rho = 0.103)$			$\theta = 1.0 (\rho = 0.000)$		
			LR	KM	LR	LR	KM	LR	LR	KM	LR	LR	KM	KM
0.35	0.8	<i>n</i>	78	58	161	146	259	260	293	301				
		PSF	.834(.052)	.805(.041)	.816(.051)	.813(.055)	.819(.052)	.801(.052)	.909(.057)	.801(.041)				
		Moran	.614(.049)	.510(.057)	.700(.050)	.647(.051)	.782(.055)	.764(.055)	.800(.045)	.824(.058)				
		<i>n</i>	105	77	215	196	347	348	393	403				
0.3	0.9	PSF	.936(.051)	.913(.041)	.901(.048)	.901(.057)	.901(.041)	.903(.057)	.916(.046)	.905(.054)				
		Moran	.777(.054)	.634(.051)	.803(.054)	.773(.050)	.881(.054)	.870(.055)	.909(.053)	.901(.051)				
		<i>n</i>	48	33	87	76	136	133	154	154				
		PSF	.892(.051)	.810(.042)	.834(.048)	.804(.056)	.818(.041)	.805(.052)	.812(.051)	.804(.058)				
0.25	0.8	Moran	.683(.044)	.535(.047)	.715(.054)	.626(.055)	.784(.050)	.754(.053)	.820(.055)	.811(.053)				
		<i>n</i>	65	44	117	101	183	178	206	207				
		PSF	.955(.041)	.907(.042)	.913(.047)	.907(.051)	.905(.056)	.902(.059)	.913(.053)	.904(.049)				
		Moran	.818(.047)	.685(.057)	.843(.049)	.763(.056)	.881(.051)	.881(.052)	.919(.053)	.900(.057)				
0.25	0.9	<i>n</i>	35	22	55	45	82	77	92	89				
		PSF	.923(.040)	.815(.036)	.840(.049)	.807(.049)	.829(.054)	.803(.054)	.819(.049)	.805(.055)				
		Moran	.747(.043)	.593(.056)	.739(.047)	.640(.057)	.798(.046)	.783(.052)	.826(.051)	.814(.049)				
		<i>n</i>	47	29	73	60	110	103	123	119				
0.25	0.9	PSF	.968(.040)	.910(.037)	.933(.048)	.910(.051)	.924(.048)	.904(.053)	.911(.046)	.902(.052)				
		Moran	.868(.049)	.708(.057)	.839(.045)	.787(.058)	.900(.049)	.880(.047)	.902(.052)	.909(.053)				

Note: n is corresponding to Table 1