Contemporary Clinical Trials xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Contemporary Clinical Trials

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

Sample size calculation for before–after experiments with partially overlapping cohorts

Song Zhang [a,*], Jing Cao [b], Chul Ahn [c]

[a] Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, TX, United States
[b] Department of Statistical Science, Southern Methodist University, Dallas, TX, United States
[c] Department of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas TX 75390-9066, United States

ARTICLE INFO

Article history:
Received 26 February 2015
Received in revised form 4 September 2015
Accepted 20 September 2015
Available online xxxx

Keywords:
Sample size
Clinical trial
Before–after study
Experimental design
Binary outcome

ABSTRACT

We investigate sample size calculation for before–after experiments where the outcome of interest is binary and the enrolled subjects contribute a mixed type of data: some subjects contribute complete pairs of before- and after-intervention outcomes, while some subjects contribute incomplete data (before-intervention only or after-intervention only). We use the GEE approach to derive a closed-form sample size formula by treating the incomplete observations as missing data in a generalized linear model. The impacts of various designing factors are appropriately accounted for in the sample size formula, including intervention effect, baseline response rate, within-subject correlation, and distribution of missing values in the before- and after-intervention periods. We illustrate sample size estimation using a real application example. We conduct simulation studies to demonstrate that the proposed sample size maintains the nominal power and type I error under a wide spectrum of trial configurations.

1. Introduction

This study is motivated by a before-and-after experiment to assess how an evidence-based colorectal cancer (CRC) prevention outreach program improves the screening rate in a socially and economically disadvantaged community. The general idea of the experiment goes as follows: in the targeted community, a baseline survey is performed on a random sample of screen-eligible adults at a local safety-net health system. A patient is asked whether he/she is willing to participate and complete a colorectal screening procedure. Then a one-year awareness campaign is conducted on colorectal cancer screening through the use of client reminders (such as letters alerting patients for need of screening), small media (such as letters discussing importance of screening), and reducing structural barriers to screening (such as making screening more convenient). After the campaign, the random survey will be repeated on the same population to evaluate the effect of the awareness campaign. The outcome of primary interest is binary, with 0 for no and 1 for yes.

The complication in this experiment is that, due to the limited size of the local safety-net health system, there is a significant overlapping between the subjects surveyed before and after intervention. For example, suppose that we randomly sample 1000 subjects at each time point, 600 of them will appear twice in the survey. As a result, a total of $n = 1400$ unique subjects will be involved in this study, among whom, we have paired observations from $n_1 = 600$ subjects, pre-intervention observations only from $n_2 = 400$ subjects, and post-intervention observation only from $n_3 = 400$ subjects. This study does not follow the typical before-and-after experimental design where each subject contributes a pair of observations, one before the intervention and one after the intervention. The reason is that, compared with performing two straightforward random samplings, tracking down every subject to obtain observations both at baseline and after intervention requires a significant extra amount of funding and manpower, which might become prohibitively expensive for a large-scale population study. Furthermore, in a socially and economically disadvantaged community, there is a great presence of homelessness and unstable housing. Even if we design the study to obtain paired outcomes from each subject, there would always be a significant amount of missing values in the collected data set. Thus, it is meaningful to develop a sample size approach for studies that involves a mixture type of data: some subjects contributing before-intervention measurements only, some after-intervention only, and some pairs of before- and after-intervention measurements.

There has been relatively limited development in the statistical inference based on paired binary outcomes with incomplete data. Ekbohm [5], Choi and Stablein [1], and Thomson [22] proposed estimators for the proportional difference based on the large sample theory. Shih [18] investigated maximum likelihood estimation and likelihood ratio test for this type of data. Tang and Tang [21], and Tang et al. [20] proposed nonparametric exact testing and estimating approaches. In this paper we present a sample size calculation method for experiments with paired binary outcomes, which appropriately accounts for the impact of missing values in the before- or after-intervention measurement.

* Corresponding author.
E-mail addresses: song.zhang@utsouthwestern.edu (S. Zhang), jcao@mail.smu.edu (J. Cao), chul.ahn@utsouthwestern.edu (C. Ahn).

http://dx.doi.org/10.1016/j.cct.2015.09.015
1551-7144/

Traditionally researchers have tried to accommodate missing values through a crude adjustment. It starts by calculating the sample size assuming no missing data, denoted by n_0. When every subject contributes a complete pair of outcomes, the McNemar's test is the most popular approach to detect the before–after difference [2], and sample size calculation for the McNemar's test is well established in statistical literature [12,19,3,8,6]. Once n_0 is obtained, the final sample size is calculated by n_0/w, where w is the proportion of subjects who are expected to contribute complete data among all enrolled subjects.

We propose to adjust for missing data through a generalized estimating equation (GEE) approach [9]. GEE has long been recognized as a robust method to model correlated data and accommodate missing values in studies involving longitudinal and clustered observations [23,13]. Sample size calculation based on the GEE approach has been explored by many researchers. For example, Liu and Liang [10] developed a sample size formula based on a generalized score test. Rochon [16] proposed a sample size formula using a non-central version of the Wald χ^2 test statistics. Jung and Ahn [7] investigated sample size calculation to compare rates of change between two treatment groups. Zhang and Ahn [24] developed a sample size formula for the test of time-averaged difference accounting for missing values. In this study we present a closed-form sample size formula for before–after studies with partially overlapping cohorts. When there is no missing data, the proposed sample size is very close to that calculated based on the McNemar's test. When there is missing data, however, because the proposed sample size appropriately accounts for partial information from incomplete pairs, it can lead to a substantial saving compared with the crude adjustment approach. The sample size formula explicitly shows the factors that affect the impact of missing values.

The rest of the paper is organized as follows. In Section 2 we derive the sample size formula based on the GEE approach. Simulation studies were presented in Section 3. We demonstrate the proposed method using a real application example in Section 4. Finally, we discuss limitations and potential extensions in Section 5.

2. A GEE sample size approach

We first present the derivation under no missing data. Let y_{it} be the binary outcome (0 for no and 1 for yes) from subject $i(i = 1,\ldots, n)$ at time t. We use $t = 0$ and 1 to denote the before- and after-intervention periods, respectively. We model y_{it} by a logistic regression model:

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \beta_1 + \beta_2 t = X_{it}'\beta \tag{1}$$

where $p_{it} = P(y_{it} = 1), \beta_1 = \log(\frac{p_{i0}}{1-p_{i0}})$ models the log-transformed baseline odds and $\beta_2 = \log(\frac{p_{i1}}{1-p_{i1}}) - \log(\frac{p_{i0}}{1-p_{i0}})$ is the log-transformed odds ratio between the after- and before-intervention responses. Thus the intervention effect is represented by β_2, and we are interested in testing the null hypothesis H_0: $\beta_2 = 0$ versus H_a: $\beta_2 \neq 0$. We present the model in a matrix form, with $X_{it} = (1, t)'$ being the vector of covariates at time t and $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ being the vector of regression parameters. We further define $\rho = \mathrm{Corr}(y_{i0}, y_{i1})$ to be the within-subject correlation coefficient. We assume the responses to be independent across different subjects, $\mathrm{Corr}(y_{it}, y_{i't'}) = 0$ for $i \neq i'$. Thus the statistical properties of paired outcomes (y_{i0}, y_{i1}) are fully described by (β_1, β_2, ρ).

Table 1
Probabilities of pre- and post-intervention outcomes.

	Post-intervention		
Pre-intervention	No	Yes	
No	h_{00}	h_{01}	$1-p_0$
Yes	h_{10}	h_{11}	p_0
	$1-p_1$	p_1	

First we estimate the intervention effect through the GEE approach. Under an independent working correlation structure, the GEE estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$ is obtained by solving

$$S_n(\beta) = \sum_{i=1}^{n}\sum_{t=0}^{1}[y_{it} - p_{it}(\beta)]X_{it} = 0.$$

Here $p_{it}(\beta) = \exp(X_{it}'\beta)/[1 + \exp(X_{it}'\beta)]$ is implied by Eq. (1). The Newton–Raphson algorithm can be employed to obtain a numerical solution. At the $(m+1)$th iteration,

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \boldsymbol{A}_n^{-1}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)S_n\left(\hat{\boldsymbol{\beta}}^{(m)}\right),$$

where $A_n(\beta) = -\partial S_n(\beta)/\partial\beta$. Liang and Zeger [9] showed that $\sqrt{n}(\hat{\beta} - \beta)$ is approximately normal with mean zero and the variance is consistently estimated by $\Sigma_n = nA_n^{-1}(\hat{\beta})V_n(\hat{\beta})A_n^{-1}(\hat{\beta})$, where

$$V_n\left(\hat{\beta}\right) = \sum_{i=1}^{n}\left[\left(\sum_{t=0}^{1}\hat{\epsilon}_{it}X_{it}\right)\left(\sum_{t=0}^{1}\hat{\epsilon}_{it}X_{it}\right)'\right]$$

with $\hat{\epsilon}_{it} = y_{it} - p_{it}(\hat{\beta})$. Letting $\hat{\sigma}_2^2$ be the (2,2)th element of Σ_n, we reject H_0: $\beta_2 = 0$ if $|\sqrt{n}\hat{\beta}_2/\hat{\sigma}_2| > z_{1-\alpha/2}$. Here α is the significance level and $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution.

Let σ_2^2 be the true variance of the GEE estimator $\hat{\beta}_2$ under true parameters (β_1, β_2, ρ). If the alternative hypothesis is true, $\beta_2 \neq 0$, in order to achieve a testing power of $1-\gamma$ with a type I error of α, the required sample size is solved from equation $P(|\sqrt{n}\beta_2/\sigma_2| > z_{1-\alpha/2}|H_a) = 1-\gamma$. The solution is

$$n = \frac{\sigma_2^2\left(z_{1-\alpha/2} + z_{1-\gamma}\right)^2}{\beta_2^2}. \tag{2}$$

In the following theorem we present a closed-form expression for σ_2^2, which leads to a closed-form GEE sample size formula under no missing data.

Theorem 1. Let $p_t = p_{it}(\beta), t = 0, 1$, be the true response rates shared by all subjects. We define $\tau_t^2 = p_t(1-p_t)$. As $n \to \infty$, σ_2^2 has a closed-form

$$\sigma_2^2 = \frac{\tau_0^2 + \tau_1^2 - 2\rho\tau_0\tau_1}{\tau_0^2\tau_1^2}. \tag{3}$$

Proof. See Appendix A. □

2.1. A sample size to accommodate missing data

To accommodate the scenarios that a portion of subjects are likely to miss the before- or after-intervention assessment, we introduce $\delta_{it} = 0/1$ to indicate that the outcome of the ith subject at time t $(t = 0, 1)$ is missing/observed. The probability that a subject completes the outcome at time t is denoted by $q_t = E(\delta_{it})$. We impose the constraints that $P(\delta_{i0} = \delta_{i1} = 0) = 0$, i.e., each subject contributes at least one of the before- or after-intervention outcomes. Under this constraint, it can be shown that the proportion of subjects with complete pairs of outcomes is $P(\delta_{i0} = \delta_{i1} = 1) = q_0 + q_1 - 1$. Thus we impose the second constraint, $q_0 + q_1 > 1$. We demonstrate that, in the presence of incomplete pairs, the GEE estimator of β_2 still has a closed-form expression for variance.

Please cite this article as: S. Zhang, et al., Sample size calculation for before–after experiments with partially overlapping cohorts, Contemp Clin Trials (2015), http://dx.doi.org/10.1016/j.cct.2015.09.015