Computational Statistics and Data Analysis 74 (2014) 39–51

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

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

Pei-Fang Su [a], Chung-I Li [b], Yu Shyr [c,*]

[a] Department of Statistics, National Cheng Kung University, Tainan 70101, Taiwan
[b] Department of Applied Mathematics, National Chiayi University, Chiayi 60004, Taiwan
[c] Center for Quantitative Sciences, Vanderbilt University, Nashville, TN 37232, USA

ARTICLE INFO

Article history:
Received 18 December 2012
Received in revised form 30 October 2013
Accepted 21 December 2013
Available online 4 January 2014

Keywords:
Logrank test
Kaplan–Meier statistic
Hazard function
Paired observation
Positive stable frailty model

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.

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 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 visits). 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,

* Corresponding author.
E-mail address: yu.shyr@vanderbilt.edu (Y. Shyr).
URL: https://medschool.vanderbilt.edu/cqs (Y. Shyr).

0167-9473/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.12.006

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 ith $(i = 1, \ldots, n)$ 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 $N_{ki}(t) = \Delta_{ki} I(X_{ki} \le t)$ and $Y_k(t) = \sum_{i=1}^{n} Y_{ki}(t)$ count the number of individuals still at risk at time t, where $Y_{ki}(t) = I(X_{ki} \ge t)$. 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

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