Communications for Statistical Applications and Methods
2018, Vol. 25, No. 3, 321–328

https://doi.org/10.29220/CSAM.2018.25.3.321
Print ISSN 2287-7843 / Online ISSN 2383-4757

Sample size calculation for comparing time-averaged responses in K-group repeated binary outcomes

Jijia Wang[a], Song Zhang[b], Chul Ahn[1,b]

[a]Department of Statistical Science, Southern Methodist University, USA;
[b]Department of Clinical Sciences, UT Southwestern Medical Center, USA

Abstract

In clinical trials with repeated measurements, the time-averaged difference (TAD) may provide a more powerful evaluation of treatment efficacy than the rate of changes over time when the treatment effect has rapid onset and repeated measurements continue across an extended period after a maximum effect is achieved (Overall and Doyle, *Controlled Clinical Trials*, **15**, 100–123, 1994). The sample size formula has been investigated by many researchers for the evaluation of TAD in two treatment groups. For the evaluation of TAD in multi-arm trials, Zhang and Ahn (*Computational Statistics & Data Analysis*, **58**, 283–291, 2013) and Lou *et al.* (*Communications in Statistics-Theory and Methods*, **46**, 11204–11213, 2017b) developed the sample size formulas for continuous outcomes and count outcomes, respectively. In this paper, we derive a sample size formula to evaluate the TAD of the repeated binary outcomes in multi-arm trials using the generalized estimating equation approach. This proposed sample size formula accounts for various correlation structures and missing patterns (including a mixture of independent missing and monotone missing patterns) that are frequently encountered by practitioners in clinical trials. We conduct simulation studies to assess the performance of the proposed sample size formula under a wide range of design parameters. The results show that the empirical powers and the empirical Type I errors are close to nominal levels. We illustrate our proposed method using a clinical trial example.

Keywords: time-averaged difference, multi-arm trials, sample size formula

1. Introduction

In clinical trials with repeated measurements, comparing treatments based on the time-averaged difference (TAD), defined as the difference in the average of longitudinally measured responses between treatment groups, is often considered a meaningful metric for the treatment effect. Overall and Doyle (1994) suggested that TAD can provide a more powerful evaluation of treatment efficacy than the rate of changes over time when the treatment effect has rapid onset and repeated measurements are obtained across an extended period after the maximum effect has been achieved. Many sample size formulas have been developed for the inference of TAD between two treatment groups (Overall and Doyle, 1994; Diggle *et al.*, 2013; Zhang and Ahn, 2012; Lou *et al.*, 2017a). However, sample size calculation for the comparison of time-averaged responses among multiple treatment groups has received less attention in the literature. Randomized trials with multiple treatment arms are widely used in practice (Parmar *et al.*, 2014). They increase the chance of finding an effective treatment as well as reduce the cost and time requirement of clinical trials by testing more treatments simultaneously. For multi-arm trials with continuous outcomes, Zhang and Ahn (2013) presented the sample size formula

[1] Corresponding author: Department of Clinical Sciences, UT Southwestern Medical Center, 5323 Harry Hines Blvd, E5.506, Dallas, TX 75390, USA. E-mail: Chul.Ahn@UTSouthwestern.edu

Published 31 May 2018 / journal homepage: http://csam.or.kr

to compare the time-averaged responses based on the generalized estimating equation (GEE) method (Liang and Zeger, 1986). This sample size formula took into account arbitrary missing patterns, various correlation structures, and unbalanced randomization. Lou *et al.* (2017b) further extended the sample size approach to multi-arm trials with repeatedly measured count outcomes.

In this paper, we investigate sample size calculation for the comparison of time-averaged responses among $K \geq 3$ groups where a binary outcome is repeatedly measured over the study period. The paper is arranged as follows. In Section 2, we briefly review the GEE method for the inference of TAD in multi-arm trials with a repeatedly measured binary outcome. In Section 3, we derive a closed-form sample size formula for the assessment of TAD among treatment groups over time using the GEE method. We demonstrate that this formula is flexible to account for various missing patterns and correlation structures. In Section 4, we conduct simulation studies to assess the performance of the proposed sample size formula under various practical settings. In Section 5, we illustrate the proposed method using a clinical trial example. Section 6 concludes with the discussion.

2. Generalized estimating equation estimator

Suppose in a clinical trial a total of n subjects are enrolled and randomly assigned to one of K treatment groups. Each subject is scheduled to obtain J measurements over the study period. Let Y_{kij} be the binary response measurement obtained at time $t_j (j = 1, \ldots, J)$ from subject $i (i = 1, \ldots, n_k)$ of the kth treatment group ($k = 1, \ldots, K$), where n_k denotes the number of subjects assigned to the kth treatment group. We use $r_k = n_k/n$ to denote the proportion of subjects assigned to the kth treatment group. To make inference about the TAD among the K treatment groups, we model Y_{kij} with the following logistic model:

$$\begin{aligned} Y_{kij} &\sim \text{Bernoulli}(p_k), \\ \text{logit}(p_k) = \log\left(\frac{p_k}{1-p_k}\right) &= b_k, \quad \text{for } k = 1, \ldots, K. \end{aligned}$$

We have $E(Y_{kij}) = p_k = e^{b_k}/(1 + e^{b_k})$. That is, $\boldsymbol{b} = (b_1, \ldots, b_K)'$ represents the time-averaged response on the log-odds scale for the K treatment groups. As a binary variable, it is obvious that $\text{Var}(Y_{kij}) = p_k(1-p_k)$. Furthermore, we model the within-subject correlation among measurements obtained from the same subject by $\text{corr}(Y_{kij}, Y_{kij'}) = \rho_{jj'}$ with $\rho_{jj} = 1$. We assume Y_{kij}'s to be independent across subjects.

The GEE estimator of $\boldsymbol{b}$, denoted by $\hat{\boldsymbol{b}} = (\hat{b}_1, \ldots, \hat{b}_K)'$, can be obtained from the following equation:

$$\hat{b}_k = \log\left(\frac{\sum_{i=1}^{n_k}\sum_{j=1}^{J} y_{kij}/(n_k J)}{1 - \sum_{i=1}^{n_k}\sum_{j=1}^{J} y_{kij}/(n_k J)}\right),$$

which is derived based on an independent working correlation structure. Liang and Zeger (1986) showed that as $n \to \infty$, $\sqrt{n}(\hat{\boldsymbol{b}} - \boldsymbol{b})$ approximately has a normal distribution with mean vector $\boldsymbol{0}$ and variance matrix $\boldsymbol{V}$. We can consistently estimate $\boldsymbol{V}$ by $\boldsymbol{V}_n = \boldsymbol{W}\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{b}})\boldsymbol{\Sigma}_n\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{b}})\boldsymbol{W}$, where $\boldsymbol{W}$ is a

diagonal matrix with diagonal elements being $(1/\sqrt{r_1}, \ldots, 1/\sqrt{r_K})$,

$$\boldsymbol{A}_n(\boldsymbol{b}) = \begin{pmatrix} \frac{1}{n_1}\sum_{i=1}^{n_1}\sum_{j=1}^{J}\frac{e^{b_1}}{(1+e^{b_1})^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2}\sum_{i=1}^{n_2}\sum_{j=1}^{J}\frac{e^{b_2}}{(1+e^{b_2})^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{n_K}\sum_{i=1}^{n_K}\sum_{j=1}^{J}\frac{e^{b_K}}{(1+e^{b_K})^2} \end{pmatrix},$$

and

$$\boldsymbol{\Sigma}_n = \begin{pmatrix} \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\sum_{j=1}^{J}\hat{\varepsilon}_{1ij}\right)^2 & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2}\sum_{i=1}^{n_2}\left(\sum_{j=1}^{J}\hat{\varepsilon}_{2ij}\right)^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{n_K}\sum_{i=1}^{n_K}\left(\sum_{j=1}^{J}\hat{\varepsilon}_{Kij}\right)^2 \end{pmatrix}.$$

Here $\hat{\varepsilon}_{kij} = y_{kij} - e^{\hat{b}_k}/(1 - e^{\hat{b}_k)}$ denotes the residual.

To compare the time-averaged responses among K treatment groups, the null hypotheses of interest is $H_0 : b_1 = \cdots = b_K$. The test statistic is

$$Z = \frac{\boldsymbol{C}'\hat{\boldsymbol{b}}}{\sqrt{\mathrm{Var}\left(\boldsymbol{C}'\hat{\boldsymbol{b}}\right)}}, \tag{2.1}$$

where $\boldsymbol{C} = (c_1, \ldots, c_K)'$ is a vector denoting a contrast of treatment effects with $\sum_{k=1}^{K} c_k = 0$. For example, one reasonable specification would be $\boldsymbol{C} = (-1, 1/(K-1), \ldots, 1/(K-1))'$. The null hypothesis is rejected if $|Z| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution.

3. Sample size calculation

As $n \to \infty$, let $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ denote the limits of the $\boldsymbol{A}_n$ and $\boldsymbol{\Sigma}_n$, respectively. Then $\boldsymbol{V}_n$ converges to $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{A}^{-1}(\hat{\boldsymbol{b}})\boldsymbol{\Sigma}\boldsymbol{A}^{-1}(\hat{\boldsymbol{b}})\boldsymbol{W}$. Given true treatment effects $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, the required sample size to reject the null hypothesis with power $1-\gamma$ at significance level α is

$$n = \frac{\left(z_{1-\frac{\alpha}{2}} + z_{1-\gamma}\right)^2 \boldsymbol{C}'\boldsymbol{V}\boldsymbol{C}}{\left(\boldsymbol{C}'\boldsymbol{\theta}\right)^2}. \tag{3.1}$$

Missing data are frequently encountered in clinical trials with repeated measurements. We now show that a closed-form extension of (3.1) can be obtained to account for missing data. Let Δ_{kij} be the missing indicator, which takes value 0/1 for a missed/observed outcome measurement. Then the

generalized expressions of $\boldsymbol{A}_n$ and $\boldsymbol{\Sigma}_n$ that accommodate missing data are

$$\boldsymbol{A}_n(\boldsymbol{b}) = \begin{pmatrix} \frac{1}{n_1}\sum_{i=1}^{n_1}\sum_{j=1}^{J}\frac{\Delta_{1ij}e^{b_1}}{\left(1+e^{b_1}\right)^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2}\sum_{i=1}^{n_2}\sum_{j=1}^{J}\frac{\Delta_{2ij}e^{b_2}}{\left(1+e^{b_2}\right)^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{n_K}\sum_{i=1}^{n_K}\sum_{j=1}^{J}\frac{\Delta_{Kij}e^{b_K}}{\left(1+e^{b_K}\right)^2} \end{pmatrix},$$

and

$$\boldsymbol{\Sigma}_n = \begin{pmatrix} \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\sum_{j=1}^{J}\Delta_{1ij}\hat{\varepsilon}_{1ij}\right)^2 & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2}\sum_{i=1}^{n_2}\left(\sum_{j=1}^{J}\Delta_{2ij}\hat{\varepsilon}_{2ij}\right)^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{n_K}\sum_{i=1}^{n_K}\left(\sum_{j=1}^{J}\Delta_{Kij}\hat{\varepsilon}_{Kij}\right)^2 \end{pmatrix}.$$

We assume that the missing probabilities only depend on time. In addition, we define $\delta_j = E(\Delta_{kij})$ to be the probability of a subject having a measurement at time t_j and $\delta_{jj'} = E(\Delta_{kij}\Delta_{kij'})$ to be the probability of a subject having measurements simultaneously at t_j and $t_{j'}$. Note that $\delta_{jj} = \delta_j$. We can use probabilities δ_j and $\delta_{jj'}$ to describe various types of missing patterns. Then, as $n \to \infty$, we have

$$\boldsymbol{A}(\boldsymbol{b}) = \sum_{j=1}^{J}\delta_j \begin{pmatrix} \frac{e^{b_1}}{\left(1+e^{b_1}\right)^2} & 0 & \cdots & 0 \\ 0 & \frac{e^{b_2}}{\left(1+e^{b_2}\right)^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{e^{b_K}}{\left(1+e^{b_K}\right)^2} \end{pmatrix}$$

and

$$\boldsymbol{\Sigma} = \sum_{j=1}^{J}\sum_{j'=1}^{J}\delta_{jj'}\rho_{jj'} \begin{pmatrix} \frac{e^{b_1}}{\left(1+e^{b_1}\right)^2} & 0 & \cdots & 0 \\ 0 & \frac{e^{b_2}}{\left(1+e^{b_2}\right)^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{e^{b_K}}{\left(1+e^{b_K}\right)^2} \end{pmatrix}.$$

The general variance $\boldsymbol{V}$ can be expressed as

$$\boldsymbol{V} = \boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{W} = \frac{\sum_{j=1}^{J}\sum_{j'=1}^{J}\delta_{jj'}\rho_{jj'}}{\left(\sum_{j=1}^{J}\delta_j\right)^2} \begin{pmatrix} \frac{\left(1+e^{b_1}\right)^2}{r_1e^{b_1}} & 0 & \cdots & 0 \\ 0 & \frac{\left(1+e^{b_2}\right)^2}{r_2e^{b_2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{\left(1+e^{b_K}\right)^2}{r_Ke^{b_K}} \end{pmatrix}.$$

Plugging $\boldsymbol{V}$ and $\boldsymbol{C} = (-1, 1/(K-1), \ldots, 1(K-1))'$ into Equation (3.1), the generalized sample size formula that accommodates various correlation structures and missing data patterns is

$$n = \frac{\left(z_{1-\frac{\alpha}{2}} + z_{1-\gamma}\right)^2 \sum_{j=1}^{J} \sum_{j'=1}^{J} \delta_{jj'} \rho_{jj'}}{\left(\theta_1 - \frac{1}{K-1} \sum_{k=2}^{K} \theta_k\right)^2 \left(\sum_{j=1}^{J} \delta_j\right)^2} \left[\frac{\left(1 + e^{b_1}\right)^2}{r_1 e^{b_1}} + \left(\frac{1}{K-1}\right)^2 \sum_{k=2}^{K} \frac{\left(1 + e^{b_k}\right)^2}{r_k e^{b_k}} \right]. \tag{3.2}$$

4. Simulation studies

To assess the performance of the proposed sample size method, we conduct simulation under different parameter settings. Suppose subjects are randomly assigned to one of four treatment groups ($K = 4$). Each subject is planned to obtain $T = 6$ measurements at scheduled times ($t_j = j - 1$ for $j = 1, \ldots, 6$). We investigate three missing patterns: independent missing (IM), monotone missing (MM), and mixed missing (MIX). Under IM, missing measurements occur independently over time with $\delta_{jj'} = \delta_j \delta_{j'}$ for $j \neq j'$. Under MM, a subject missing a measurement at time t_j will miss all the following measurements, such that $\delta_{jj'} = \delta_{j'}$ for $j \leq j'$. IM represents scenarios where subjects miss clinical visits due to random reasons. MM represents scenarios where subjects randomly drop out during the study period. In real clinical trials, it is likely that subject dropout and missing visits both occur, which implies a mixture of IM and MM, denoted by MIX. Let $(\delta_1^{(\mathrm{IM})}, \ldots, \delta_J^{(\mathrm{IM})})$ and $(\delta_1^{(\mathrm{MM})}, \ldots, \delta_J^{(\mathrm{MM})})$ be the marginal probabilities of obtaining a measurement under the IM and MM patterns, respectively. Moreover, let $\delta_{jj'}^{(\mathrm{IM})}$ and $\delta_{jj'}^{(\mathrm{MM})}$ be the corresponding joint probabilities under these two patterns. Suppose the proportion of subjects who follow IM and MM patterns are w and $1 - w$, then for mixed missing, we have

$$\begin{aligned} \delta_j^{(\mathrm{MIX})} &= w\delta_j^{(\mathrm{IM})} + (1-w)\,\delta_j^{(\mathrm{MM})}, \\ \delta_{jj'}^{(\mathrm{MIX})} &= w\delta_{jj'}^{(\mathrm{IM})} + (1-w)\,\delta_{jj'}^{(\mathrm{MM})}. \end{aligned}$$

In simulation studies, we set $w = 0.5$ for the MIX pattern. For marginal observation probabilities, we assume $\boldsymbol{\delta}^{(IM)} = \boldsymbol{\delta}^{(MM)} = \boldsymbol{\delta} = (\delta_1, \ldots, \delta_J)$, where four sets of values are explored:

$$\begin{aligned} \boldsymbol{\delta}_1 &= (1.00, 1.00, 1.00, 1.00, 1.00, 1.00)\,, \\ \boldsymbol{\delta}_2 &= (1.00, 0.95, 0.90, 0.85, 0.80, 0.75)\,, \\ \boldsymbol{\delta}_3 &= (1.00, 0.99, 0.96, 0.91, 0.84, 0.75)\,, \\ \boldsymbol{\delta}_4 &= (1.00, 0.91, 0.84, 0.79, 0.76, 0.75)\,. \end{aligned}$$

These settings correspond to four scenarios. We assume no missing data at t_1. $\boldsymbol{\delta}_1$ indicates complete data throughout the study. $\boldsymbol{\delta}_2$ indicates that the missing probabilities have a linear trend with 5% increase in dropout rate at each subsequent time point. $\boldsymbol{\delta}_3$ indicates an accelerating trend missing probabilities toward the end of study. $\boldsymbol{\delta}_4$ indicates a trend opposite to that of $\boldsymbol{\delta}_3$. Note that $\boldsymbol{\delta}_2 - \boldsymbol{\delta}_4$ have the same dropout rate (25%) at the end of study.

The simulation study also explores two correlation structures: compound symmetry (CS, $\rho_{jj'} = \rho$ for $j \neq j'$) and auto-regressive (AR(1), $\rho_{jj'} = \rho^{|t_j - t_{j'}|}$) with $\rho = 0.3, 0.5$. We set the Type I error and power at $\alpha = 0.05$ and $1 - \gamma = 0.8$, respectively. For simplicity, we assume the balanced design with $r_1 = \cdots = r_K = 1/K$. However, the proposed sample size formula is applicable to any unbalance design. We consider two alternative hypotheses. The first alternative hypothesis is that the first group

Table 1: Required sample size (empirical power, empirical Type I error) under $H_1 : \theta_1 = 0, \theta_2 = \theta_3 = \theta_4 = 0.5$

	$\boldsymbol{\delta}$	CS		AR(1)	
		$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.3$	$\rho = 0.5$
	$\boldsymbol{\delta}_1$	284 (0.8074, 0.0508)	397 (0.7996, 0.0546)	188 (0.8083, 0.0550)	266 (0.8032, 0.0544)
IM	$\boldsymbol{\delta}_2$	300 (0.8060, 0.0546)	413 (0.8042, 0.0507)	205 (0.8047, 0.0547)	283 (0.8035, 0.0519)
	$\boldsymbol{\delta}_3$	295 (0.8043, 0.0522)	408 (0.8065, 0.0511)	201 (0.8098, 0.0573)	280 (0.7983, 0.0524)
	$\boldsymbol{\delta}_4$	305 (0.8096, 0.0526)	418 (0.8032, 0.0481)	209 (0.8056, 0.0555)	286 (0.8055, 0.0532)
MM	$\boldsymbol{\delta}_2$	312 (0.8079, 0.0555)	433 (0.8068, 0.0499)	212 (0.8081, 0.0524)	297 (0.8037, 0.0526)
	$\boldsymbol{\delta}_3$	301 (0.8082, 0.0521)	417 (0.8041, 0.0517)	205 (0.8038, 0.0518)	287 (0.8060, 0.0520)
	$\boldsymbol{\delta}_4$	323 (0.8049, 0.0488)	449 (0.8076, 0.0542)	219 (0.8095, 0.0562)	307 (0.8021, 0.0486)
MIX	$\boldsymbol{\delta}_2$	306 (0.8099, 0.0547)	423 (0.8035, 0.0551)	208 (0.8053, 0.0515)	290 (0.8043, 0.0503)
	$\boldsymbol{\delta}_3$	298 (0.8070, 0.0481)	413 (0.7985, 0.0505)	203 (0.8037, 0.0561)	283 (0.8060, 0.0535)
	$\boldsymbol{\delta}_4$	314 (0.8079, 0.0563)	433 (0.8014, 0.0534)	214 (0.8076, 0.0563)	297 (0.8094, 0.0506)

CS = compound symmetry; AR = auto-regressive; IM = independent missing; MM = monotone missing; MIX = mixed missing.

is control group. The others are treatment groups with same treatment effect. Specifically, $H_1 : \theta_1 = 0, \theta_2 = \theta_3 = \theta_4 = 0.5$. The second alternative hypothesis is that the treatment groups are ordered by treatment effects. Specifically, $H_1 : \theta_1 = 0, \theta_k = \theta_1 + (k-1)\,\Delta_0$ for $k = 2, 3, 4$ and $\Delta_0 = 0.25$. For each combination of design parameters (missing pattern, marginal observation probability $\boldsymbol{\delta}$, correlation structure, correlation ρ, alternative hypothesis H_1), the simulation is conducted as:

1. Calculate the required sample size (n) based on Equation (3.2).

2. Generate n samples under null hypothesis and alternative hypothesis separately. Each sample contains n multivariate Bernoulli random variable $\boldsymbol{Y}_{ki} = (Y_{ki1}, \ldots, Y_{kiJ})^{'}$ with correlation parameter $\rho_{jj'}$ under the assumed correlation structure. The correlated binary vectors are generated by the method of Emrich and Piedmonte (1991).

3. Create incomplete data sets. Generate missing indicators based on the specified missing pattern and the marginal observation probabilities.

4. Calculate $\hat{\boldsymbol{b}}$, $\boldsymbol{A}_n$, $\boldsymbol{\Sigma}_n$, and $\boldsymbol{W}$, and the test statistic Z based on Equation (2.1).

5. Repeat Steps 2–4 for $L = 10{,}000$ times. The empirical Type I error and empirical power are estimated by $\sum_{l=1}^{L} I\left(|Z| > z_{1-\alpha/2}\right)/L$ under the null hypothesis and alternative hypothesis, respectively.

Tables 1 and 2 summarize the required sample sizes and their corresponding empirical Type I errors and empirical powers under different combinations of simulation parameters. Table 1 is obtained under alternative hypotheses $H_1 : \theta_1 = 0, \theta_2 = \theta_3 = \theta_4 = 0.5$, and Table 2 under $H_1 : \theta_1 = 0, \theta_2 = 0.25, \theta_3 = 0.5, \theta_4 = 0.75$. We have several observations: (1) The empirical powers and Type I errors are close to the nominal levels, which are 0.8 and 0.05, respectively. Therefore, the proposed sample size formula has a good performance over a wide range of design parameter settings; (2) The sample sizes under the AR(1) are always smaller than those under the CS correlation structure, with other design parameters being the same; (3) The sample size increases with the correlation (ρ), which is obvious from the sample size formula (3.2); (4) Given the same marginal observation probabilities, the order of the required sample sizes under different missing pattern is MM > MIX > IM. Under MM missing data tend to concentrate in successive observations from a few subjects, while under IM missing data are randomly distributed across subjects over the study period. The MM missing pattern

Table 2: Required sample size (empirical power, empirical Type I error) under $H_1 : \theta_1 = 0, \theta_2 = 0.25, \theta_3 = 0.5, \theta_4 = 0.75$

	δ	CS		AR(1)	
		$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.3$	$\rho = 0.5$
	δ_1	285 (0.8005, 0.0516)	399 (0.8054, 0.0501)	189 (0.8045, 0.0536)	267 (0.8050, 0.0507)
IM	δ_2	301 (0.8000, 0.0529)	414 (0.8067, 0.0504)	205 (0.8026, 0.0567)	284 (0.8087, 0.0504)
	δ_3	296 (0.8032, 0.0500)	410 (0.8088, 0.0522)	201 (0.8024, 0.0536)	281 (0.8090, 0.0549)
	δ_4	306 (0.8020, 0.0537)	419 (0.8010, 0.0497)	209 (0.8078, 0.0561)	287 (0.8046, 0.0518)
MM	δ_2	312 (0.8027, 0.0523)	434 (0.8055, 0.0530)	212 (0.8059, 0.0575)	297 (0.8002, 0.0504)
	δ_3	301 (0.8094, 0.0489)	419 (0.8091, 0.0551)	205 (0.8087, 0.0526)	288 (0.8020, 0.0518)
	δ_4	324 (0.8077, 0.0527)	450 (0.8008, 0.0500)	220 (0.8075, 0.0531)	308 (0.8033, 0.0545)
MIX	δ_2	307 (0.8035, 0.0512)	424 (0.8001, 0.0505)	209 (0.8039, 0.0529)	291 (0.8084, 0.0505)
	δ_3	299 (0.8079, 0.0495)	414 (0.8089, 0.0503)	203 (0.8058, 0.0551)	284 (0.8010, 0.0559)
	δ_4	315 (0.8083, 0.0522)	435 (0.8074, 0.0536)	215 (0.8085, 0.0525)	297 (0.8077, 0.0506)

CS = compound symmetry; AR = auto-regressive; IM = independent missing; MM = monotone missing; MIX = mixed missing.

leads to greater information loss (hence greater sample size requirement) than IM, with MIX lies in between. (5) Despite the same dropout rate at the end of study, the order of required sample sizes is $\boldsymbol{\delta}_4 > \boldsymbol{\delta}_2 > \boldsymbol{\delta}_3$, because the overall proportion of missing values is the greatest under $\boldsymbol{\delta}_4$.

5. Example

PASS sample size software manual (Pass14, 2015) illustrated the sample size estimation to test proportions in a repeated measurement design between two treatment groups. Here, we show sample size estimation to test proportions among three treatment groups. An investigator wants to design a study that compares the efficacy of a prophylactic treatment for the common cold with two active drugs and a placebo. The null hypothesis is that there is no difference in the proportion of patients who get sick among three treatment groups. Patients will be randomly assigned to one of three treatment groups with an equal probability, and followed monthly from September to April (beginning in October, hence $J = 7$) to determine the patient's disease status (present or absent). The study investigated if there is an overall difference in the proportion of patients who get sick among three treatment groups. A baseline of 60% disease rate for the common cold is estimated based on previous studies. It is expected that the disease rate will continue to be 60% in the placebo group. Suppose that a clinically meaningful difference in efficacy is 30% relative reduction in disease rate in two active medication groups, which corresponds to a disease rate of 42%. The hypothesis of interest is then $H_0 : b_1 = b_2 = b_3$ versus $H_1 : b_1 = \text{logit}(60\%) = 0.4055, b_2 = b_3 = \text{logit}(42\%) = -0.3228$. We would like to calculate the sample size that can detect the difference in treatment effect with Type I error $\alpha = 0.05$ and power $1 - \gamma = 0.8$ under a balanced design. We set the measurement times at $t_j = j - 1 (j = 1, \ldots, 7)$. The observation probabilities are assumed to follow a linear trend with a 30% dropout at the end of study, $\boldsymbol{\delta} = (1, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70)$. Under the AR(1) correlation structure with $\rho = 0.5$, the sample sizes required under the IM, MM, and MIX (assuming a balanced mixture of IM and MM) patterns are 104, 110, and 107, respectively. However, the required sample sizes under the IM, MM, and MIX are 165, 175, and 170 under the CS correlation structure.

6. Discussion

In this study, we derived a sample size formula to compare the time-averaged responses of repeated binary outcomes among K groups. This proposed sample size formula can accommodate arbitrary

correlation structures, missing patterns, marginal observation probabilities, and unbalanced experimental designs. We develop a sample size formula based on the GEE method because: (1) It has been widely used to analyze data from clinical trials with longitudinal/repeated observations; (2) It is robust to the misspecification of correlation structure (Liang and Zeger, 1986; Jung and Ahn, 2003); (3) It is flexible to accommodate missing data (Zeger *et al.*, 1988). Our simulation studies show that the empirical powers and the empirical Type I errors are very close to the nominal levels under a wide range of design parameters. When we compare the time-averaged responses among K groups, a larger correlation is always associated with a larger sample size, which is obvious in Equation (3.2) and discussed in Lou *et al.* (2017a).

References

Diggle PJ, Heagerty P, Liang KY, and Zeger SL (2013). *Analysis of Longitudinal Data* (2nd ed.), Oxford University Press, Oxford.

Emrich L and Piedmonte M (1991). A method for generating high-dimensional multivariate binary variates, *The American Statistician*, **45**, 302–304.

Jung SH and Ahn C (2003). Sample size estimation for GEE method for comparing slopes in repeated measurements data, *Statistics in Medicine*, **22**, 1305–1315.

Liang KY and Zeger SL (1986). Longitudinal data analysis for discrete and continuous outcomes using Generalized Linear Models, *Biometrika*, **84**, 3–32.

Lou Y, Cao J, Zhang S, and Ahn C (2017a). Sample size calculations for time-averaged difference of longitudinal binary outcomes, *Communications in Statistics-Theory and Methods*, **46**, 344–353.

Lou Y, Cao J, and Ahn C (2017b). Sample size estimation for comparing rates of change in K-group repeated count outcomes, *Communications in Statistics-Theory and Methods*, **46**, 11204–11213.

Parmar M, Carpenter J, and Sydes MR (2014). More multiarm randomised trials of superiority are needed, *The Lancet*, **384**, 283–284.

PASS14 (2015). *Power Analysis and Sample Size Software*, NCSS LLC.

Overall J and Doyle S (1994). Estimating sample sizes for repeated measurement design, *Controlled Clinical Trials*, **15**, 100–123.

Zhang S and Ahn C (2012). Sample size calculations for the time-averaged differences in the presence of missing data, *Contemporary Clinical Trials*, **33**, 550–556.

Zeger SL, Liang KY, and Albert PS (1988). Models for longitudinal data: a generalized estimating equation approach, *Biometrics*, **44**, 1049–1060.

Zhang S and Ahn C (2013). Sample size calculation for comparing time-averaged responses in k-group repeated-measurement studies, *Computational Statistics & Data Analysis*, **58**, 283–291.

Received March 29, 2018; Revised April 16, 2018; Accepted April 16, 2018