Published in final edited form as:

Stat Methods Med Res. 2017 December ; 26(6): 2543–2551. doi:10.1177/0962280215601137.

Sample Size Considerations for Split-Mouth Design

Hong Zhu, Song Zhang, and Chul Ahn

Division of Biostatistics, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, USA

Abstract

Split-mouth designs are frequently used in dental clinical research, where a mouth is divided into two or more experimental segments that are randomly assigned to different treatments. It has the distinct advantage of removing a lot of inter-subject variability from the estimated treatment effect. Methods of statistical analyses for split-mouth design have been well developed. However, little work is available on sample size consideration at the design phase of a split-mouth trial, although many researchers pointed out that the split-mouth design can only be more efficient than a parallel-group design when within-subject correlation coefficient is substantial. In this paper, we propose to use the generalized estimating equation (GEE) approach to assess treatment effect in split-mouth trials, accounting for correlations among observations. Closed-form sample size formulas are introduced for the split-mouth design with continuous and binary outcomes, assuming exchangeable and “nested exchangeable” correlation structures for outcomes from the same subject. The statistical inference is based on the large sample approximation under the GEE approach. Simulation studies are conducted to investigate the finite-sample performance of the GEE sample size formulas. A dental clinical trial example is presented for illustration.

Keywords

continuous and binary outcomes; dental clinical trial; GEE; sample size; split-mouth

1. Introduction

In dental clinical trials, clinicians have the option of randomizing treatments over individuals (mouth level) or over segments in the mouth (segment level) within each individual. In the former case, all segments/sites of an individual receive the same treatment, which is called the parallel-group design. In the latter case, the split-mouth design is adopted, where the mouth is divided into two or more experimental segments that are randomly assigned to different treatments. For example, Morrow et al.¹ enrolled 23 patients in a split-mouth trial for the treatment of gingivitis. Four sites located on the left or right side (segment) of a patient’s mouth were randomly assigned to either the experimental treatment (chlorhexidine) or control (saline). The split-mouth design has the advantage of removing a lot of inter-subject variability from the estimated treatment effect, and potentially requires fewer subjects than a parallel-group trial with the same power. Statistical methods for the analysis

of quantitative outcomes arising from the split-mouth design have been developed, as reviewed in Lesaffre et al.² Mixed-effect ANOVA models, generalized mixed-effect models, and generalized estimating equation (GEE) technique³ can be used to test or estimate the treatment effect, adjusting for the clustering of data. Particularly, Donner⁴ proposed an adjusted chi-square test for analyzing binary data in a split-mouth design, and Donner⁵ discussed the problem using GEE method to further correct for imbalance in baseline measurements. Split-mouth trials are not recommended when contamination between sites is suspected and when finding matching sites is impossible in subjects.⁶ In addition, some researchers pointed out that the split-mouth design can only be more efficient than the parallel-group design when the within-patient correlation coefficient is substantial, where efficiency is determined in terms of the number of measurements needed.⁷ However, relatively little attention has been paid in literature to the sample size consideration for the split-mouth design and a rigorous comparison with the parallel-group design is not available. This paper intends to fill this gap by developing closed-form sample size formulas to help clinicians on effectively designing a split-mouth study.

The GEE method has been widely used to model repeated measurement data or clustered data due to its robustness against mis-specification of the true correlation structure and ability to accommodate missing data. Sample size calculation based on the GEE method has been explored by many researchers in literature. Liu and Liang⁸ developed a sample size formula based on a generalized score test. Jung and Ahn^{9,10} proposed sample size approaches to comparing rates of change for repeated continuous and binary measurements between two treatment groups. Zhang and Ahn¹¹ and Lou et al.¹² investigated sample size calculation for time-averaged differences for continuous and binary outcomes from repeated measurement studies, in presence of missing data. In this article, we derive closed-form sample size formulas based on the GEE approach, which properly account for correlations among observations in the split-mouth design, for both continuous and binary outcomes. We also explore the connection between the sample size requirement for the split-mouth design and that for the parallel-group design, and demonstrate how correlations among observations can influence their relative efficiency. There are other application areas in health sciences, such as dermatology and certain animal studies,^{13,14} which also utilizes the split-mouth design, or more accurately, the split-cluster design. Nevertheless, we adopt the terminology of the split-mouth design here, and the proposed methods can be directly applied to the split-cluster design when naturally occurring clusters such as multiple sites or organs in the same subject are assigned to different treatments.

The remainder of the article is arranged as follows. In Section 2, we present the data structure and notations related to the split-mouth design, and propose a GEE sample size approach based on a marginal linear regression model for continuous outcomes. In Section 3, a GEE sample size approach based on a marginal logistic regression model for binary outcomes from the split-mouth design is presented. Closed-form sample size formulas are derived based on the large sample approximation under the GEE approach. Section 4 describes simulation studies conducted to investigate performances of the GEE sample size formulas for continuous and binary outcomes with exchangeable and “nested exchangeable”¹⁵ correlation structures. It shows that power levels as predicted by the GEE sample size formula generally agree well with the empirical power. In Section 5, the

proposed sample size method is illustrated with a dental clinical trial example. Some recommendations at the design stage of a split-mouth trial and potential extensions of the proposed methods are discussed in Section 6.

2. Statistical model and sample size approach for continuous outcomes

Consider a split-mouth design where two treatments are involved. In the split-mouth design, each subject's mouth is divided into two segments, and segments of m_{1j} and m_{2j} sites within the j th subject are randomly assigned to receive either experimental or control treatments, where $m_{1j} + m_{2j} = m_j$ and $j = 1, \dots, n$. Let Y_{ijl} denote the continuous outcome on site l of subject j under treatment i , for $i = 1, 2$, $j = 1, \dots, n$ and $l = 1, \dots, m_{ij}$. We further assume that there is a common correlation ρ among outcomes $(Y_{ijk}, Y_{i'j'k'})$ within the same subject, where $i \neq i'$ or $j \neq j'$. Observed outcomes from different subjects are assumed to be independent. Let $r_{1j} = 1$ and $r_{2j} = 0$ denote the experimental and control treatments, respectively.

To make an inference on the differential effect of treatment on Y_{ijk} , we assume a linear regression model,

$$Y_{ijl} = \beta_1 + \beta_2 r_{ij} + \varepsilon_{ijl}, \quad (1)$$

where β_1 is the intercept, β_2 quantifies the differential effect of treatment, and ε_{ijl} denotes random error. It is common to assume that $E(\varepsilon_{ijl}) = 0$ and $\text{Var}(\varepsilon_{ijl}) = \sigma^2$. Our primary interest is to test the null hypothesis $H_0: \beta_2 = 0$, accounting for the within-subject correlation. Let $\beta = (\beta_1, \beta_2)'$. For the convenience of discussion, model (1) can be re-written as

$$Y_j = X_j \beta + \varepsilon_j,$$

where for subject j , Y_j is an $m_j \times 1$ vector of outcomes, ε_j is an $m_j \times 1$ vector of random errors, and X_j is an $m_j \times 2$ design matrix

$$X_j = \begin{pmatrix} 1_{m_{1j}} & 1_{m_{1j}} \\ 1_{m_{2j}} & 0_{m_{2j}} \end{pmatrix},$$

and 1_m is an $m \times 1$ vector of 1's and 0_m is an $m \times 1$ vector of 0's.

Under the model assumption, the true correlation matrix is given by $R_0 = (1 - \rho)I_{m_{1j} + m_{2j}} + \rho J_{m_{1j} + m_{2j}}$ (exchangeable), where I is an identity matrix and J is a square matrix of 1's. We use the independent working correlation structure to derive the GEE estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$, which is obtained by solving equation $S_n(\beta) = 0$ with

$$S_n(\beta) = n^{-1/2} \sum_j (\mathbf{Y}_j - \mathbf{X}_j \beta)^T \mathbf{X}_j.$$

Liang and Zeger³ showed that $n^{1/2}(\hat{\beta} - \beta)$ is approximately normal with mean 0 and the variance is consistently estimated by

$$\Sigma_n = \sigma^2 \left(n^{-1} \sum_j \mathbf{X}_j^T \mathbf{X}_j \right)^{-1} \left(n^{-1} \sum_j \mathbf{X}_j^T R_0 \mathbf{X}_j \right) \left(n^{-1} \sum_j \mathbf{X}_j^T \mathbf{X}_j \right)^{-1}.$$

The robust variance of $n^{1/2}\hat{\beta}_2$ is the (2,2)th element of Σ_n , denoted by σ_2^2 . We reject H_0 if

$\left| \frac{n^{1/2}\hat{\beta}_2}{\sigma_2} \right| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $100(1-\frac{\alpha}{2})$ th percentile of a standard normal distribution. Sample size formula for the split-mouth design can be developed based on this result.

For illustration, we assume a balanced design where $m_{1j} = m_{2j} = k$. By some algebra, it can be shown that, as $n \rightarrow \infty$,

$$\begin{aligned} \left(n^{-1} \sum_j \mathbf{X}_j^T \mathbf{X}_j \right)^{-1} &\rightarrow \begin{pmatrix} \frac{1}{k} & -\frac{1}{k} \\ -\frac{1}{k} & \frac{2}{k} \end{pmatrix}, \\ n^{-1} \sum_j \mathbf{X}_j^T R_0 \mathbf{X}_j &\rightarrow \begin{pmatrix} 2k \{1 + (2k-1)\rho\} & k \{1 + (2k-1)\rho\} \\ k \{1 + (2k-1)\rho\} & k \{1 + (k-1)\rho\} \end{pmatrix}. \end{aligned}$$

The (2,2)th element of Σ_n is simplified as $\sigma_2^2 = \frac{2\sigma^2(1-\rho)}{k}$. Then, given type I error α , power $1 - \gamma$, and the true value of differential treatment effect β_2 the required total number of subjects for the split-mouth design is

$$n_s = \frac{2\sigma^2(1-\rho)(z_{1-\frac{\alpha}{2}} + z_{1-\gamma})^2}{k\beta_2^2}, \quad (2)$$

where k is the number of sites per segment. Since each subject contributes $2k$ sites in a split-mouth trial, the total number of sites is $m_s = 2kn_s$. On the other hand, the required total number of subjects for the parallel-group design with a cluster size of k is¹⁶

$$n_p = \frac{\{1 + (k-1)\rho\}}{k} \times \frac{4\sigma^2(z_{1-\frac{\alpha}{2}} + z_{1-\gamma})^2}{\beta_2^2}.$$

In the parallel-group design the total number of sites is $m_p = kn_p$. Therefore, in terms of the number of subjects, the relative efficiency of the split-mouth design over the parallel-group design can be expressed as

$$\frac{n_p}{n_s} = \frac{2\{1+(k-1)\rho\}}{1-\rho},$$

from which we find that the relative efficiency of the split-mouth design increases with the within-subject correlation ρ . The relative efficiency of the split-mouth design in terms of the number of sites is

$$\frac{m_p}{m_s} = \frac{n_p}{2n_s} = \frac{1+(k-1)\rho}{1-\rho},$$

Further, for the special case of one site per segment ($k = 1$), the relative efficiency of the split-mouth design in terms of the number of subjects becomes

$$\frac{n_p}{n_s} = \frac{2}{1-\rho},$$

which was discussed in Wang and Bakhai.¹⁷

Additionally, we can employ a more general correlation structure by assuming a common correlation ρ among outcomes (Y_{ijh} , $Y_{ijl'}$) observed in the same segment, referred to as the intra-segment correlation, while the correlation among outcomes (Y_{1jh} , $Y_{2jl'}$) observed in different segments within the same subject is given by ρ_{12} , referred to as the inter-segment correlation. Consequently, both the intra-segment correlation and the inter-segment correlation need to be taken into account for testing the null hypothesis $H_0: \beta_2 = 0$. Under the new model assumption, the true correlation matrix (“nested exchangeable”) is

$$R_1 = \begin{pmatrix} (1-\rho)I_{m_{1j}} + \rho J_{m_{1j}} & \rho_{12}J_{m_{1j} \times m_{2j}} \\ \rho_{12}J_{m_{2j} \times m_{1j}} & (1-\rho)I_{m_{2j}} + \rho J_{m_{2j}} \end{pmatrix}.$$

The combinations (ρ, ρ_{12}) , for which R_1 is positive definite, can be determined using the technique in Teerenstra et al.¹⁵ Similarly, the robust variance of $n^{1/2}\hat{\beta}_2$ is the (2,2)th element of

$$\sum_n^1 = \sigma^2 \left(n^{-1} \sum_j \mathbf{X}_j^T \mathbf{X}_j \right)^{-1} \left(n^{-1} \sum_j \mathbf{X}_j^T R_1 \mathbf{X}_j \right) \left(n^{-1} \sum_j \mathbf{X}_j^T \mathbf{X}_j \right)^{-1}.$$

For a balanced design where $m_{1j} = m_{2j} = k$, as $n \rightarrow \infty$,

$$n^{-1} \sum_j \mathbf{X}_j^T R_1 \mathbf{X}_j \rightarrow \begin{pmatrix} 2k \{1+(k-1)\rho+k\rho_{12}\} & k \{1+(k-1)\rho+k\rho_{12}\} \\ k \{1+(k-1)\rho+k\rho_{12}\} & k \{1+(k-1)\rho\} \end{pmatrix}$$

and the (2,2)th element of $\sum_n^1$ is $\sigma_{21}^2 = \frac{2\sigma^2 \{1+(k-1)\rho-k\rho_{12}\}}{k}$. Thus, given type I error α , power $1 - \gamma$, and the true value of differential treatment effect β_2 , the required total number of subjects for the split-mouth design is

$$n_s = \frac{2\sigma^2 \{1+(k-1)\rho-k\rho_{12}\} (z_{1-\frac{\alpha}{2}} + z_{1-\gamma})^2}{k\beta_2^2}, \quad (3)$$

where k is the number of sites per segment. When the true correlation is exchangeable ($\rho = \rho_{12}$), the sample size formula in (3) reduces to the one in (2).

3. Statistical model and sample size approach for binary outcomes

In this section, we discuss sample size consideration for the split-mouth design with binary outcomes. Let Y_{ijl} denote the binary response on site l of subject j receiving treatment i , where $Y_{ijl} = 1$ denotes a “success” and $Y_{ijl} = 0$ denotes a “failure” for $i = 1, 2, j = 1, \dots, n$ and $l = 1, \dots, m_{ij}$. We assume that the intra-segment correlation is ρ and the inter-segment correlation is ρ_{12} .

To make an inference on the differential effect of treatment, we assume the following logistic regression model: $Y_{ijl} \sim \text{Bernoulli}(p_{ijl})$ and

$$\text{logit} \{ \Pr(Y_{ijl}=1) \} = \log \left(\frac{p_{ijl}}{1-p_{ijl}} \right) = \beta_1 + \beta_2 r_{ij}, \quad (4)$$

$r_{1j} = 1$ and $r_{2j} = 0$ indicate the experimental and control treatments, respectively, β_1 is the log-transformed odds for the control group, and β_2 is the log-transformed odds ratio between the experimental and control treatments, representing the treatment difference on the outcome. The primary interest is to test the null hypothesis $H_0: \beta_2 = 0$ accounting for both the intra-segment and inter-segment correlations. We can apply the GEE using the robust variance approach and an independent working correlation matrix to test $H_0: \beta_2 = 0$. Model (4) can be re-written as

$$p_{ijl}(\beta) = \frac{e^{\beta' Z_{ijl}}}{1 + e^{\beta' Z_{ijl}}},$$

where $\beta = (\beta_1, \beta_2)'$ and $Z_{ijl} = (1, r_{ij})$. Under the model assumption, the true correlation matrix is “nested exchangeable”. By the GEE method, an estimator $\hat{\beta}$ is obtained by solving equation $S_n^*(\beta) = 0$ with

$$S_n^*(\beta) = n^{-1/2} \sum_{j=1}^n \sum_{i=1}^2 \sum_{l=1}^{m_{ij}} \{Y_{ijl} - p_{ijl}(\beta)\} Z_{ijl}.$$

The equation is solved by the Newton-Raphson algorithm. At the m th iteration,

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + n^{-1/2} A_n^{-1} \left(\hat{\beta}^{(m-1)} \right) S_n^* \left(\hat{\beta}^{(m-1)} \right),$$

where

$$A_n(\beta) = -n^{-1/2} \frac{\partial S_n^*(\beta)}{\partial \beta} = n^{-1} \sum_{j=1}^n \sum_{i=1}^2 \sum_{l=1}^{m_{ij}} p_{ijl} (1 - p_{ijl}) \begin{pmatrix} 1 & r_{ij} \\ r_{ij} & r_{ij}^2 \end{pmatrix}.$$

Liang and Zeger³ showed that $n^{1/2}(\hat{\beta} - \beta)$ is approximately normal with mean 0 and the variance is consistently estimated by $\sum_n^* = A_n^{-1}(\hat{\beta}) V_n(\hat{\beta}) A_n^{-1}(\hat{\beta})$, where

$$V_n(\hat{\beta}) = n^{-1} \sum_{j=1}^n \left(\sum_{i=1}^2 \sum_{l=1}^{m_{ij}} \hat{\epsilon}_{ijl} Z_{ijl} \right)^{\otimes 2},$$

with $\hat{\epsilon}_{ijl} = Y_{ijl} - p_{ijl}(\hat{\beta})$, and $c^{\otimes 2} = cc^T$ for a vector c . We reject H_0 if $\left| \frac{n^{1/2} \hat{\beta}_2}{\sigma_{2*}} \right| > z_{1-\alpha/2}$, where σ_{2*}^2 is the (2,2)th element of $\sum_n^*$.

We consider a balanced design where $m_{1j} = m_{2j} = k$. To facilitate the discussion, let $P_1 = \theta^{\beta_1 + \beta_2} / (1 + \theta^{\beta_1 + \beta_2})$ and $P_2 = \theta^{\beta_1} / (1 + \theta^{\beta_1})$ denote the true success rates in the experimental and control groups, respectively. Define $Q_1 = 1 - P_1$ and $Q_2 = 1 - P_2$. The null hypothesis $H_0: \beta_2 = 0$ is equivalent to $H_0: P_1 = P_2$.

Theorem 1

As $n \rightarrow \infty$, $\sum_n^* = A_n^{-1}(\hat{\beta}) V_n(\hat{\beta}) A_n^{-1}(\hat{\beta}) \rightarrow \sum^*$ and the (2,2)th element of $\sum^*$ has a closed form

$$\sigma_{2*}^2 = \frac{\{1+(k-1)\rho\}(P_1Q_1+P_2Q_2)-2k\rho_{12}\sqrt{P_1Q_1P_2Q_2}}{kP_1Q_1P_2Q_2}.$$

The proof of Theorem 1 is presented in the Appendix. Given type I error α , power $1 - \gamma$, and the true value of differential treatment effect β_2 the required total number of subjects for the split-mouth design is

$$n_s = \frac{\sigma_{2*}^2 (z_{1-\frac{\alpha}{2}} + z_{1-\gamma})^2}{\beta_2^2}, \quad (5)$$

When the true correlation is exchangeable ($\rho = \rho_{12}$) and $P_1 = P_2 = P$ (P is the true overall success rate), the sample size formula (5) can be simplified as

$$n_s = \frac{2(1-\rho)(z_{1-\frac{\alpha}{2}} + z_{1-\gamma})^2}{kP(1-P)\beta_2^2},$$

where k is the number of sites per segment. For the parallel-group design with a cluster size of k for binary outcomes, the required total number of subjects is

$$n_p = \frac{\{1+(k-1)\rho\}}{k} \times \frac{4(z_{1-\frac{\alpha}{2}} + z_{1-\gamma})^2}{P(1-P)\beta_2^2}.$$

Therefore, with the exchangeable correlation structure, we have

$$\frac{n_p}{n_s} = \frac{2\{1+(k-1)\rho\}}{1-\rho},$$

which is consistent with the relative efficiency of the split-mouth design over the parallel-group design that we find for the continuous outcomes, in terms of the number of subjects.

4. Simulation study

The first set of the simulation is to demonstrate the effect of various design configurations on the sample size for the split-mouth design with continuous outcomes. The nominal levels of type I error and power are set at $\alpha = 0.05$ and $1 - \gamma = 0.8$, respectively. We consider both exchangeable ($\rho = \rho_{12}$) and “nested exchangeable” ($\rho \neq \rho_{12}$) correlation structures, where values of ρ range from 0.1, 0.15 to 0.2, and values of ρ_{12} range from 0.05, 0.1 to 0.15. Assuming a balanced design with $m_{1j} = m_{2j} = k = 3$, three sites in each segment are assigned to either experimental or control treatment. We set the true values of regression coefficients

$\beta = (\beta_1, \beta_2)' = (0.3, 0.2)'$ and variance $\sigma^2 = 0.5$ or 1, where $(\beta_2, \sigma^2) = (0.2, 1)$ indicates an effect size of 0.2 comparing experimental treatment with control treatment. We assess the performance of the proposed sample size approach for continuous outcomes for each combination of $(\sigma^2, \rho, \rho_{12})$. The simulation procedure is as follows: (i) Estimate sample size n_s based on equation (3); (ii) Generate 5000 null ($\beta_2 = 0$) data sets and 5000 alternative ($\beta_2 = 0.2$) data sets, each containing n_s subjects. For subject j , generate a vector of outcomes Y_j from the Model $Y_j = X_j\beta + \varepsilon_j$ where a vector of random errors ε_j is generated from a multivariate normal distribution with mean 0, variance $\sigma^2 = 0.5$ or 1, and correlation matrix R_0 (exchangeable) or R_1 ("nested exchangeable"); (iii) For each data set, estimate $\hat{\beta}_2$ and σ_2 (or σ_{21}); (iv) Calculate the empirical type I error and empirical power as the proportion of

$$\left| \frac{n_s^{1/2} \hat{\beta}_2}{\sigma_2} \right| > z_{1-0.05/2} \text{ (or } \left| \frac{n_s^{1/2} \hat{\beta}_2}{\sigma_{21}} \right| > z_{1-0.05/2}) \text{ under the null and alternative hypotheses.}$$

Table 1 presents the sample size estimate, empirical power and empirical type I error from the simulation. The empirical powers and type I errors are generally close to their nominal levels, which indicates a good performance of the proposed method. With all other factors fixed, for the "nested exchangeable" correlation ($\rho \neq \rho_{12}$), the sample size increases as the intra-segment correlation ρ increases, or the inter-segment correlation ρ_{12} decreases. For the exchangeable correlation ($\rho = \rho_{12}$), the sample size increases as ρ decreases. The statistical inference under the GEE method is based on a large sample approximation. It is thus important to examine the performance of the proposed method in some small-sample-size scenarios. In Table 1, we have explored scenarios where sample size can be as small as 49 and the corresponding empirical power remains close to the nominal level of 0.8. This provides assurance to researchers that the proposed method is widely applicable to split-mouth clinical trials, even when the sample size is relatively small.

The second set of the simulation is to evaluate the performance of the GEE sample size formula for binary outcomes, under various design configurations: true success rate for control P_2 of 0.1 or 0.2; true treatment effect $= P_1 - P_2$ of 0.05 or 0.1; true intra-segment correlation ρ from 0.1, 0.15 to 0.2; true inter-segment correlation ρ_{12} from 0.05, 0.1 to 0.15. We assume a balanced design with $m_{1j} = m_{2j} = k = 3$. For each $(P_1, P_2, \rho, \rho_{12})$ we generate 5000 null data sets and 5000 alternative data sets with sample size n_s , where the vector of binary outcomes Y_j for subject j is generated using the approach described by Obuchowski.¹⁸ The empirical power and type I error are calculated as proportions of times that the alternative hypothesis and null hypothesis are rejected based on the GEE approach, when fitting the logistic regression model (4) to the simulated data. Table 2 summarizes the simulation results, including the required sample size, empirical power and empirical type I error under different combinations of design factors $(P_1, P_2, \rho, \rho_{12})$. We have studied scenarios with a wide range of sample size, varying from 53 to 457. It shows that a larger treatment effect leads to a smaller sample size requirement. Similar to continuous outcomes, the sample size increases as ρ increases or ρ_{12} decreases, when $\rho \neq \rho_{12}$, with all other factors fixed. The empirical type I errors are all close to the nominal level of 0.05. The empirical power tends to be smaller than the nominal level when the sample size is relatively small, and gets closer to the nominal level as the sample size increases. The possible explanation

might be related to normal approximation. When sample size is small, the normal approximation might not be very satisfactory.

5. Example

Morrow et al.¹ reported a split-mouth trial of chlorhexidine in the treatment of gingivitis, where the left and right sides of a patient's upper and lower jaws were randomly assigned to chlorhexidine or a control treatment. Each treatment was applied to four sites located on the left and right sides of the upper and lower jaws ($m_{1j} = m_{2j} = k = 4$). The trial enrolled 23 orthodontic patients, and the proportions of patients having plaque in chlorhexidine and control groups are estimated as $\hat{P}_1 = 0.89$ and $\hat{P}_2 = 0.77$, respectively, with the intra-segment correlation estimated as $\hat{\rho} = 0.070$ and the inter-segment correlation estimated as $\hat{\rho}_{12} = 0.039$. Correspondingly, we have the log-transformed odds for the control group $\hat{\beta}_1 = 1.21$, and the log-transformed odds ratio $\hat{\beta}_2 = 0.88$. An investigator would like to conduct a split-mouth trial to study the effect of a new drug as a treatment of gingivitis, following a similar study design where the new drug and a control treatment are randomly assigned to each segment with four sites of a patient's mouth. To test the hypotheses $H_0: \beta_2 = 0$ versus $H_1: \beta_2$

With 80% power at a 5% significant level, the number of subjects needs to be determined. Based on the preliminary data from the aforementioned trial, we assume that the design factors are $P_2 = 0.77$, $\rho = 0.070$ and $\rho_{12} = 0.039$. By the proposed method for binary outcomes, the required total number of subjects to detect a treatment effect of $\beta_2 = P_1 - P_2 = 0.10$ ($\beta_2 = 0.69$) is $n_s = 84$ for a split-mouth trial. We also estimate the sample size for a parallel-group trial of the same cluster size of $k = 4$, and the required total number of subjects is $n_p = 135$. In terms of the number of subjects, the relative efficiency of the split-mouth design over the parallel-group design is $\frac{n_p}{n_s} = 1.61$. Further, to detect a treatment effect of $\beta_2 = P_1 - P_2 = 0.15$ ($\beta_2 = 1.23$), the required total number of subjects is $n_s = 35$ for a split-mouth trial, and the required total number of subjects for the corresponding parallel-group trial is $n_p = 48$.

6. Discussion

In this paper, we present closed-form sample size formulas for the split-mouth design with continuous and binary outcomes. To our knowledge, it is the first attempt to systematically investigate the sample size methods for designing a split-mouth study. Marginal linear and logistic regression models with the GEE approach are employed to account for correlations among observations in split-mouth trials. The main contribution of this paper is to provide closed-form sample size formulas, which allows deeper insight into the impact of various design factors (treatment effect size, intra-segment and inter-segment correlations, etc.) on the sample size. We also theoretically derive the relative efficiency of the split-mouth design over the parallel-group design. The relative efficiency increases with the inter-segment correlation ρ_{12} , while it decreases with the intra-segment correlation ρ . The statistical inference under the GEE approach is based on the asymptotic properties, thus the sample size formulas are generally applicable for large sample sizes. Our simulation shows that the nominal power and type I error are preserved over a wide range of sample sizes.

Clinical trials with multiple or repeated measurements often encounter missing data, and in a split-mouth trial, there may be missing observations at some sites of a patient. A common practice to account for missing data is to estimate sample size by $n/(1 - q)$, where n is the sample size (number of subjects or sites) calculated assuming no missing and q is the expected missing rate. However, as shown in Zhang and Ahn,¹¹ this crude adjustment may be unsatisfactory, as missing data might cause less information loss if the correlations among repeated measurements are high. On the other hand, the GEE approach has the advantage of utilizing incomplete observations. One possible extension of the current work is to incorporate missing data into sample size consideration. Finally, in this paper we have concentrated on the split-mouth design with continuous and binary outcomes. In the future, we will extend the proposed methods to the split-mouth design with categorical, count and survival outcomes. The aforementioned extensions demand significant methodological development and will be pursued in separate studies.

Acknowledgments

This work was supported in part by the Cancer Center Support Grant from the National Cancer Institute (5P30CA142543) awarded to the Harold C. Simmons Cancer Center at the University of Texas Southwestern Medical Center. The authors thank the editor and two reviewers for their constructive comments that have improved the initial version of this paper.

References

1. Morrow D, Wood DP, Speechley M. Clinical effect of subgingival chlorhexidine irrigation on gingivitis in adolescent orthodontic patients. *Am J Orthod Dentofac.* 1992; 101:408–413.
2. Lesaffre E, Philstrom B, Needleman I, et al. The design and analysis of split-mouth studies: what statisticians and clinicians should know. *Stat Med.* 2009; 28:3470–3482. [PubMed: 19536744]
3. Liang K, Zeger S. Longitudinal data analysis for discrete and continuous outcomes using generalized linear models. *Biometrika.* 1986; 84:3–32.
4. Donner A, Klar N, Zou G. Methods for the statistical analysis of binary data in split-cluster designs. *Biometrics.* 2004; 60:919–925. [PubMed: 15606412]
5. Donner A, Zou G. Methods for the statistical analysis of binary data in split-smouth designs with baseline measurements. *Stat Med.* 2007; 26:3476–3486. [PubMed: 17149800]
6. Pandis N. Sample calculation for split-mouth designs. *Am J Orthod Dentofac.* 2012; 141:818–819.
7. Hujoei P, Loesche W. Efficiency of split-mouth designs. *J Clin Periodontol.* 1990; 17:722–728. [PubMed: 2262586]
8. Liu G, Liang K. Sample size calculations for studies with correlated observations. *Biometrics.* 1997; 53:937–947. [PubMed: 9290224]
9. Jung S, Ahn C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Stat Med.* 2003; 22:1305–1315. [PubMed: 12687656]
10. Jung S, Ahn C. Sample size for a two-group comparison of repeated binary measurements using GEE. *Stat Med.* 2005; 24:2583–2596. [PubMed: 16118812]
11. Zhang S, Ahn C. Sample size calculation for time-averaged differences in the presence of missing data. *Contemp Clin Trials.* 2012; 33:550–556. [PubMed: 22553832]
12. Lou Y, Cao J, Zhang S, et al. Sample size calculations for time-averaged difference of longitudinal binary outcomes. *Commun Stat A-Theor.* in press.
13. Bigby M, Godenne A. Continuing medical education: understanding and evaluating clinical trials. *J Am Acad Dermatol.* 1986; 34:555–593.
14. Weiss RA. Comparison of endovenous radiofrequency versus 810 nm diode laser occlusion of large veins in an animal model. *Dermatol Surg.* 2002; 28:56–61. [PubMed: 11991272]

15. Teerenstra S, Lu B, Preisser JS, et al. Sample size considerations for GEE analyses of three-level cluster randomized trials. *Biometrics*. 2010; 66:1230–1237. [PubMed: 20070297]
16. Donner, A., Klar, N. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.
17. Wang, D., Bakhai, A. Clinical Trials in Practice. A Practical Guide to Design, Analysis and Reporting, chapter 10. London: Remedica; 2006.
18. Obuchowski NA. On the comparison of correlated proportions for clustered data. *Stat Med*. 1998; 17:1495–1507. [PubMed: 9695194]

Appendix : Proof of Theorem 1

We consider a balanced design where $m_{1j} = m_{2j} = k$. First of all, $A_n(\hat{\beta})$ can be split into two parts for the treatment and control as

$$\begin{aligned} A_n(\hat{\beta}) &= \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^2 \sum_{l=1}^k p_{ijl}(\hat{\beta}) (1 - p_{ijl}(\hat{\beta})) \begin{pmatrix} 1 & r_{ij} \\ r_{ij} & r_{ij}^2 \end{pmatrix} \\ &= \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^2 \sum_{l=1}^k p_{ijl}(\beta) (1 - p_{ijl}(\beta)) \begin{pmatrix} 1 & r_{ij} \\ r_{ij} & r_{ij}^2 \end{pmatrix} + o_p(1) \\ &= \frac{1}{n} \sum_{j=1}^n \sum_{l=1}^k p_{1jl}(1 - p_{1jl}) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \frac{1}{n} \sum_{j=1}^n \sum_{l=1}^k p_{2jl}(1 - p_{2jl}) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + o_p(1). \end{aligned}$$

Applying the central limit theorem, as $n \rightarrow \infty$, $A_n(\hat{\beta})$ converges to

$$A = P(1 - P) \begin{pmatrix} k(P_1 Q_1 + P_2 Q_2) & k P_1 Q_1 \\ k P_1 Q_1 & k P_1 Q_1 \end{pmatrix}.$$

Next, we separate $V_n(\hat{\beta})$ into two parts for the treatment and control as

$$\begin{aligned} V_n(\hat{\beta}) &= \frac{1}{n} \sum_{j=1}^n \left(\sum_{i=1}^2 \sum_{l=1}^k \hat{\varepsilon}_{ijl} Z_{ijl} \right)^{\otimes 2} = \frac{1}{n} \sum_{j=1}^n \left\{ \sum_{l=1}^k \begin{pmatrix} \varepsilon_{1jl} + \varepsilon_{2jl} \\ \varepsilon_{1jl} \end{pmatrix} \right\}^{\otimes 2} + o_p(1) \\ &= \frac{1}{n} \sum_{j=1}^n \sum_{l=1}^k \sum_{l'=1}^k \begin{pmatrix} \varepsilon_{1jl} \varepsilon_{1jl'} + \varepsilon_{1jl} \varepsilon_{2jl'} + \varepsilon_{2jl} \varepsilon_{2jl'} + \varepsilon_{2jl} \varepsilon_{1jl'} & \varepsilon_{1jl} \varepsilon_{1jl'} + \varepsilon_{2jl} \varepsilon_{1jl'} \\ \varepsilon_{1jl} \varepsilon_{1jl'} + \varepsilon_{1jl} \varepsilon_{2jl'} & \varepsilon_{1jl} \varepsilon_{1jl'} \end{pmatrix} + o_p(1). \end{aligned}$$

By the central limit theorem, as $n \rightarrow \infty$, $V_n(\hat{\beta})$ converges to V , where

$$\begin{aligned} V_{11} &= \{k + (k^2 - k)\rho\}(P_1 Q_1 + P_2 Q_2) + 2k^2 \rho_{12} \sqrt{P_1 Q_1 P_2 Q_2}, \\ V_{12} = V_{21} &= \{k + (k^2 - k)\rho\} P_1 Q_1 + k^2 \rho_{12} \sqrt{P_1 Q_1 P_2 Q_2}, \\ V_{22} &= \{k + (k^2 - k)\rho\} P_1 Q_1. \end{aligned}$$

A few steps of algebra show that the (2,2)th element of $\Sigma^* = A^{-1} V A^{-1}$ is

$$\sigma_{2*}^2 = \frac{\{1+(k-1)\rho\}(P_1Q_1+P_2Q_2)-2k\rho_{12}\sqrt{P_1Q_1P_2Q_2}}{kP_1Q_1P_2Q_2}.$$

Table 1

Sample size (empirical power, empirical type I error) for simulation with continuous outcomes, type I error=0.05, power=0.8.

σ^2	σ	$\sigma_{12} = 0.05$	$\sigma_{12} = 0.1$	$\sigma_{12} = 0.15$
0.5	0.1	69 (0.810, 0.058)	59 (0.795, 0.061)	49 (0.791, 0.062)
	0.15	75 (0.810, 0.047)	65 (0.795, 0.057)	56 (0.809, 0.054)
	0.2	82 (0.820, 0.057)	72 (0.805, 0.053)	62 (0.804, 0.056)
1	0.1	137 (0.799, 0.051)	118 (0.808, 0.052)	98 (0.809, 0.056)
	0.15	150 (0.798, 0.053)	131 (0.785, 0.049)	111 (0.793, 0.057)
	0.2	163 (0.798, 0.047)	144 (0.810, 0.056)	124 (0.791, 0.051)

Table 2

Sample size (empirical power, empirical type I error) for simulation with binary outcomes, type I error=0.05, power=0.8.

P_1	P_2	ρ	$\rho_{12} = 0.05$	$\rho_{12} = 0.1$	$\rho_{12} = 0.15$
0.15	0.1	0.1	244 (0.793, 0.054)	209 (0.766, 0.047)	175 (0.730, 0.049)
		0.15	267 (0.826, 0.051)	232 (0.791, 0.050)	198 (0.759, 0.049)
		0.2	290 (0.823, 0.052)	256 (0.806, 0.049)	221 (0.780, 0.056)
0.2	0.1	0.1	73 (0.836, 0.056)	63 (0.802, 0.052)	53 (0.759, 0.053)
		0.15	80 (0.845, 0.053)	70 (0.830, 0.054)	60 (0.795, 0.053)
		0.2	87 (0.862, 0.048)	77 (0.846, 0.046)	67 (0.813, 0.055)
0.25	0.2	0.1	384 (0.826, 0.052)	330 (0.805, 0.053)	275 (0.768, 0.054)
		0.15	421 (0.844, 0.054)	366 (0.823, 0.051)	311 (0.791, 0.053)
		0.2	457 (0.847, 0.050)	403 (0.834, 0.050)	348 (0.805, 0.049)
0.3	0.2	0.1	104 (0.822, 0.053)	89 (0.790, 0.051)	75 (0.762, 0.054)
		0.15	114 (0.839, 0.053)	99 (0.819, 0.051)	85 (0.787, 0.053)
		0.2	124 (0.841, 0.057)	109 (0.819, 0.051)	95 (0.801, 0.055)