Published in final edited form as:

Contemp Clin Trials. 2018 January ; 64: 274–280. doi:10.1016/j.cct.2015.09.015.

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

Song Zhang^{a,*}, Jing Cao^b, and Chul Ahn^c

^aDepartment of Clinical Sciences, UT Southwestern Medical Center, Dallas, TX, United States

^bDepartment of Statistical Science, Southern Methodist University, Dallas, TX, United States

^cDepartment of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas TX 75390-9066, United States

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.

Keywords

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

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

*Corresponding author: song.zhang@utsouthwestern.edu (S. Zhang).

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.

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 $i(i=1, \dots, n)$ 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:

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

where $p_{it} = P(y_{it}=1)$, $\beta_1 = \log \left(\frac{p_{i0}}{1 - p_{i0}} \right)$ models the log-transformed baseline odds and $\beta_2 = \log \left(\frac{p_{i1}}{1 - p_{i1}} \right) - \log \left(\frac{p_{i0}}{1 - p_{i0}} \right)$ 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 $\beta=(\beta_1, \beta_2)'$ being the vector of regression parameters. We further define $\rho = \text{Corr}(y_{i0}, y_{i1})$ to be the within-subject correlation coefficient. We assume the responses to be independent across different subjects, $\text{Corr}(y_{it}, y_{i't'})=0$ for $i \neq i'$. Thus the statistical properties of paired outcomes (y_{i0}, y_{i1}) are fully described by (β_1, β_2, ρ) .

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

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

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

$\sum_n = n A_n^{-1}(\hat{\beta}) V_n(\hat{\beta}) A_n^{-1}(\hat{\beta})$, where

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

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- α /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

$P(|\sqrt{n}\hat{\beta}_2 / \sigma_2| > z_{1-\alpha/2} | H_a) = 1 - \gamma$. The solution is

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

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 $n \rightarrow \infty$, σ_2^2 has a closed-form

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

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 i th 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 $P(\delta_0=\delta_1=0)=0$, 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 $P(\delta_0=\delta_1=1)=q_0+q_1-1$. 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.

Theorem 2—Let q_t ($t=0,1$) be the proportion of subjects who have before- and after-intervention outcomes, respectively. As $n \rightarrow \infty$, σ_2^2 has a closed form,

$$\sigma_2^2 = \frac{q_0\tau_0^2 + q_1\tau_1^2 - 2(q_0 + q_1 - 1)\rho\tau_0\tau_1}{q_0q_1\tau_0^2\tau_1^2}.$$

Proof—See Appendix B.

Thus we have a general sample size formula based on GEE that accommodates potential incomplete observations:

$$n_{GEE} = \frac{[q_0\tau_0^2 + q_1\tau_1^2 - 2(q_0 + q_1 - 1)\rho\tau_0\tau_1] (z_{1-\alpha/2} + z_{1-\gamma})^2}{q_0q_1\tau_0^2\tau_1^2\beta_2^2}. \quad (4)$$

Eq. (4) suggests that besides observation probabilities q_0 and q_1 , the impact of incomplete observations depends on three factors: the baseline response rate p_0 through τ_0^2 , the after-intervention response rate p_1 through τ_1^2 , and within-subject correlation ρ . With all other factors fixed, a stronger correlation is associated with a smaller sample size.

Using the GEE method, Liu and Liang [10], and Pan [15] developed sample size calculation for clinical trials under the general scenario of correlated binary outcomes, which includes the before–after experiments investigated in this study. However, these methods do not account for missing data. Dang et al. [4] proposed a sample size calculation for correlated binary outcomes based on the generalized liner mixed model (GLIMMIX) approach, which can accommodate missing data. However, their approach assumes the observations to be independent between the intervention and control groups. It is inapplicable to the before–after study where outcomes under intervention and control are observed from the same subjects, and thus correlated.

It is noteworthy that under the scenario of before–after studies with complete data ($q_0=q_1=1$), the proposed sample size is equivalent to that developed by Pan [15]. Importantly, under this scenario, the GEE estimators of regression parameters and their

variances are the same whether the working correlation is independent or equals to the true correlation [11], i.e., there is no information loss by using an independent working correlation. As a result, Eqs. (21) and (20) by Pan [15], which are the variances of $\hat{\beta}_2$ under the independent working correlation and true working correlation, respectively, both equal to Eq. (3) of σ_2^2 in this study.

2.2. Sample size based on the McNemar's test

For the special case that all subjects contribute complete pairs of outcomes ($q_0=q_1=1$), the required sample size can be calculated based on the McNemar's test [3]:

$$n_{MN} = \frac{\left[z_{1-\alpha/2} \sqrt{h_{01} + h_{10}} + z_{1-\gamma} \sqrt{h_{01} + h_{10} - (h_{01} - h_{10})^2} \right]^2}{(h_{01} - h_{10})^2}, \quad (5)$$

where

$$h_{uv} = P(y_{i0}=u, y_{i1}=v), \text{ for } u=0, 1 \text{ and } v=0, 1,$$

are joint probabilities of paired outcomes (y_{i0}, y_{i1}) . The relationship between h_{uv} and (p_0, p_1) is presented in Table 1. Note that (p_0, p_1, ρ) are exchangeable with the joint probabilities. We first have

$$h_{11} = E(y_{i0}y_{i1}) = E(y_{i0})E(y_{i1}) + Cov(Y_{i0}, Y_{i1}) = p_0p_1 + \rho \sqrt{p_0(1-p_0)p_1(1-p_1)},$$

and the other probabilities: $h_{10}=p_0-h_{11}$, $h_{01}=p_1-h_{11}$, and $h_{00}=1-h_{11}-h_{10}-h_{01}$.

Note that n_{MN} only requires the specification of (h_{01}, h_{10}) . Suppose we have two scenarios: $(h_{01}=0.2, h_{10}=0.1, h_{11}=0.05)$ versus $(h_{01}=0.2, h_{10}=0.1, h_{11}=0.25)$. Due to equal values of (h_{01}, h_{10}) , n_{MN} would produce equal sample sizes based on Eq. (5). For n_{GEE} , the design configurations are different: $(p_0=0.15, p_1=0.25, \rho=0.08)$ versus $(p_0=0.35, p_1=0.45, \rho=0.39)$, hence different sample sizes based on Eq. (4).

In Fig. 1, we compare the sample sizes calculated based on n_{MN} and n_{GEE} under complete observations ($q_0=q_1=1$). The title of each graph presents the discordant probabilities (h_{10}, h_{01}) under investigation. The horizontal axis shows the value of h_{11} . Note that the valid ranges of h_{11} are different across graphs. Despite the changing values of h_{11} , the sample size based on n_{MN} remains the same within each graph. From each combination of (h_{10}, h_{01}, h_{11}) we can obtain the corresponding (p_0, p_1, ρ) , and then the GEE sample size n_{GEE} can be calculated based on Eq. (4). The vertical axis shows the sample size ratio, n_{GEE}/n_{MN} . The solid curve plots the sample size ratio under various design configurations. It is not smooth due to the integer constraint on sample sizes. The horizontal dashed line indicates where $n_{GEE}/n_{MN}=1$. The legend shows the value of n_{MN} . It can be shown that as h_{11} increases over its valid range, the corresponding within-subject correlation ρ follows a $\cap$ -shape. For

example, under $(h_{10}=0.2, h_{01}=0.3)$, as h_{11} moves from 0 to 0.5, ρ first increases from -0.17 to 0.41 and then decreases to -0.17 . Recall that with all other factors fixed, n_{GEE} is negatively associated with ρ . Thus we observed a general *U*-shape in the sample size ratio curves. Among all the scenarios explored in Fig. 1, the sample sizes based on the McNemar's test have a ranges of 234 to 2039, and in most cases n_{GEE} and n_{MN} are close, within $\pm 1\%$ to each other. Thus we conclude that under complete data, the sample size obtained based on the McNemar's test is similar to that derived by the proposed GEE method. The GEE sample size method, however, provides greater flexibility to accommodate incomplete data.

3. Simulation

We conduct simulation studies to examine whether the proposed sample size n_{GEE} maintains the desired power and type I error, with incomplete observations in the before- and after-intervention measurements. We specify design configurations by (p_0, ρ, q_0, q_1) with $\rho = p_1 - p_0$ representing the intervention effect. There is a one-to-one correspondence between (p_0, ρ) and logistic regression coefficients (β_1, β_2) , but we present (p_0, ρ) due to its straightforward interpretation. We examine two levels of baseline response rate, $p_0=0.15$ and 0.4 ; two levels of intervention effect, $\rho=0.05$ and 0.1 ; and four levels of within-subject correlation, $\rho=-0.15, 0, 0.15, 0.3$. We choose the above four levels of ρ because they cover negative correlation, independence, and positive correlation, and they are within the valid ranges of correlation under all pairs of (p_0, ρ) considered. We also consider four missing patterns, denoted by

$$\begin{aligned} \mathbf{q}^{(1)} &= (q_0=1, q_1=1); \quad \mathbf{q}^{(2)} = (q_0=0.85, q_1=0.85); \\ \mathbf{q}^{(3)} &= (q_0=0.9, q_1=0.8); \quad \mathbf{q}^{(4)} = (q_0=0.8, q_1=0.9). \end{aligned}$$

Here $\mathbf{q}^{(1)}$ represents the scenario of no missing data, $\mathbf{q}^{(2)}$ the scenario of balanced distribution of missing values in the before- and after-intervention outcomes, and $\mathbf{q}^{(3)}$ and $\mathbf{q}^{(4)}$ the scenarios of unbalanced distributions of missing values.

For each combination of $(p_0, \rho, \mathbf{q})$, we calculate the required sample size n_{GEE} . Then 5000 data sets are generated, each containing n_{GEE} pairs of binary outcomes following (p_0, ρ) . We impose incomplete data in the before- and after-intervention outcomes according to missing pattern $\mathbf{q}$. We obtain the GEE estimator of β_2 and test hypothesis $H_0: \beta_2=0$. The empirical type I error and power are evaluated as the proportion of times that H_0 is rejected under the null and alternative hypothesis, respectively. The simulation study is programmed in R. The R code is available upon request from the first author.

The simulation results are presented in Table 2. Each cell shows the sample size (empirical power, empirical type I error) under trial configuration $(p_0, \rho, \mathbf{q})$. Furthermore, under each row corresponding to complete data ($\mathbf{q}^{(1)}$), we have presented in $[\cdot]$ the sample sizes calculated based on the approach by Liu and Liang [10]. We have several observations. Firstly, the trial configurations explored in Table 2 lead to a wide range in sample size, 181 to 2039, but the empirical power and type I error are all close to the nominal levels. Thus we demonstrate that the proposed sample size method maintains the nominal power and type I

error under a wide spectrum of trial configurations. Secondly, the within-subject correlation ρ has a great impact on sample size. Under $p_0=0.4$, $\rho=0.05$, and no missing data, as ρ increases from -0.15 to 0.3 , the required sample size decreases from 1767 to 1076, a 39% reduction. Thirdly, it is also noteworthy that for trials with a binary outcome, the baseline response rate p_0 has a profound impact on sample size requirement. For example, with the same intervention effect $\rho=0.05$, correlation $\rho=0.15$, and missing pattern $\mathbf{q}^{(4)}=(q_0=0.8, q_1=0.9)$, as the baseline response rate p_0 change from 0.15 to 0.4, the sample size increases 69%, from 938 to 1588. The reason is that variance $\tau^2_0=p_0(1-p_0)$ has a $\cap$ with respect to p_0 , increasing as p_0 changing from 0 to 0.5 and decreasing as p_0 changing from 0.5 to 1. Fourthly, under complete data, the proposed sample sizes generally agree with those calculated by Liu and Liang [10], although the latter are slightly smaller. The sample size by Liu and Liang [10] is calculated based on a quasi-score test statistic. It appears to be more efficient under the scenario of complete data. However, the proposed sample size approach offers greater flexibility in accounting for missing data. Finally, the proposed method leads to significant saving in sample size compared with the crude adjustment for missing data (n_0/w), as we described in Section 1. For example, under ($p_0=0.4$, $\rho=0.1$, $\rho=0.15$), the sample size under complete data is $n_0=332$. Furthermore, under missing patterns $\mathbf{q}^{(2)}-\mathbf{q}^{(4)}$, the proportions of patients who are expected to contribute complete pairs of outcomes are equal, $w=q_0+q_1-1=0.7$. Thus with the traditional adjustment for incomplete observations, the required sample size is $332/0.7=474$, much larger than the sample sizes under the proposed approach. In real clinical practice, the saving in sample size not only leads to more efficient utilization of resources, but also leads to a smaller number of patients exposed to the potential risk of experimental treatment.

We have performed an additional simulation study where the missing patterns considered are

$$\mathbf{q}^{(1*)}=(q_0^*=1, q_1^*=1); \mathbf{q}^{(2*)}=(q_0^*=0.72, q_1^*=0.72); \\ \mathbf{q}^{(3*)}=(q_0^*=0.86, q_1^*=0.58); \mathbf{q}^{(4*)}=(q_0^*=0.58, q_1^*=0.86),$$

comparable to the motivating example described in Section 1. The results are presented in Table 3. Similar conclusions can be obtained as those based on Table 2.

4. Example

The Task Force on Community Preventive Services has recommended that cancer control programs use a multi-pronged approach to improve community demand for and access to cancer screening [14]. Recommendations included: 1) use of client reminders, such as mails advising patients that they are due for screening, 2) use of small media, such as informational letters on the importance of screening, and 3) reduction of structural barriers to screening, such as making screening more convenient, and eliminating complex administrative procedures and need for multiple clinic visits. We would like to investigate if the one-year implementation of CRC prevention program will significantly improve the rate of willingness to participate and complete a colorectal cancer screening at a local safety-net health system. The current average colorectal cancer screening rate is 21% in the local safety-net hospital. Thus we assume the rate of willingness to participate and complete a

colorectal cancer screening to be $p_0=21\%$ at baseline. We also assume that the rate will be improved to 30% after implementation of CRC prevention program ($=30\%-21\%=9\%$). There is no preliminary data on within-subject correlation for CRC screening. An unpublished dataset on a survey about breast cancer screening showed a correlation around 0.1. Hence we consider a range of ρ : 0.05, 0.1, 0.15, 0.2. To test the hypotheses $H_0: p_0=p_1$ versus $H_1: p_0 \neq p_1$ with 80% power at 5% two-sided significance level, the number of subjects needs to be determined. Under the assumption of no missing data, the numbers of subjects required are 353 (352), 335 (334), 316 (316), 298 (298) under the proposed (McNemar) approach, given within-subject correlation $\rho=0.05, 0.1, 0.15, 0.2$, respectively.

When 20% of patients do not respond to the survey before or after the intervention ($q_0=q_1=0.8$), the required sample sizes are 447, 430, 413, 395 based on the proposed sample size, given $\rho=0.05, 0.1, 0.15, 0.2$. If we only include the portion ($q_0+q_1-1=60\%$) of patients with complete pairs of observations into analysis, we would need to enroll 557 unique patients.

The above example shows that different values of ρ have great impact on sample size requirement. In practice, the information or data to estimate design parameters is usually limited. The advantage of the proposed method is that the sample size formula Eq. (4) has a closed form, which allows investigators to evaluate the impact of different factors analytically. We suggest to consider a range of plausible values for design factors (such as ρ, p_0, q_0, q_1), and conduct sensitivity analysis to assess the operational characteristics of clinical trials when the actual design factors deviate from the assumed values. The sensitivity analysis also allows us to determine a sample size that achieves acceptable performance under majority of plausible scenarios.

5. Discussion

In this study we use the GEE approach to derive a closed form sample size formula for before–after experiments with partially overlapping cohorts. The incomplete observations (before-intervention only and after-intervention only) are accounted for as missing data in a generalized linear model. The proposed sample size formula is flexible to accommodate arbitrary types of missing patterns, including complete data as a special case. We further demonstrate that the proposed sample size is very close to that calculated based on the McNemar's test when all subjects contribute complete pairs of observations, but leads to significant saving in sample size when some subjects only contribute incomplete data. One potential limitation is that the GEE approach is based on large sample approximation, which means that the proposed sample size might not perform well when the required sample size is small due to, say, a large intervention effect. Our simulation study explores a wide spectrum of trial configurations and shows that the proposed sample size preserves the nominal power and type I error when the sample sizes range from 2039 to 181. One of our future research topics is to investigate the extension of nonparametric sample size approaches to the scenarios of partially overlapping cohorts.

Zhang et al. [25] considered a special scenario in before- and after-intervention studies where all subjects participated in the pre-intervention assessment but some did not come back for

the after-intervention assessment. That is, missing data only occur in the post-intervention measurements. Their sample size formula is a special case of Eq. (4) by setting $q_0=1$. The proposed sample size can also be applied to paired experiments. For example, Romeo et al. [17] presented a study where each patient was examined for brain involvement in myotonic dystrophy type I by two devices: single photon emission tomography (SPECT) and positron emission tomography (PET). The primary goal is to compare the diagnostic performance of SPECT and PET. The sample size formula Eq. (4) can be employed to account for the possibility that missing data can occur for either devices.

Acknowledgments

The work was supported in part by NIH grants 1UL1TR001105 and 1R03AG039689-01A1, AHRQ grant R24HS22418, CPRIT grants RP110562-C1 and RP120670-C1, and NSF grant IIS-1302497-02.

Appendix A Proof of Theorem 2

First we rewrite Σ_n as $\sum_n = [A_n(\hat{\beta})/n]^{-1} [V_n(\hat{\beta})/n] [A_n(\hat{\beta})/n]^{-1}$. The detailed expression of $A_n(\beta)$ is

$$A_n(\beta) = -\partial S_n(\beta) / \partial \beta = \sum_{i=1}^n \sum_{t=0}^1 p_{it}(\beta) [1 - p_{it}(\beta)] X_{it} X'_{it}.$$

As $n \rightarrow \infty$, it is straightforward that

$$\begin{aligned} A_n(\hat{\beta})/n &\rightarrow A(\beta) = \sum_{t=0}^1 \tau_t^2 X_{it} X'_{it} = \begin{pmatrix} \tau_0^2 + \tau_1^2 & \tau_1^2 \\ \tau_1^2 & \tau_1^2 \end{pmatrix}, \\ V_n(\hat{\beta})/n &\rightarrow V(\beta) = \sum_{t=0}^1 \sum_{t'=0}^1 \rho_{tt'} \tau_t \tau_{t'} X_{it} X'_{it'} = \begin{pmatrix} \tau_0^2 + 2\rho\tau_0\tau_1 + \tau_1^2 & 2\rho\tau_0\tau_1 + \tau_1^2 \\ \tau_1^2 & \tau_1^2 \end{pmatrix}. \end{aligned}$$

Here we have defined $\rho_{tt}=1$ if $t=t'$, and $\rho_{tt}=\rho$ otherwise. The derivation uses the fact that $\text{Var}(y_{it})=p(1-p)$ and $\text{Corr}(y_{it}, y_{it'})=\rho_{tt'}$. Thus the variance matrix of the GEE estimator, Σ_n , approaches to

$$\Sigma = A^{-1}(\beta) V(\beta) A^{-1}(\beta).$$

Because

$$A^{-1}(\beta) = \frac{1}{\tau_0^2 \tau_1^2} \begin{pmatrix} \tau_1^2 & -\tau_1^2 \\ -\tau_1^2 & \tau_0^2 \tau_1^2 \end{pmatrix},$$

a few steps of algebra show that the (2,2)th element of Σ is

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

Appendix B Proof of Theorem 2.1

With the inclusion of δ_{it} ($t=0,1$), the equations in the previous section are updated accordingly,

$$\begin{aligned} S_n(\beta) &= \sum_{i=1}^n \sum_{t=0}^1 \delta_{it} [y_{i0} - p_{i0}(\beta)] X_{i0}, \\ A_n(\beta) &= \sum_{i=1}^n \sum_{t=0}^1 \delta_{it} p_{it}(\beta) [1 - p_{it}(\beta)] X_{it} X'_{it}, \\ V_n(\hat{\beta}) &= \sum_{i=1}^n \left[\left(\sum_{t=0}^1 \delta_{it} \hat{e}_{it} X_{it} \right) \left(\sum_{t=0}^1 \delta_{it} \hat{e}_{it} X_{it} \right)' \right] \end{aligned}$$

As $n \rightarrow \infty$, it is straightforward that

$$\begin{aligned} A_n(\hat{\beta})/n &\rightarrow A(b) = \begin{pmatrix} q_0\tau_0^2 + q_1\tau_1^2 & q\tau_1^2 \\ q\tau_1^2 & q\tau_1^2 \end{pmatrix}, \\ V_n(\hat{\beta})/n &\rightarrow V(b) = \begin{pmatrix} q_0\tau_0^2 + 2(q_0 + q_1 - 1)\rho\tau_0\tau_1 + q_1\tau_1^2 & (q_0 + q_1 - 1)\rho\tau_0\tau_1 + q_1\tau_1^2 \\ (q_0 + q_1 - 1)\rho\tau_0\tau_1 + q_1\tau_1^2 & q_1\tau_1^2 \end{pmatrix}, \end{aligned}$$

After a few steps, we have

$$\sigma_2^2 = \frac{q_0\tau_0^2 + q_1\tau_1^2 - 2(q_0 + q_1 - 1)\rho\tau_0\tau_1}{q_0q_1\tau_0^2\tau_1^2}.$$

References

1. Choi SC, Stablein DM. Practical tests for comparing two proportions with incomplete data. *Appl Stat.* 1982; 31:256–262.
2. Cochran WG. The comparison of percentages in matched samples. *Biometrika.* 1950; 37(4):256–266. [PubMed: 14801052]
3. Connor RJ. Sample size for testing differences in proportions for the paired-sample design. *Biometrics.* 1987; 43(1):207–211. [PubMed: 3567305]
4. Dang Q, Mazumdar S, Houck P. Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. *Comput Methods Prog Biomed.* 2008; 91(2):122–127.
5. Ekbohm G. On testing the equality of proportions in the paired case with incomplete data. *Psychometrika.* 1982; 47(1):115–118.
6. Julious SA, Campbell MJ, Altman DG. Estimating sample sizes for continuous, binary, and ordinal outcomes in paired comparisons: practical hints. *J Biopharm Stat.* 1999; 9(2):241–251. [PubMed: 10379691]
7. Jung S, Ahn CW. Sample size for a two-group comparison of repeated binary measurements using GEE. *Stat Med.* 2005; 24(17):2583–2596. [PubMed: 16118812]
8. Lachin JM. Power and sample size evaluation for the McNemar test with application to matched case–control studies. *Stat Med.* 1992; 11(9):1239–1251. [PubMed: 1509223]
9. Liang K, Zeger SL. Longitudinal data analysis for discrete and continuous outcomes using generalized linear models. *Biometrika.* 1986; 84:3–32.
10. Liu G, Liang K. Sample size calculations for studies with correlated observations. *Biometrics.* 1997; 53(3):937–947. [PubMed: 9290224]
11. McDonald B. Estimating logistic regression parameters for bivariate binary data. *J R Stat Soc Ser B.* 1993; 55:391–397.

12. Miettinen OS. The matched pairs design in the case of all-or-none responses. *Biometrics*. 1968; 24(2):339–352. [PubMed: 5683874]
13. Norton EC, Bieler GS, Ennett ST, Zarkin GA. Analysis of prevention program effectiveness with clustered data using generalized estimating equations. *J Consult Clin Psychol*. 1996; 64(5):919–926. [PubMed: 8916620]
14. Task Force on Community Preventive Services. Recommendations for client- and provider-directed interventions to increase breast, cervical, and colorectal cancer screening. *Am J Prev Med*. 2008; 35:S21–S25. [PubMed: 18541184]
15. Pan W. Sample size and power calculations with correlated binary data. *Control Clin Trials*. 2001; 22(3):211–227. [PubMed: 11384786]
16. Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. *Stat Med*. 1998; 17(14):1643–1658. [PubMed: 9699236]
17. Romeo V, Pegoraro E, Squarzanti F, Sorarù G, Ferrati C, Ermani M, Zucchetta P, Chierichetti F, Angelini C. Retrospective study on pet-spect imaging in a large cohort of myotonic dystrophy type 1 patients. *Neurol Sci*. 2010; 31(6):757–763. [PubMed: 20842397]
18. Shih WJ. Maximum likelihood estimation and likelihood ratio test with incomplete pairs. *J Stat Comput Simul*. 1987; 21:187–194.
19. Shork M, Williams G. Number of observations required for the comparison of two correlated proportions. *Commun Stat Simul Comput*. 1980; 9(4):349–357.
20. Tang M, Ling M, Tian G. Exact and approximate unconditional confidence intervals for proportion difference in the presence of incomplete data. *Stat Med*. 2009; 28(4):625–641. [PubMed: 19035467]
21. Tang M, Tang N. Exact tests for comparing two paired proportions with incomplete data. *Biom J*. 2004; 46(1):72–82.
22. Thomson PC. A hybrid paired and unpaired analysis for the comparison of proportions. *Stat Med*. 1995; 14(13):1463–1470. [PubMed: 7481184]
23. Zeger SL, Liang K, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*. 1988; 44(4):1049–1060. [PubMed: 3233245]
24. Zhang S, Ahn C. Sample size calculation for time-averaged differences in the presence of missing data. *Contemp Clin Trials*. 2012; 33(3):550–556. [PubMed: 22553832]
25. Zhang S, Cao J, Ahn C. A GEE approach to determine sample size for pre- and post-intervention experiments with dropout. *Comput Stat Data Anal*. 2014; 69:114–121.

Fig. 1.

The title of each graph presents the discordancy probabilities (h_{10}, h_{01}) explored. The horizontal axis shows the value of h_{11} . The valid ranges of h_{11} are different over graphs. The vertical axis shows the ratio between sample sizes, n_{GEE}/n_{MN} , under complete observations ($q_0=q_1=1$). The GEE sample size are calculated based on configurations (p_0, p_1, ρ) , each implied by a particular combination of (h_{10}, h_{01}, h_{11}) . The horizontal dashed line indicates where the sample sizes are equal between the two methods. The legend shows the value of n_{MN} . The curves of sample size ratio are not smooth due to the integer constraint on sample sizes.

Table 1

Probabilities of pre- and post-intervention outcomes.

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

Table 2

Performance of sample sizes based on the GEE method.

(ρ, ρ_0)	$\rho = -0.15$	$\rho = 0$	$\rho = 0.15$	$\rho = 0.3$
$\mathbf{q}^{(1)}$ (0.15, 0.05)	1048(0.796, 0.050) [1038]	912(0.812, 0.053) [903]	776(0.808, 0.053) [769]	641(0.804, 0.050) [634]
$\mathbf{q}^{(2)}$	1205(0.794, 0.051)	1073(0.803, 0.051)	942(0.809, 0.053)	810(0.814, 0.051)
$\mathbf{q}^{(3)}$	1216(0.805, 0.051)	1084(0.810, 0.048)	952(0.803, 0.045)	820(0.803, 0.050)
$\mathbf{q}^{(4)}$	1202(0.814, 0.049)	1070(0.799, 0.050)	938(0.801, 0.045)	806(0.800, 0.045)
$\mathbf{q}^{(1)}$ (0.15, 0.10)	294(0.810, 0.047) [284]	256(0.808, 0.047) [248]	219(0.818, 0.050) [211]	181(0.812, 0.052) [175]
$\mathbf{q}^{(2)}$	338(0.815, 0.052)	301(0.814, 0.047)	265(0.807, 0.049)	228(0.814, 0.050)
$\mathbf{q}^{(3)}$	342(0.812, 0.050)	306(0.817, 0.047)	269(0.814, 0.051)	233(0.822, 0.048)
$\mathbf{q}^{(4)}$	336(0.810, 0.047)	299(0.811, 0.047)	262(0.806, 0.051)	226(0.813, 0.054)
$\mathbf{q}^{(1)}$ (0.40, 0.05)	1767(0.810, 0.049) [1761]	1536(0.804, 0.052) [1531]	1306(0.793, 0.054) [1301]	1076(0.800, 0.055) [1072]
$\mathbf{q}^{(2)}$	2031(0.795, 0.054)	1807(0.810, 0.053)	1584(0.803, 0.048)	1361(0.798, 0.048)
$\mathbf{q}^{(3)}$	2039(0.792, 0.047)	1815(0.787, 0.044)	1591(0.809, 0.052)	1367(0.812, 0.050)
$\mathbf{q}^{(4)}$	2036(0.790, 0.053)	1812(0.807, 0.048)	1588(0.801, 0.051)	1364(0.802, 0.054)
$\mathbf{p}^{(1)}$ (0.40, 0.10)	449(0.804, 0.052) [443]	390(0.806, 0.053) [385]	332(0.798, 0.057) [327]	273(0.810, 0.056) [270]
$\mathbf{p}^{(2)}$	516(0.795, 0.047)	459(0.805, 0.047)	403(0.806, 0.048)	346(0.794, 0.047)
$\mathbf{p}^{(3)}$	518(0.801, 0.049)	461(0.801, 0.051)	404(0.805, 0.051)	348(0.803, 0.050)
$\mathbf{p}^{(4)}$	517(0.794, 0.057)	460(0.801, 0.046)	403(0.800, 0.047)	347(0.804, 0.053)

Each cell shows the sample size (empirical power, empirical type I error) under trial configuration (ρ_0, ρ) . Here ρ_0 is the baseline response rate, $\rho = \rho_1 - \rho_0$ represents the intervention effect, ρ denotes within-subject correlation, and $\mathbf{q}$ indicates missing data pattern. For comparison, under the rows corresponding to complete data ($\mathbf{q}^{(1)}$), we present in [·] the sample sizes calculated based on the approach by Liu and Liang [10].

Table 3

Performance of sample sizes based on the GEE method under a higher proportion of missing data.

(η, ρ)	$\rho = -0.15$	$\rho = 0$	$\rho = 0.15$	$\rho = 0.3$
$q^{(1*)}$ (0.15, 0.05)	1048(0.796, 0.050)	912(0.812, 0.053)	776(0.808, 0.053)	641(0.804, 0.050)
$q^{(2*)}$	1382(0.809, 0.055)	1267(0.807, 0.045)	1152(0.808, 0.055)	1036(0.808, 0.051)
$q^{(3*)}$	1465(0.808, 0.046)	1346(0.805, 0.044)	1226(0.813, 0.050)	1106(0.810, 0.048)
$q^{(4*)}$	1408(0.800, 0.052)	1288(0.798, 0.050)	1168(0.794, 0.050)	1048(0.798, 0.051)
$q^{(1*)}$ (0.15, 0.10)	294(0.810, 0.047)	256(0.808, 0.047)	219(0.818, 0.050)	181(0.812, 0.052)
$q^{(2*)}$	388(0.811, 0.045)	356(0.813, 0.056)	324(0.811, 0.049)	292(0.810, 0.048)
$q^{(3*)}$	416(0.821, 0.048)	383(0.815, 0.054)	350(0.826, 0.048)	317(0.824, 0.048)
$q^{(4*)}$	389(0.802, 0.051)	356(0.794, 0.048)	323(0.805, 0.047)	289(0.789, 0.043)
$q^{(1*)}$ (0.40, 0.05)	1767(0.810, 0.049)	1536(0.804, 0.052)	1306(0.793, 0.054)	1076(0.800, 0.055)
$q^{(2*)}$	2329(0.801, 0.046)	2134(0.803, 0.049)	1938(0.797, 0.054)	1743(0.802, 0.047)
$q^{(3*)}$	2427(0.797, 0.052)	2224(0.797, 0.048)	2021(0.804, 0.049)	1818(0.804, 0.048)
$q^{(4*)}$	2414(0.806, 0.048)	2211(0.787, 0.051)	2008(0.802, 0.048)	1804(0.803, 0.043)
$q^{(1*)}$ (0.40, 0.10)	449(0.804, 0.052)	390(0.806, 0.053)	332(0.798, 0.057)	273(0.810, 0.056)
$q^{(2*)}$	592(0.799, 0.053)	542(0.807, 0.048)	492(0.808, 0.052)	443(0.805, 0.057)
$q^{(3*)}$	617(0.799, 0.056)	566(0.811, 0.050)	514(0.808, 0.053)	462(0.808, 0.049)
$q^{(4*)}$	613(0.795, 0.047)	561(0.799, 0.047)	509(0.799, 0.049)	458(0.802, 0.050)

Simulation under the same setting as that in Table 2 except that the proportion of missing data are higher, comparable to the motivating example in Section 1.