Published in final edited form as:
Stat Med. 2013 May 30; 32(12): 2140–2154. doi:10.1002/sim.5678.

Sample size estimation in educational intervention trials with subgroup heterogeneity in only one arm

Denise Esserman[a,b], **Yingqi Zhao**[a], **Yiyun Tang**[c], and **Jianwen Cai**[a,*]

[a]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

[b]Department of Medicine, Division of General Medicine and Clinical Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, U.S.A.

[c]Pfizer Oncology, Clinical Leads - Oncology, La Jolla, CA

Abstract

We present closed form sample size and power formulas motivated by the study of a psycho-social intervention in which the experimental group has the intervention delivered in teaching subgroups while the control group receives usual care. This situation is different from the usual clustered randomized trial since subgroup heterogeneity only exists in one arm. We take this modification into consideration and present formulas for the situation in which we compare a continuous outcome at both a single point in time and longitudinally over time. In addition, we present the optimal combination of parameters such as the number of subgroups and number of time points for minimizing sample size and maximizing power subject to constraints such as the maximum number of measurements that can be taken (i.e. a proxy for cost).

Keywords

Sample size; heterogeneous subgroups; clinical trials; longitudinal data

1. Introduction

Often times besides treatment regimens having strong physical side effects (e.g. anemia) they also have strong psychological side effects (e.g. depression), which could lead to non-adherent medication taking behaviors. In the case of Hepatitis C treatment, as is probably true for most diseases, non-adherence to medication leads to a decreased chance of reaching a sustained virologic response (SVR), i.e. “cure” [1][2]. Therefore, a proposed method for helping to deal with the psychological side-effects is a therapy intervention where patients meet in groups, facilitated by a psychologist, over the course of treatment. They learn ways of coping with the side-effects of treatment and form a support group to help manage these side-effects with the primary goal of keeping medication taking adherence high, and thus increase their chances of SVR. To test this psycho-social intervention, subjects would be randomized to receive treatment, and therefore be clustered within teaching subgroup, or control, i.e. usual care where each individual would be their own subgroup (i.e. no clusters). Therefore, sample size calculations for this design would only need to factor in subgroup cluster effects for the intervention group.

*Correspondence to: Department of Biostatistics, CB#7420, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, U.S.A. cai@bios.unc.edu.

Standard sample size formulas for individuals in a randomized trial with a continuous outcome assume independence between subjects. It turns out that simply applying the standard methods will result in an underestimation of sample size if subgroup heterogeneity exists [3]. Donner *et al* [4] show that the standard sample size estimates should be inflated by a factor $1 + (n - 1)\rho$ to provide the same statistical power if the individual randomized studies were carried out, where ρ is the intracluster correlation coefficient (ICC) describing the relationship of the between to within cluster variance, and n is the average cluster size.

Other work with clustered randomized trials with a continuous outcome has mainly focused on the completely clustered randomized design, where both the treatment and the control arms have subgroup heterogeneity. Hoover [3] provides methods to compare a single measure between two interventions where the magnitude of the subgroup heterogeneity is allowed to vary between the arms. In the appendix of this article, Hoover provides a one-sided approach which allows for the control group to have a small (possibly no) heterogeneous effect, but also assumes the intervention will not be harmful. Heo and Leon [5] consider sample size requirements for cluster randomized trials where there are three level hierarchical data. Their model allows for reduction to two level and one level data, however, they do not discuss this reduction in only one of the arms. Liu *et al.* [6] provide power and sample size procedures for clustered repeated measurements using generalized estimating equations. Here randomization into the two arms of the study is cluster based. Teerenstra *et al.* [7] provide sample size and power formulas for 3-level cluster randomized trials and provide some guidance for number of clusters, number of subjects per cluster and number of evaluations; again, assuming clustering in both groups.

Since we are dealing with the situation where we have subgroup heterogeneity within the experimental group, but no subgroup heterogeneity within the control group, methods that assume clustering in both groups as discussed above will overestimate the needed sample size, while methods that completely ignore clustering in both groups will underestimate the needed sample size. Therefore, in Section 2 we proposed modified approaches to sample size and power calculations to accommodate the situation where subgroup heterogeneity exists in only one arm of the trial. More specifically, we discuss a modified t-test approach in Section 2.1, expanding on the methodology introduced by Hoover [3], but allowing for the fact that the intervention could possibly be harmful (two-sided test); we address the longitudinal setting in Section 2.2; and we discuss optimal allocation in Section 2.3. In Section 3, we present simulation studies comparing the empirical and estimated power and type I error rates for the tests derived in Sections 2.1 and 2.2 and present the power curves when trying to optimize resources in the longitudinal setting. In Section 4, we present an example and examine ways to maximize power given limited resources. Finally, in Section 5 we provide a brief discussion of the methods and results and give suggestions for areas of future research.

2. General Methodology

2.1. Single Measurement

Below we discuss sample size calculations for the difference in the mean responses between two arms, one which has subgroup heterogeneity and the other which does not. The primary interest is testing whether the intervention works, i.e. whether there is a difference in the means of the two arms. If we simply use the traditional two-sample t-test and ignore the clustering in the intervention arm, we utilize more information than we actually have and will therefore, overestimate the power, resulting in an insufficient sample size to reach the desired results.

Similar to the notation used by Hoover [3], we first assume $k_E > 1$ subgroups in the experimental arm with subgroup size n_i for the i^{th} subgroup, $i = 1, \ldots, k_E$. Therefore, the total sample size in the experimental arm is given by $n_E = \sum_{i=1}^{k_E} n_i$ and n_C represents the total number of subjects in the control arm. Let Y_k^C, $k = 1, \ldots, n_C$ denote the outcome for the k^{th} subject in the control arm. Assuming that for individuals in the control arm, the model can be expressed as $Y_k^C = \mu_0 + \varepsilon_k$, where μ_0 is the pre-intervention mean outcome and the ε_k ($k = 1, \ldots, n_C$) denote the errors which are independent and identically distributed normal random variables with mean 0 and variance $\sigma_{0,C}^2$, accounting for the individual heterogeneity in the control arm. Let Y_{ij}^E represent the outcome for the j^{th} subject within the i^{th} subgroup in the experimental arm, $i = 1, \ldots, k_E$, $j = 1, \ldots, n_i$. For the experimental arm, in addition to the individual heterogeneity, we need to take into account the heterogeneous treatment cluster effects. The model can be written as $Y_{ij}^E = \mu_0 + \delta + b_i + \varepsilon_{ij}$, where δ is the treatment effect due to experimental intervention (i.e. if δ is different from 0, then on average patients in the intervention arm will have responses different from that of the control arm.), the ε_{ij} ($i = 1, \ldots, k_E$, $j = 1, \ldots, n_i$) are assumed to be independently and normally distributed with mean 0 and variance $\sigma_{0,E}^2$, where the individual error may be different from that in the control arm, and b_i represents the random effect in each subgroup i, independently and normally distributed with mean 0 and variance σ_E^2, where the magnitude of the variation σ_E^2 will depend on the performances of different therapists or different group dynamics.

Hoover [3] presented several approaches to compare two arms, both with subgroup heterogeneity. We consider methods for the setting with only one arm having subgroup heterogeneity. If we are interested in detecting a clinically meaningful difference δ, we define a modified t-test, which allows for different variances in the two groups under the null. Let $\overline{Y}_{i\cdot}^E = \sum_{j=1}^{n_i} Y_{ij}^E / n_i$ denote the mean for subgroup i in the experimental arm and $\overline{Y}_{SG}^E = \sum \overline{Y}_{i\cdot}^E / k_E$ denote the sample mean of experimental arm which weights each subgroup (SG) equally. If we let $s_{E,SG}^2 = \sum_{i=1}^{k_E} (\overline{Y}_{i\cdot}^E - \overline{Y}_{SG}^E)^2 / (k_E - 1)$, then $s_{E,SG}^2 / k_E$ estimates $\left(\sigma_{0,E}^2 / \tilde{n}_E + \sigma_E^2\right) / k_E$ with $\tilde{n}_E = (\sum_{i=1}^{k_E} 1 / k_E n_i)^{-1}$, the variance of $\overline{Y}_{SG}^E$. Note that if the inverse of n_i do not vary greatly, $s_{E,SG}^2$ can be approximated with a chi-square distribution with $k_E - 1$ degrees of freedom. We let $\overline{Y}^C$ represent the sample mean of the control arm and estimate the variance of $\overline{Y}_C$, $\sigma_{0,C}^2 / n_C$, with s_C^2 / n_C, where $s_C^2 = \sum_{i=1}^{n_C} (Y_i - \overline{Y}^C)^2 / (n_C - 1)$. The modified t-test statistic is then given by the following:

$$t_{mod} = \frac{\left| \overline{Y}_{SG}^E - \overline{Y}^C \right|}{\sqrt{\frac{s_{E,SG}^2}{k_E} + \frac{s_C^2}{n_C}}}.$$

The null hypothesis ($H_0 : \delta = 0$) is rejected for values of $t_{mod} > t_r^{\alpha/2}$, where $t_r^{\alpha/2}$ denotes the $(1 - \alpha/2)^{th}$ percentile of the t-distribution with r degrees of freedom where r comes from Satterthwaite's approximation [8], given by

$$r=\frac{\left(\frac{s_{E,SG}^2}{k_E}+\frac{s_C^2}{n_C}\right)^2}{\frac{s_{E,SG}^4}{k_E^2(k_E-1)}+\frac{s_C^4}{n_C^2(n_C-1)}}.$$

Thus, a close approximation to the power of the modified t-test, 1 − β, is given by the following:

$$\beta=t_{\tilde{r},\Psi}^{-1}\left(t_{\tilde{r}}^{\alpha/2}\right), \text{ with } \tilde{r}=\frac{U_E^2\frac{k_E+1}{k_E-1}+2U_EU_C+U_C^2\frac{n_C+1}{n_C-1}}{U_E^2\frac{k_E+1}{(k_E-1)^2}+U_C^2\frac{n_C+1}{(n_C-1)^2}},$$

where $U_E=(\sigma_{0,E}^2/\tilde{n}_E+\sigma_E^2)/k_E$, $U_C=\sigma_{0,C}^2/n_C$, and the non-centrality parameter $\Psi=\delta/\sqrt{U_E+U_C}$ [9]. Since the design effect is $1+(n-1)\rho$, where n is the planned average subgroup size, the effective sample size for the nk_E subjects in the experimental arm is thus $nk_E/(1+(n-1)\rho)$, where $\rho=\sigma_E^2/(\sigma_{0,C}^2+\sigma_E^2)$. We set $n_C=nk_E/(1+(n-1)\rho)$. For $\tilde{r}\geq 120$, we can approximate the t-distribution with a standard normal distribution [3]. Hence to detect a clinically meaningful difference δ between the two group mean responses with 1 − β power at α significance level, assuming an average subgroup size of n in the experimental arm, the required minimal number of subgroups k_E is the smallest integer k_E satisfying

$$k_E \geq \frac{(\Phi^{-1}(1-\alpha/2)+\Phi^{-1}(1-\beta))^2(\sigma_{0,E}^2/n+\sigma_E^2+(1+(n-1)\rho)\sigma_{0,C}^2/n)}{\delta^2}. \quad (1)$$

If we assume that the individual effect $\sigma_{0,C}^2=\sigma_{0,E}^2=\sigma_0^2$ for simplicity, and express the difference in terms of a standardized effect size, $\Delta_\delta=\delta/\sigma_0$, k_E is then given by

$$k_E \geq \frac{(\Phi^{-1}(1-\alpha/2)+\Phi^{-1}(1-\beta))^2(1/n+\rho/(1-\rho)+(1+(n-1)\rho)/n)}{\Delta_\delta^2},$$

where Φ denotes the cumulative distribution of the standard normal. In the situation where $\tilde{r}<120$, the number of subgroups should be determined directly from (1) by adjusting k_E until the power achieves the desired level.

2.2. Longitudinal Measurements

If instead of a single time point, each subject will be measured repeatedly over a period of time, measurements from the same subject could be correlated and therefore, these correlations must be accounted for when computing sample size for a repeated measures study design. Under these circumstances, we can fit a mixed-effects linear model for the purpose of testing the difference in outcome between the experimental and control arms over time. The resulting model will have three levels of data for the experimental arm, but only two levels in the control arm, since subjects in the control arm are independent. Heo and Leon [5] provided the power and sample size formulae to detect the interaction effects between intervention and time based on maximum likelihood estimates for a perfectly balanced design (i.e. the same number of subgroups in the two arms as well as the same number of subjects per subgroup) assuming clustering in both arms. In this subsection we

will provide the formulae for sample size and power using similar methodology under our setting, i.e. only one arm with subgroup heterogeneity.

Recall that for the experimental arm, the intervention is delivered by teaching subgroups, indexed by i, $i = 1, \ldots, k_E$, with j subjects, $j = 1, \ldots, n_i$, nested within each subgroup. $n = \sum_{i=1}^{k_E} n_i / k_E$ is the planned average subgroup size. In this multi-level setting, the first level includes repeated measures on subjects, the second level includes subjects, and the third level includes therapists/groups. Subjects in the control arm are independent (i.e, no variation stems from different teaching subgroups), hence, no level three random effects exist. We abuse notation slightly by letting $i = nk_E + 1, \ldots, nk_E + n_C$ with $j \equiv 1$ index subjects in the control arm. Note that i denotes subgroups in the experimental arm while in the control arm, i actually indexes subjects, since each subject forms a subgroup. We assume subjects are observed n_T times over the course of the study at some common time points. Let l, $l = 1, \ldots, n_T$, index the repeated measurements and let Y_{ijl} be the l^{th} response of the j^{th} subject in the i^{th} teaching subgroup (experimental arm), or the l^{th} response of the i^{th} subject (control arm), and T_l represent the measurement time of Y_{ijl} (measured as time since enrollment in the study). In addition, let Trt_i be the treatment indicator with $Trt_i = 1$ for subgroup i ($i = 1, \ldots, k_E$) in the experimental arm and $Trt_i = 0$ for patient i ($i = nk_E + 1, \ldots, nk_E + n_C$) in the control arm.

The primary interest is testing whether the treatment effect varies over time (i.e. the rate of change in the outcome of the subjects in the experimental arm is different from that in the control arm). If we let η^E and η^C denote the rates of change in the experimental and control arm, respectively, we can express the null hypothesis as:

$$H_0: \gamma = \eta^E - \eta^C = 0.$$

An unbiased estimate of γ is given by $\hat{\gamma} = \hat{\eta}^E - \hat{\eta}^C$ [5], where

$$\widehat{\eta}^E = \frac{\sum_{i=1}^{k_E} \sum_{j=1}^{n_i} \sum_{l=1}^{n_T} \left(T_l - \overline{T}\right) \left(Y_{ijl} - \overline{Y}^E\right)}{\sum_{i=1}^{k_E} \sum_{j=1}^{n} \sum_{l=1}^{n_T} \left(T_l - \overline{T}\right)^2},$$

$$\widehat{\eta}^C = \frac{\sum_{i=nk_E+1}^{nk_E+n_C} \sum_{l=1}^{n_T} \left(T_l - \overline{T}\right) \left(Y_{i1l} - \overline{Y}^C\right)}{\sum_{i=nk_E+1}^{nk_E+n_C} \sum_{l=1}^{n_T} \left(T_l - \overline{T}\right)^2},$$

$\overline{T} = \sum_{l=1}^{n_T} T_l / n_T$ is the mean time point and $Var(T) = \sum_{l=1}^{n_T} (T_l - \overline{T})^2 / n_T$ is the variance of the time variable T. In planning we want an equal number of participants in each subgroup which is common in practice; we thus assume an equal subgroup size n in all formulas from this point on.

A mixed level mixed-effects linear model can be fit as follows:

$$Y_{ijl} = \beta_0 + \beta_1 Trt_i + \beta_2 T_l + \gamma Trt_i \times T_l + b_i \times Trt_i + b_{j(i)} + \varepsilon_{ijl}, \quad (2)$$

where β_0 and $\beta_0 + \beta_1$ represent the pre-intervention main effects for the control group and experimental group, respectively, β_2 represents the main effect for time, and γ is the interaction effect between intervention and time we are interested in testing. The ε_{ijl} are the error terms, normally distributed as $N(0, \sigma_0^2)$; the $b_{j(i)}$ which are assumed to follow a normal distribution with mean 0 and variance σ_2^2, represent the random effects at level two, the subject level; and the b_i, the level three random effects for subgroups, are distributed as $N(0, \sigma_E^2)$. It is assumed that the b_i and $b_{j(i)}$ are independent of each other and the ε_{ijl}.

Based on (11), it can be shown that $E(Y_{ijl}) = \beta_0 + \beta_1 + (\beta_2 + \gamma)T_l$ and $Var(Y_{ijl})=\sigma_E^2+\sigma_2^2+\sigma_0^2$ for participants in the experimental arm and $E(Y_{ijl}) = \beta_0 + \beta_2 T_l$ and $Var(Y_{ijl})=\sigma_2^2+\sigma_0^2$ for participants in the control arm. Therefore, the ICC among repeated subgroup observations for the experimental arm is $\rho_2=Corr(Y_{ijl}, Y_{ij'l'})=\sigma_E^2/(\sigma_E^2+\sigma_2^2+\sigma_0^2)$ and the correlation for observations from a given subject from the experimental arm is

$$\rho_1=Corr(Y_{ijl}, Y_{ijl'})=\frac{\sigma_E^2+\sigma_2^2}{\sigma_E^2+\sigma_2^2+\sigma_0^2}, \quad (3)$$

and for the control arm is

$$Corr(Y_{i1l}, Y_{i1l'})=\frac{\sigma_2^2}{\sigma_2^2+\sigma_0^2}.$$

Note that a more general model can be considered, which allows the random effects for subgroup and subject levels to interact with time. The correlation between observations based on the more general model can be derived in a similar way. See the Appendix for details. For practical purposes, we stay with the current model to derive the sample size formula. The variance of $\hat{\gamma}$ can therefore be written as

$$Var(\widehat{\gamma})=Var(\widehat{\eta}^E)+Var(\widehat{\eta}^C)=\frac{\sigma_0^2}{nk_E n_T Var(T)}+\frac{\sigma_0^2}{n_C n_T Var(T)}.$$

The second equation can be obtained via expansion of $Var(\hat{\eta}^E)$ and $Var(\hat{\eta}^C)$ separately, using the specific form of variance and covariance between different subjects. Interested readers can refer to Heo and Leon [5] for more details.

Based on (3), we have $\sigma_0^2=(1-\rho_1)\sigma^2$, where $\sigma^2=\sigma_E^2+\sigma_2^2+\sigma_0^2$. Therefore, given the total variance for the experimental arm σ^2, the test statistic can be constructed as:

$$D=\frac{\widehat{\gamma}}{se(\widehat{\gamma})}=\frac{\sqrt{n_T Var(T)}(\widehat{\eta}^E-\widehat{\eta}^C)}{\sigma\sqrt{(1-\rho_1)(\frac{1}{nk_E}+\frac{1}{n_C})}}.$$

.

According to the large sample theory, as the sample size increases, the test statistic D will approach a standard normal distribution under the null. Under the alternative, $(\hat{\gamma}-\gamma)/se(\hat{\gamma}) \sim N(0, 1)$. Thus the power for the test statistic D is given by

$$\Phi\left(\frac{|\Delta_\gamma|\sqrt{n_T Var(T)}}{\sqrt{(1-\rho_1)(\frac{1}{nk_E}+\frac{1}{n_C})}}-\Phi^{-1}(1-\alpha/2)\right), \quad (4)$$

where $\Delta_\gamma = \gamma/\sigma$ is the standardized effect size for the slope difference, which is the difference between the rates of change in two groups scaled by the standard deviation.

Note that the ICC for the repeated subgroup measurements in the experimental arm $\rho_2=\sigma_E^2/\sigma^2$, where σ_E^2 is the variance component among clusters. The effective sample size for nk_E subjects is $nk_E/(1 + (n - 1)\rho_2)$, and we set it equal to n_C, i.e.

$$nk_E=n_C(1+(n-1)\rho_2). \quad (5)$$

Based on Equation (4), we can obtain the required number of teaching subgroups k_E given the other parameters. For a desired statistical power $1 - \beta$ at significance level α, k_E is the smallest integer such that

$$k_E \geq \frac{(\Phi^{-1}(1-\alpha/2)+\Phi^{-1}(1-\beta))^2(1-\rho_1+(1-\rho_1)(1+(n-1)\rho_2))}{nn_T var(T)\Delta_\gamma^2}, \quad (6)$$

and n_C is the smallest integer such that $n_C = nk_E/(1 + (n - 1)\rho_2)$. On the other hand, we can also calculate n_T as the smallest integer such that

$$n_T \geq \frac{(\Phi^{-1}(1-\alpha/2)+\Phi^{-1}(1-\beta))^2(1-\rho_1+(1-\rho_1)(1+(n-1)\rho_2))}{nk_E var(T)\Delta_\gamma^2}.$$

Holding all other factors constant, the relationship between n_T and k_E is reciprocal, i.e, to achieve the same power, we can reduce the number of repeated measures n_T while increasing the number of subgroups in the experimental arm or vice versa.

2.3. Allocation of Resources

Often times in planning a study, we not only need to consider the power/sample size requirements, but also need to take into account available resources (e.g. cost). For a fixed number of subgroups k_E in the experimental arm, larger subgroup sizes and/or more measurement time points can increase power; however, the costs will also be increased. We attempt to find a combination of subgroup sizes and time points which maximize the power to detect a clinically meaningful effect when the budget is fixed and/or minimize the study costs as long as the desired power is achieved. For simplicity, we consider the situation when the number of subjects in each subgroup equals n in the experimental arm. We assume that the total number of measurements that can be taken for the entire study is n_M, where

$$n_M=(n_C+nk_E)n_T. \quad (7)$$

Given this constraint (used as a proxy for controlling cost), our goal is to maximize the power given by (4) under certain scenarios.

By plugging $nk_E = (1 + (n - 1)\rho_2)n_C$ into (7), we have the following constraints on the relationship between n, k_E and n_C

$$n_C = \frac{1}{(2+(n-1)\rho_2)} \cdot \frac{n_M}{n_T}, \quad (8)$$

$$k_E = \frac{1+(n-1)\rho_2}{2+(n-1)\rho_2} \cdot \frac{n_M}{n_T n} \quad (9)$$

If we pre-specify the subgroup size (n) and the number of planned repeated measurements (n_T), the number of subgroups in the experimental arm and the number of subjects in the control arm are determined. On the other hand, if we are willing to be more flexible in choosing the subgroup size, but need to fix the number of clusters k_E in advance, n is determined by solving (9)

$$n = \left\lfloor \frac{(\rho_2 - 2)k_E n_T + n_M \rho_2 + \sqrt{((2-\rho_2)k_E n_T - n_M \rho_2)^2 - 4k_E \rho_2 (\rho_2 - 1) n_M n_T}}{2 n_T k_E \rho_2} \right\rfloor, \quad (10)$$

for fixed k_E, n_T and n_M, where $\lfloor x \rfloor$ is the largest integer not greater than x.

Figure 1 provides a visual display of the interrelationship among k_E, n_T, n and ρ_2 imposed by (5) and (7). According to Figure 1, the total number of subjects required for the experimental group is reduced with larger k_E or n_T. From Figure 1(a), we can see that there is an increment in the control group size when increasing the number of subgroups k_E while holding n_T fixed. On the other hand, the number of subjects in the control group will decrease as more repeated measures are chosen holding k_E fixed in Figure 1(b). Considering k_E as a function of the effects of ρ_2, according to Figure 1(c), we find that more subgroups in the experimental arm will be needed as ρ_2 increases, assuming that n_M and n are fixed. In this situation, we would need to recruit more subjects in the experimental group and less to the control group. Given fixed total number of measurements (n_M) and number of repeated measures (n_T), Figure 1(d) shows that less subgroups are required when more participants are included in each subgroup and the total number of subjects will be increased in the experimental arm, while decreased in the control arm. With the constraints imposed on the combination of (n, k_E, n_C, n_T), we want to find combinations which give us better power. Plugging (8) and (9) into (4), we can obtain the power under constraints (5) and (7). If the interest is in choosing the best combination of (n_T, k_E) to achieve the most power for given ρ_1, ρ_2, n_M, Δ_γ and $Var(T)$ from several possible combinations of (n_T, k_E), we can use (10) to calculate the corresponding subgroup size for each combination of (n_T, k_E). We then calculate the corresponding power based on (4). Thus the combination that yields the best power can be chosen accordingly.

Generally, we can increase power by either increasing the number of measurement times (n_T) or the number of subgroups (k_E) for a fixed subgroup size (n) in the experimental group, yet that may not always be feasible. Practical concerns for the cost of conducting longer trials or enrolling extra subgroups and therapists must be considered. We provide detailed illustrations in Section 3.3, where under the constraint of a fixed number of total measurements, we can identify some equivalent combinations in terms of power.

3. Simulations

3.1. Modified t-test power

The simulation studies presented below were conducted to verify the power formula for the modified t-test given by (1). In addition, we were interested in comparing the modified t-

test, denoted as method I, in which subgroup heterogeneity exists only in the experimental arm, with method II, in which we ignore all subgroup heterogeneity and assume all subjects are independent, i.e. the standard t-test, and method III, in which we consider subgroup heterogeneity in both arms using the method described in [3]. To study the performances of different tests, without loss of generality, we assumed an equal subgroup size $n = 10$, and there were $k_E = 10$ subgroups in total. Two different values for ICC $\rho = 0$ and 0.2 were considered. To generate data, we first calculate n_C based on setting $n_C = n_E/(1 + (n - 1)\rho)$ for a given combination (step 1). More specifically, for each combination, we follow these steps:

1. Calculate the sample size in the control arm given ρ and n_E;
2. Calculate the variance component σ_E for given σ_0 and ρ based on $\rho=\sigma_E^2/(\sigma_0^2+\sigma_E^2)$;
3. Generate the outcome data for the control arm $Y_j^C=\mu_0+\varepsilon_j^C$, with $Y^C=(Y_1^C,\ldots,Y_{n_C}^C)$ following a $N(\mu_0,\sigma_0^2 I_{n_C})$, in the scenario of no subgroup heterogeneity.
4. Generate the outcome data for the experimental arm $Y_{ij}^E=\mu_0+\delta+b_i^E+\varepsilon_{ij}^E$, with $Y^E=(Y_1^E,\ldots,Y_{n_E}^E)$ following a $N(\mu_0+\delta,\sigma_0^2 I_{n_E}+\Sigma_E)$, where Σ_E is a block diagonal matrix with each block consisting of $\sigma_E^2 J_n^n$, and there are k_E such blocks. J_n^n is an $n \times n$ matrix with all the entries equal to 1.
5. Conduct test with method I by considering subgroup heterogeneity in the experimental arm only;
6. Conduct test with method II using two sample t—test ignoring subgroup heterogeneity in the experimental arm;
7. Conduct test with method III by assuming subgroup heterogeneity in both arms, randomly separating the control group into k_E subgroups;
8. Retain p-values, denoted by $p_{I,s}(\delta)$, $p_{II,s}(\delta)$, and $p_{III,s}(\delta)$ for the s^{th} simulated data set (for $s = 1, 2, \ldots, 5000$) for the three methods, respectively, obtained from testing the null hypothesis $\delta = 0$;
9. Obtain the empirical power or type I error $\tilde{\phi}_m$ from 5000 simulations by

$$\tilde{\phi}_m=\sum_{s=1}^{5000}\frac{\mathbf{1}\{p_{m,s}(\delta)<\alpha\}}{5000}, m=I, II, III$$

Figure 2 presents the empirical type I error and power curves for the three methods described. The type I error rate for the modified t-test is close to the nominal level in all three ICC scenarios. As can be seen in the left panel, which corresponds to an ICC of 0, there is almost no difference between the traditional t-test and modified t-test; the two power curves from traditional and modified t-test almost overlap. Indeed, in the case where $ICC = 0$, we are testing mean difference between two groups in which subjects are independent. In this scenario, the modified t-test will reduce to a standard t-test. The middle panel summarizes the results with an ICC of 0.1 and, as expected, the type I error is inflated when the subgroup heterogeneity is ignored, while the test is conservative if we assume clustering in both groups. The right panel shows the results when ICC equals to 0.2. The type I error rate is close to the nominal level when assuming subgroup heterogeneity in both arms. This is possibly due to a small sample size required in the control arm ($n_C = 36$) when $ICC = 0.2$. When the 36 subjects are divided into 10 subgroups, the cluster size is very small relative to

the number of independent subgroups. Under such situation, the test assuming subgroup heterogeneity in both arms seems to preserve the type I error well, although slightly under powered compared to the modified t-test.

Table 1 presents a comparison between the empirical power and the theoretical power calculated from (1). Different scenarios are presented below with the simulation parameters specified as: $\delta = 0, 0.25, 0.5$; $k_E = 5, 10, 20$ and an equal subgroup size $n = 10$, corresponding to an experimental group size of $nk_E = 50, 100, 200$; and six different values for ICC: $\rho = 0.2, 0.15, 0.1, 0.05, 0.01$ and 0. Without loss of generality the pre-intervention mean level μ_0 is set at 0 and the random individual effects in both arms are generated from a standard normal distribution ($\sigma_0 = 1$). In all scenarios, the theoretical power is estimated well. Note that the number of participants to enroll in the control arm varies with the value of ρ. In addition to the power, we calculated the empirical type I error for all scenarios, where the difference between means, δ, is set to 0. The error rates are well controlled at 0.05 level. For the scenarios presented, the number of subjects per therapy subgroup is fixed at $n = 10$. If the number of participants per subgroup is decreased, the power is lower for detection of a difference given the other parameters remain the same; increasing the subgroup size will lead to more powerful results (results not shown).

3.2. Longitudinal Study to Test the Treatment Effects over Time

We conducted simulation studies to verify the power formula given by (4). Assuming an equal subgroup size n, for given n_T, T_l, n and Δ_γ, we calculated the number of subgroups needed based on (6) with 80% power and 0.05 type I error. The theoretical power is then calculated based on (4) using the calculated k_E. After generating the data, we used PROC MIXED in SAS (Cary, NC) to estimate the variance components and obtain the empirical power. Specifically, we assume equally spaced common time points with $T_l = l - 1$. To test the effect size of interaction, we formulated it in terms of the standardized between-group mean difference $\Delta_\gamma T_{end}$ at the end of trial, where $T_{end} = n_T - 1$. Scenarios with $\Delta_\gamma T_{end} = \Delta_\gamma(n_T - 1) = 0.4, 0.6$ are considered. Let $\beta_0 = \beta_1 = 0$, $\beta_2 = -1$, $\sigma^2 = \sigma_0^2 + \sigma_2^2 + \sigma_E^2 = 1$. Other simulation parameters are specified as $n_T = 3, 6, 12$, $\rho_1 = 0.4, 0.5, 0.6$, $\rho_2 = 0.05$ and the subgroup size is fixed at $n = 10$. The following steps are used for the simulations:

1. Calculate $\gamma = \sigma\Delta_\gamma$ and $Var(T)$;
2. Calculate the number of subgroups k_E in the experimental arm and the sample size n_C in the control arm;
3. Calculate the variance component σ_0^2, σ_2^2 and σ_E^2, with $\sigma_E^2 = \rho_2\sigma^2$, $\sigma_2^2 = (\rho_1 - \rho_2)\sigma^2$ and $\sigma_0^2 = \sigma^2 - (\sigma_2^2 + \sigma_E^2)$;
4. Generate treatment indicators, with $Trt_i = 1$, $i = 1, \ldots, k_E$ representing subgroups in the experimental arm, $Trt_i = 0$, $i = nk_E + 1, \ldots, nk_E + n_C$ for the subjects in the control arm.
5. Generate b_i, $i = 1, \ldots, k_E$ from $N(0, \sigma_E^2)$ independently;
6. For each b_i, $i = 1, \ldots, k_E$, generate $b_{j(i)}$ following $N(0, \sigma_2^2)$ independently for $j = 1, \ldots, n$;
7. For each combination of b_i and $b_{j(i)}$, generate ε_{ijl}, $l = 1, \ldots, n_T$ from $N(0, \sigma_0^2)$ independently;
8. Generate the outcome data for the experimental arm with

$$Y_{ijl}=\beta_0+\beta_1+\beta_2 T_l+\gamma T_l+b_i+b_{j(i)}+\varepsilon_{ijl}.$$

9. Generate $b_{j(i)}$ following $N(0,\sigma_2^2)$ independently for $i = nk_E + 1, \ldots, nk_E + n_C$, $j = 1$;
10. For each $b_{j(i)}$, $i = nk_E + 1, \ldots, nk_E + n_C$, $j = 1$, generate ε_{ijl}, $l = 1, \ldots, n_T$ from $N(0,\sigma_0^2)$ independently;
11. Generate the outcome data for the control arm with

$$Y_{ijl}=\beta_0+\beta_2 T_l+b_{j(i)}+\varepsilon_{ijl},\ j=1.$$

12. Use PROC MIXED to fit a mixed level mixed-effects linear model to the data set;
13. Retain pvalues, denoted by $p_s(\gamma)$ for the s^{th} simulated data set (for $s = 1,2, \ldots, 5000$), obtained from testing the null hypothesis $\gamma = 0$;
14. Obtain the empirical power or type I error $\tilde{\phi}_m$ from 5000 simulations by

$$\tilde{\phi}_m=\sum_{s=1}^{5000}\frac{\mathbf{1}\{p_s(\gamma)<\alpha\}}{5000}.$$

Table 2 provides simulation results for different combinations of n_T, ρ_1 and $\Delta_\gamma T_{end}$. The parameters k_E and n_C are estimated based on 80% power and 0.05 type I error. The empirical type I error rate ($\tilde{\alpha}$) and the empirical ($\tilde{\phi}$) and theoretical power (ϕ) are presented. Both the empirical type I error rate and power agree well with the theoretical values.

Simulations were also conducted to investigate the effect of each parameter on the power. We see increasing power with an increase in the number of time points measured with all the other factors fixed. There is a loss of power with smaller cluster sizes n for the same ρ_1, ρ_2, k_E and n_T. In addition, increasing ρ_2 leads to a reduction in the power as the available information is reduced with higher correlation within clusters (data not shown).

3.3. Allocation of Resources

For different combinations of (n, k_E, n_C, n_T), in order to detect a clinically meaningful effect, we can find the combinations which give us a specified power, given the fixed number of total measurements constraint. We can obtain contour plots similar to Figures 3 and 4. For example, if we are interested in identifying an effcient combination of (n_T, k_E), where the subgroup size n can vary accordingly, we compute the power for a grid of values of the parameters (n_T, k_E) subject to fixed number of total measurements. Assuming that $T_{end} = 9$, we set $T_l = T_{end}(l-1)/(n_T-1)$, $l = 1, \ldots, n_T$. The contours in Figure 3 give different levels of power for $\rho_1 = 0.4$ and $\rho_2 = 0.05$ when $n_T = 3, \ldots, 10$ and $k_E = 3, \ldots, 15$. The combinations of (n_T, k_E) with corresponding subgroup size n calculated from (10) are equivalent on the same contour in terms of the power obtained. Obviously, the power for detecting a slope difference Δ_γ of 0.04 is not sufficient with total number of measure $n_M = 500$, (i.e. no combination reaches power ≥ 80%). More measurements are required to reach a sufficient power. Therefore, we can increase n_M to 1000, where the right panel on the top shows that several different combinations of (n_T, k_E) give the power greater than 0.8. Depending on the practical considerations, we can either choose $k_E = 15$ subgroups with $n_T = 5$ measurement time points, which requires a subgroup size of $n = 7$, or we can form less subgroups in the experimental arm with less follow up sessions, say $k_E = 6$ and $n_T = 4$ while increase n to 29.

Similarly, Figure 4 can be used to find the combinations of (n, k_E), where n = 5, …, 20 and k_E = 3, …, 15, which give a specified power, with n_T determined correspondingly. In this case, we see that greater power can be reached with larger subgroup size or number of subgroups by comparing contours with the same total number of measurements (n_M). To achieve 80% power with n_M = 1000, we can either choose k_E = 15 subgroups with subgroup size n = 7, which requires n_T = 5, or we can form less subgroups in the experimental arm with bigger subgroups size, say k_E = 8 and n = 13, while keeping n_T = 5.

4. An Example

The example described below was motivated by a proposed psycho-social intervention for patients receiving treatment for Hepatitis C. As part of the usual care, all patients receiving treatment are scheduled for routine check-ups in the clinic every month for the first 6 months on treatment. Those randomized to the experimental group would receive the therapy intervention to coincide with these check-ups. Thus, measurements would be obtained at baseline, and months 1, 2, 3, 4, 5, and 6, resulting in n_T=7 and T_{end} = 6. We sought to investigate an efficient and practical design for this study making assumptions about the parameters in the model. In addition to the 7 repeated measures we also explore an n_T=5, where measurements would be obtained at baseline and months 1.5, 3, 4.5 and 6. Since time points are equally spaced, we set $T_l = l - 1$, l = 1, …, n_T for n_T = 7 and $T_l = 1.5(l - 1)$, l = 1, …, n_T for n_T = 5. Qualitative research indicates that groups of 6–10 participants are ideal to maximize group participation [10]; therefore, we explored subgroups of size 6, 8, and 10. We also assumed small and medium effect sizes, $\Delta_\gamma T_{end}$ of 0.2 and 0.5, respectively, where the between-group mean difference at the end of the trial is deemed small (medium) if it is 20% (50%) of the standard deviation. We assumed values of 0.3, 0.5 and 0.7 for ρ_1 and 0.05 for ρ_2. Table 3 gives the required number of subgroups based on (6) and the corresponding total sample size needed to achieve at least 80% power with a 5% type I error rate given all combinations of the above parameters. Note that the required sample size in the control group can be calculated using (5). As can be seen when holding all other parameters constant, to achieve 80% power: increasing n decreases the required number of subgroups; increasing ρ_1 decreases the required number of subgroups; increasing n_T decreases the required number of subgroups; and increasing $\Delta_\gamma T_{end}$ decreases the required number of subgroups.

Secondly, we fixed the total number of measurements (n_M) to 500, 1000 or 2000 and calculated the power for the above scenarios. The results are presented in Table 4. More power is associated with larger ρ_1 value, however, the investigator will have little control over it. As expected, increasing n_M, which the investigator has more control over depending on the budget resources, will result in higher power.

Table 3 indicates that to achieve an equivalent power, more follow-ups (n_T) with less number of subgroups (k_E) require more total number of measurements (n_M). Similarly, there was a slight decrease in the achieved power when increasing n_T from 5 to 7 in Table 4 for fixed n_M, ρ_1, ρ_2 and n. Therefore from a budget standpoint, it might be better to have fewer follow up sessions. With smaller n_T, the investigator will need to enroll more participants, which could be easier than having to retain smaller number of participants for more measurements. Table 3 also recommends a larger subgroup size provided that the measurement time is fixed, which requires less total number of measurements, although the difference is not substantial.

If we are interested in designing a study with ρ_1 = 0.3, ρ_2 = 0.05 and n_M = 500, to detect $\Delta_\gamma T_{end}$ = 0.5 with 80% power at .05 significance level, we first calculate k_E based on (9) for the given subgroup size n and a set of options of n_T. Then we calculate the power based on

(4). We then choose the n_T with the power closest to 80% for a given n and the results are presented in Table 5. From Table 5, we can see that if the group size is 6, we need 11 groups in the experimental group with 4 visits, equally spaced at baseline and months 2, 4 and 6. If the group size is 10, we need 9 groups with 3 visits scheduled at baseline, months 3 and 6. The latter combination might be more feasible in practice.

5. Discussion

Most of the literature addressing sample size calculations in clustered randomized trials assumes that subgroup heterogeneity occurs in both arms. We present closed form sample size and power formulas for the situation in which there is subgroup heterogeneity in only one arm of the trial. We have demonstrated through simulation that our formulas estimate the theoretical power and type I error rates well for both the modified t-test and the longitudinal setting.

We have explored how to allocate resources. We present plots in which we fix the total number of measurements which can be used as a proxy for cost. With these plots, we have demonstrated which scenarios will achieve the same power and how to maximize power. For a fixed number of subgroups (and fixed subgroup size and correlations), we can increase power by increasing the number of measurement times; similarly, for a fixed number of follow-up visits, we can increase power by increasing the number of subgroups.

In addition, we have presented a real world application of these formulas, which will become more important as more and more psycho-social interventions are developed and need to be tested. Through our simulations we demonstrate that given fixed values of the correlations, for a set power, we can decrease the required number of subgroups by increasing the size of the subgroups and the number of measurement times. It must be noted that the investigator will likely have limited control over some of the parameters and is more likely to increase power by increasing the total number of measurements that can be taken.

One concern in study planning is accounting for missing data. Since we can only know the impact and amount of missing data after the data have been observed, a common practice when designing a study is to assume no missing data and then inflate the sample size according to the expected amount of missing data. In this study, with subgroup heterogeneity in only one of the arms, we recommend assuming non-differential missing data and inflating the sample size in both the experimental and control arms using the same factor. One possible suggestion for the longitudinal setting would be to increase the total number of measurements n_M by calculating $n_M^* = n_m/$(proportion expect to observe), and then calculate the required combination of n, k_E, and n_T based on n_M^*.

This paper addresses both the simple and longitudinal settings with continuous outcomes in which subgroup size in the intervention group may very. Other things to consider for future studies may include: dichotomous outcomes; attrition over time; and the addition of other covariates to the models.

Acknowledgments

We are grateful to the editors and the reviewers for their insightful comments which have led to important improvements in the paper.

Contract/grant sponsor: National Institutes of Health grants: UL1RR025747, RO1HL57444, and PO1CA142538

Appendix

A more general mixed level mixed-effects linear model which allows the random effects for subgroup and subject levels to interact with time is:

$$Y_{ijl}=\beta_0+\beta_1 Trt_i+\beta_2 T_l+\gamma Trt_i\times T_l+b_{1,i}\times Trt_i+b_{2,i}\times Trt_i\times T_l+b_{1,j(i)}+b_{2,j(i)}\times T_l+\varepsilon_{ijl}, \quad (11)$$

where β_0 and $\beta_0 + \beta_1$ represent the pre-intervention main effects for the control group and experimental group, respectively, β_2 represents the main effect for time, and γ is the interaction effect between intervention and time we are interested in testing. The ε_{ijl} are the error terms, normally distributed as $N(0, \sigma_0^2)$; the $b_{1,j(i)}$ and $b_{2,j(i)}$ which are assumed to follow normal distribution with mean 0 and variance $\sigma_{2,1}^2$ and $\sigma_{2,2}^2$ respectively, represent the random effects at level two, the subject level; and the $b_{1,i}$ and $b_{2,i}$, the level three random effects for subgroups, are distributed as $N(0, \sigma_{E,1}^2)$ and $N(0, \sigma_{E,2}^2)$. It is assumed that the level two and level three random effects are independent of each other and the ε_{ijl}. Note that the random effects interact with the time by including $b_{2,i}$ and $b_{2,j(i)}$, considering that some individual or subgroups may benefit more from the intervention than others over time.

Based on (11), it can be shown that $E(Y_{ijl}) = \beta_0 + \beta_1 + (\beta_2 + \gamma) T_l$ and $Var(Y_{ijl})=\sigma_{E,1}^2+T_l^2\sigma_{E,2}^2+\sigma_{2,1}^2+T_l^2\sigma_{2,2}^2+\sigma_0^2$ for participants in the experimental arm and $E(Y_{ijl}) = \beta_0 + \beta_2 T_l$ and $Var(Y_{ijl})=\sigma_{2,1}^2+T_l^2\sigma_{2,2}^2+\sigma_0^2$ for participants in the control arm. Therefore, the ICC among repeated subgroup observations for the experimental arm is

$$\rho_{2,T_l,T_l'}=Corr(Y_{ijl}, Y_{ijl'})=\frac{\sigma_{E,1}^2+\sigma_{E,2}^2 T_l T_l'}{\sqrt{(\sigma_{E,1}^2+T_l^2\sigma_{E,2}^2+\sigma_{2,1}^2+T_l^2\sigma_{2,2}^2+\sigma_0^2)}\sqrt{(\sigma_{E,1}^2+T_l'^2\sigma_{E,2}^2+\sigma_{2,1}^2+T_l'^2\sigma_{2,2}^2+\sigma_0^2)}}$$

and the correlation for observations from a given subject from the experimental arm is

$$\rho_{1,T_l,T_l'}=Corr(Y_{ijl}, Y_{ijl'})=\frac{\sigma_{E,1}^2+\sigma_{E,2}^2 T_l T_l'+\sigma_{2,1}^2+\sigma_{2,2}^2 T_l T_l'}{\sqrt{(\sigma_{E,1}^2+T_l^2\sigma_{E,2}^2+\sigma_{2,1}^2+T_l^2\sigma_{2,2}^2+\sigma_0^2)}\sqrt{(\sigma_{E,1}^2+T_l'^2\sigma_{E,2}^2+\sigma_{2,1}^2+T_l'^2\sigma_{2,2}^2+\sigma_0^2)}},$$

and for the control arm is

$$Corr(Y_{i1l}, Y_{i1l'})=\frac{\sigma_{2,1}^2+\sigma_{2,2}^2 T_l T_l'}{\sqrt{\sigma_{2,1}^2+T_l^2\sigma_{2,2}^2+\sigma_0^2}\sqrt{\sigma_{2,1}^2+T_l'^2\sigma_{2,2}^2+\sigma_0^2}}.$$

References

1. McHutchinson JG, Manns M, Patel K, Poynard T, Lindsay KL, Trepo C, Dienstag J, Lee WM, Mak C, Garaud JJ, et al. Adherence to combination therapy enhances sustained response in genotype-1-infected patients with chronic hepatitis c. Gasteroenterology. 2002; 123(4):1061–1069.
2. Bronowicki JP, Ouzan D, Asselah T, Desmorat H, Zarski JP, Foucher J, Bourliere M, Renou C, Tran A, Melin P, et al. Effect of ribaviron in genotype 1 patients with hepatitis c responding to pegylated interferon alfa-2a plus ribaviron. Gasteroenterology. 2006; 131(4):1040–1048.
3. Hoover DR. Clinical trials of behavioural interventions with heterogeneous teaching subgroup effects. Statistics in Medicine. 2002; 21:1351–1364. [PubMed: 12185889]

4. Donner A, Birkett N, Buck C. Randomization by cluster sample size requirements and analysis. American Journal of Epidemiology. 1981; 116(6):906–914. [PubMed: 7315838]

5. Heo M, Leon AC. Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials. Statistics in Medicine. 2009; 28(6):1017–1027. [PubMed: 19153969]

6. Liu A, Shih W, Gehan E. Sample size and power determination for clustered repeated measurements. Statistics in Medicine. 2002; 21:1787–1801. [PubMed: 12111912]

7. Teerenstra S, Moerbeek M, van Achterberg T, Pelzer BJ, Borm GF. Sample size calculations for 3-level cluster randomized trials. Clinical Trials. 2008; 5:486–495. [PubMed: 18827041]

8. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics Bulletin. 1946; 2(6):110–114. [PubMed: 20287815]

9. DiSantostefano RL, Muller KE. A comparison of power approximations for satterthwaite's test. Communications in Statistics: Simulation and Computation. 1995; 24:583–593.

10. Morgan, DL. Focus Groups as Qualitative Research. Beverly Hills, CA: Sage University Paper Series on Qualitative Research Methods; 1988.

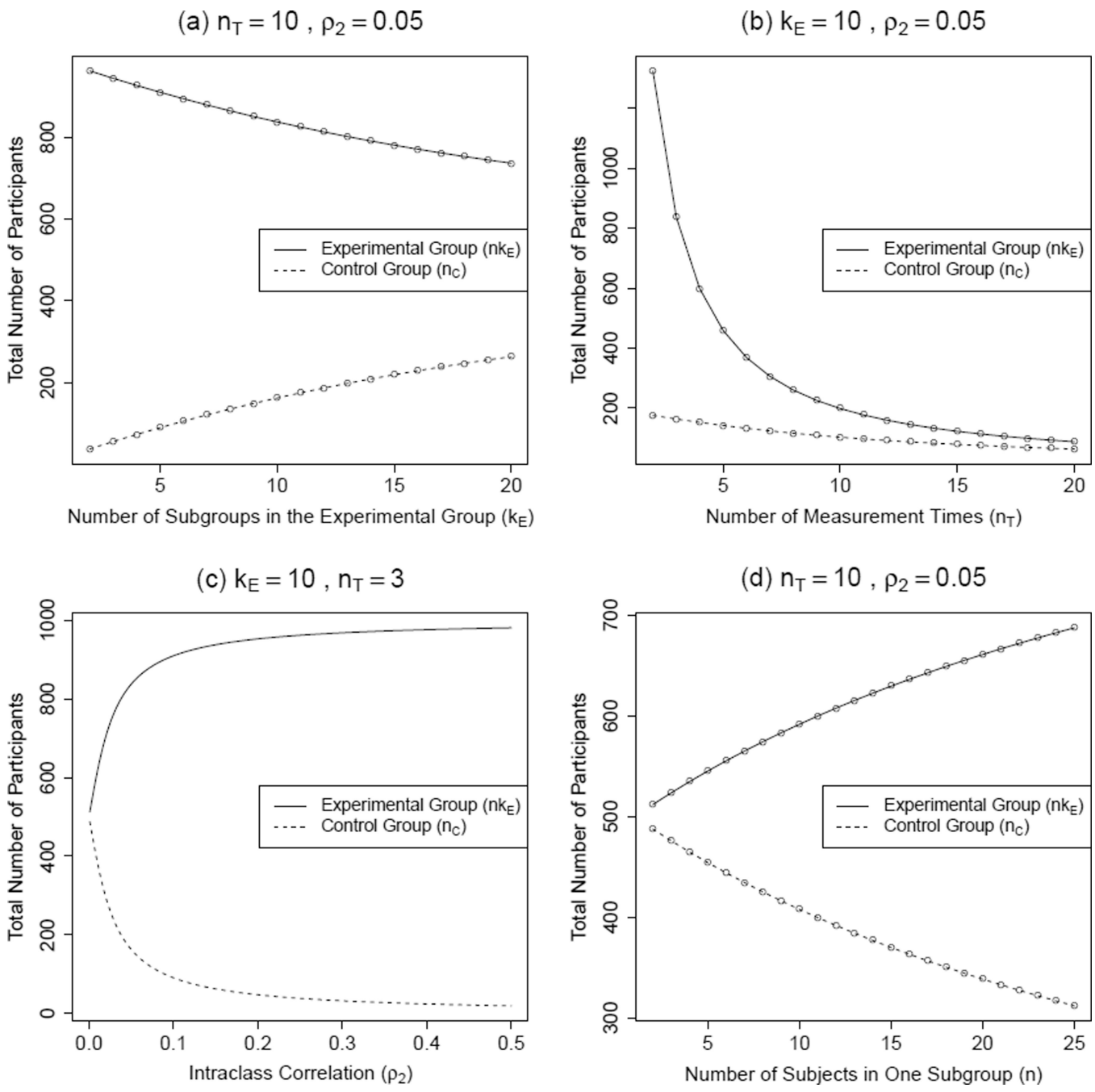

Figure 1.

The Inter-relation among k_E, n_T, ρ_2 and the total number of participants in the Experimental Arm and the Control Arm with $n_M = 3000$

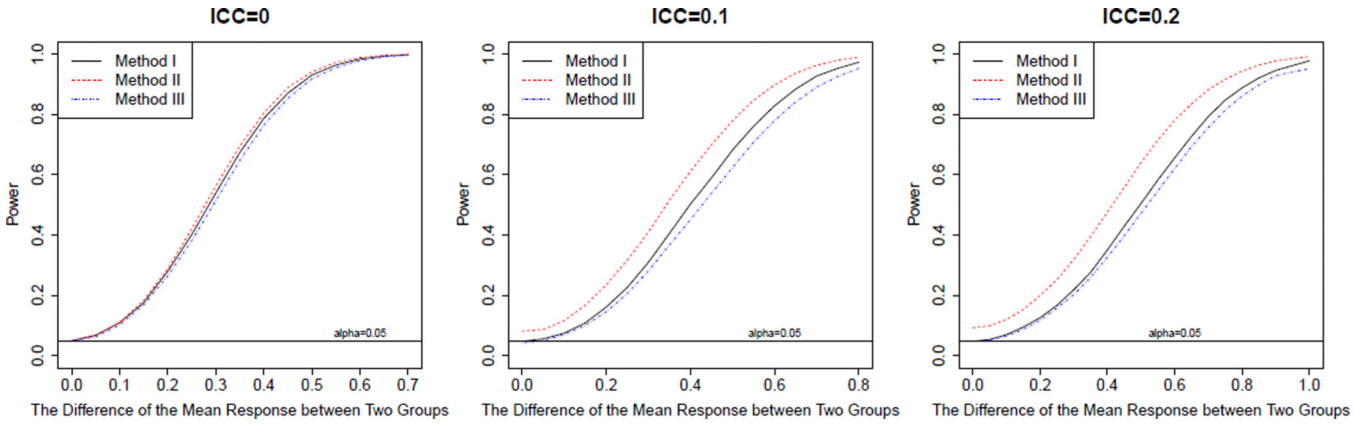

Figure 2. Power Curves for Different Test

Method I refers to the modified t-test. Method II refers to the standard t-test. Method III refers to the test considering subgroup heterogeneity in both arms. The number of subgroups in the experimental arm k_E is 10, within each subgroup there are 10 subjects. Panel on the left corresponds to the scenario where ICC is set to 0, panel in the middle corresponds to the scenario where ICC is set to 0.1, and panel on the right corresponds to the scenario where ICC is set to 0.2.

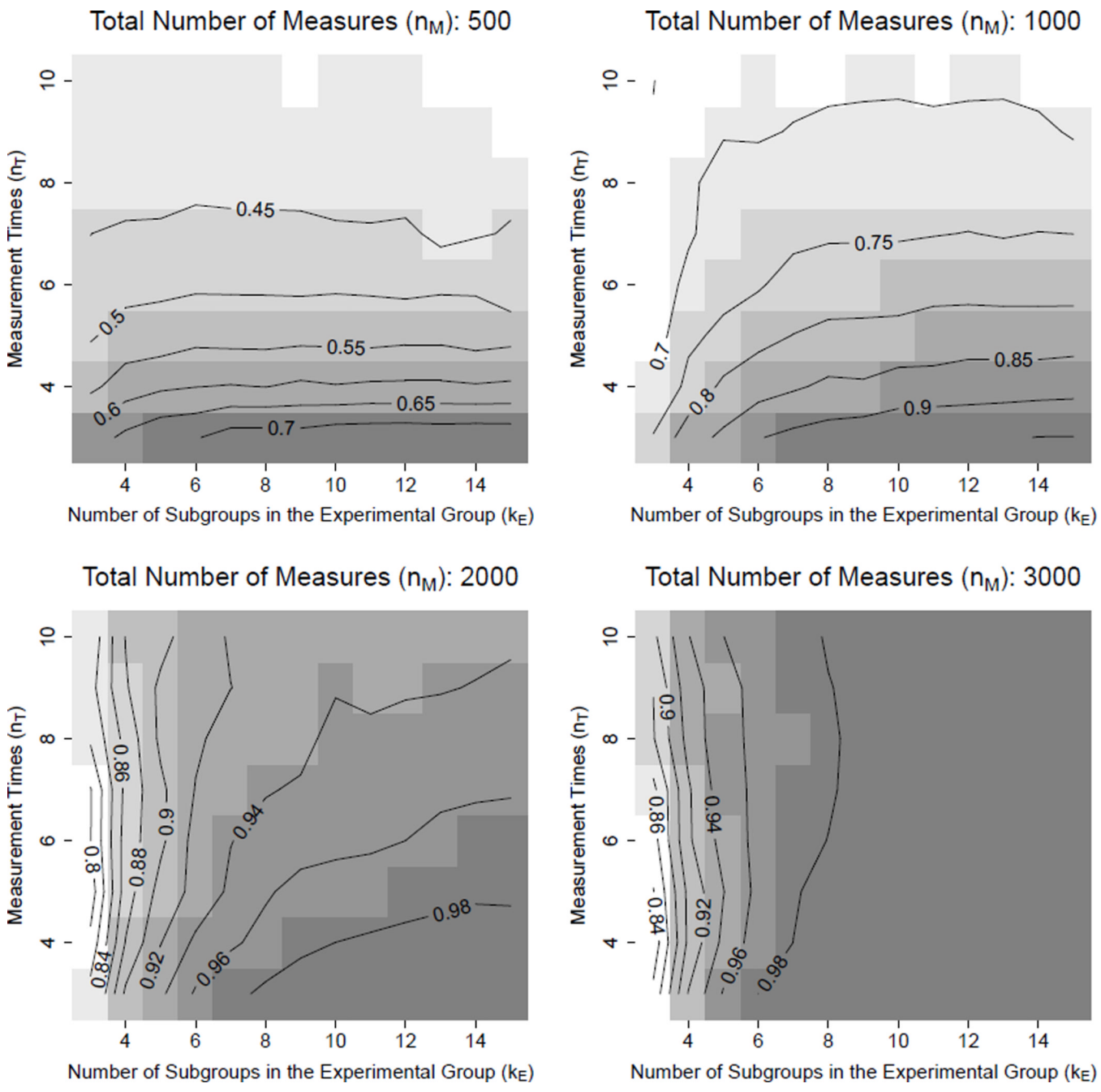

Figure 3.
Power to Detect a γ of 0.04 Standard Deviation between Two Groups with Different Combinations of (n_T, k_E) under Various Cost Constraint with $\rho_1 = 0.4$ and $\rho_2 = 0.05$

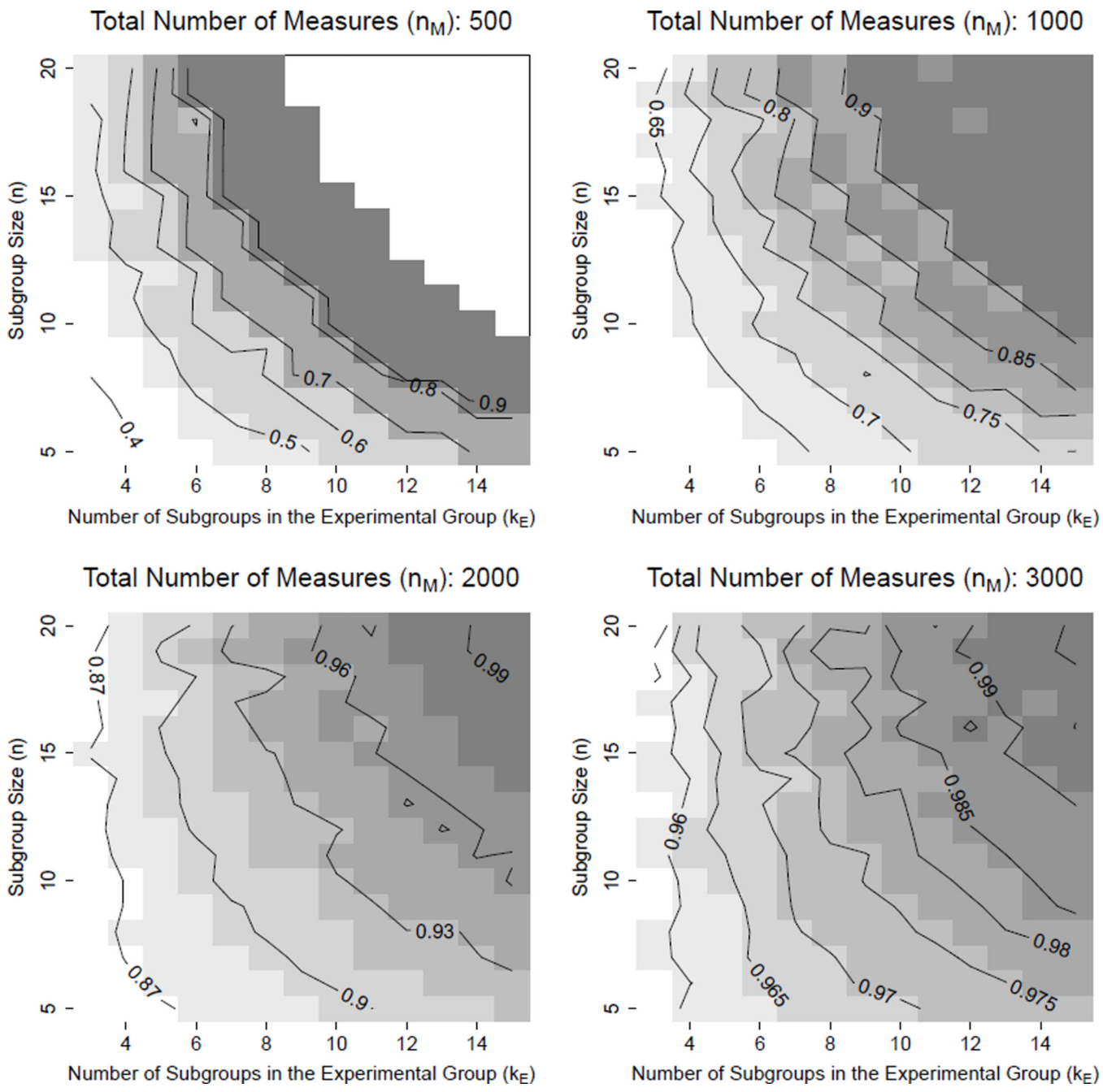

Figure 4.

Power to Detect a γ of 0.04 Standard Deviation between Two Groups with Different Combinations (n, k_E) under Various Cost Constraint with $\rho_1 = 0.4$ and $\rho_2 = 0.05$

Table 1

Comparison between Theoretical Power and Empirical Power for Modified t-test

					Type I error	Power (δ = 0.25)		Power (δ = 0.5)	
k_E	n_E	ρ	n_C	N	$\tilde{\alpha}$	ϕ	$\tilde{\phi}$	ϕ	$\tilde{\phi}$
5	50	0.2	18	68	0.046	0.11	0.10	0.26	0.26
		0.15	21	71	0.043	0.12	0.11	0.31	0.29
		0.1	26	76	0.051	0.14	0.15	0.38	0.38
		0.05	34	84	0.046	0.17	0.16	0.48	0.49
		0.01	46	96	0.049	0.21	0.20	0.61	0.62
		0	50	100	0.048	0.21	0.21	0.63	0.63
10	100	0.2	36	136	0.048	0.16	0.18	0.48	0.49
		0.15	43	143	0.054	0.19	0.19	0.57	0.53
		0.1	53	153	0.048	0.23	0.24	0.68	0.68
		0.05	69	169	0.055	0.29	0.27	0.80	0.80
		0.01	92	192	0.051	0.38	0.37	0.91	0.92
		0	100	200	0.050	0.41	0.42	0.93	0.93
20	200	0.2	71	271	0.050	0.28	0.27	0.79	0.76
		0.15	85	285	0.052	0.34	0.35	0.87	0.86
		0.1	105	305	0.050	0.41	0.40	0.93	0.93
		0.05	138	338	0.051	0.52	0.52	>0.99	0.99
		0.01	183	383	0.046	0.65	0.65	>0.99	>0.99
		0	200	400	0.050	0.69	0.69	>0.99	>0.99

Note: $N = n_C + n_E$ denotes the total sample size required. ϕ denotes the theoretical power, and $\tilde{\phi}$ denotes the empirical power. The theoretical type I error rate is set at .05, and $\tilde{\alpha}$ denotes the empirical type I error rate. Subgroup size is fixed at $n = 10$. Different scenarios are provided with varying effect sizes δ, subgroup number k_E, and ICC ρ.

Table 2

Comparison between Theoretical Power and Empirical Power for Testing Interaction Effects in Longitudinal Studies

						$\Delta_\gamma T_{end} = 0.4$						$\Delta_\gamma T_{end} = 0.6$	
n_T	ρ_1	k_E	n_E	n_C	$\tilde{\alpha}$	ϕ	$\tilde{\phi}$	k_E	n_E	n_C	$\tilde{\alpha}$	ϕ	$\tilde{\phi}$
3	0.4	15	150	104	0.045	0.82	0.81	7	70	49	0.057	0.84	0.84
	0.5	13	130	90	0.049	0.83	0.84	6	60	42	0.056	0.85	0.83
	0.6	10	100	69	0.046	0.82	0.81	5	50	35	0.051	0.86	0.87
6	0.4	11	110	76	0.049	0.83	0.83	5	50	35	0.053	0.84	0.84
	0.5	9	90	63	0.047	0.82	0.82	4	40	28	0.045	0.82	0.83
	0.6	4	40	28	0.049	0.90	0.90	4	40	28	0.045	0.90	0.90
12	0.4	7	70	49	0.049	0.85	0.85	3	30	21	0.052	0.84	0.85
	0.5	6	60	42	0.049	0.86	0.86	3	30	21	0.052	0.90	0.91
	0.6	5	50	35	0.053	0.88	0.87	2	20	14	0.053	0.84	0.84

Note: ϕ denotes the theoretical power, and $\tilde{\phi}$ denotes the empirical power. $\tilde{\alpha}$ is the empirical type I error rate. The subgroup size is fixed at $n = 10$. ICC ρ_2 is set to 0.05.

Table 3

Number of Subgroups k_E, Total Sample Size N, and Total Number of Measurements n_M Required to Achieve at Least 80% Power with 5% Type I Error Rate with $\rho_2 = 0.05$

		ρ_1	0.3			0.5			0.7		
			k_E	N	n_M	k_E	N	n_M	k_E	N	n_M
$n_T = 5$	$\Delta_\gamma T_{end} = 0.2$	$n = 6$	66	713	3565	48	519	2595	29	314	1570
		$n = 8$	52	725	3625	37	516	2580	23	321	1605
		$n = 10$	44	744	3720	31	524	2620	19	322	1610
	$\Delta_\gamma T_{end} = 0.5$	$n = 6$	11	119	595	8	87	435	5	54	270
		$n = 8$	9	126	630	6	84	420	4	56	280
		$n = 10$	7	119	595	5	85	425	3	51	255
$n_T = 7$	$\Delta_\gamma T_{end} = 0.2$	$n = 6$	57	616	4312	41	443	3101	25	270	1890
		$n = 8$	45	627	4389	32	446	3122	20	279	1953
		$n = 10$	38	643	4501	27	457	3179	16	271	1897
	$\Delta_\gamma T_{end} = 0.5$	$n = 6$	10	108	756	7	76	532	4	44	308
		$n = 8$	8	112	784	6	84	588	4	56	392
		$n = 10$	6	102	714	5	85	595	3	51	357

Note: The total sample size in the experimental group can be calculated by nk_E, and the total sample size of the control group can be obtained via (5).

Table 4

Power Table for the Example with Fix Number of Total Measurements and $\rho_2 = 0.05$

		n_T		5				7		
		ρ_1		**0.3**	**0.5**	**0.7**		**0.3**	**0.5**	**0.7**
$n_M = 500$	$\Delta_\gamma T_{end} = 0.2$	$n = 6$	$k_E = 9$	0.18	0.23	0.36	$k_E = 6$	0.15	0.19	0.29
		$n = 8$	$k_E = 7$	0.18	0.23	0.35	$k_E = 5$	0.15	0.20	0.30
		$n = 10$	$k_E = 5$	0.17	0.22	0.33	$k_E = 4$	0.15	0.20	0.30
	$\Delta_\gamma T_{end} = 0.5$	$n = 6$	$k_E = 9$	0.74	0.87	0.98	$k_E = 6$	0.64	0.78	0.94
		$n = 8$	$k_E = 7$	0.73	0.86	0.98	$k_E = 5$	0.65	0.80	0.95
		$n = 10$	$k_E = 5$	0.70	0.84	0.97	$k_E = 4$	0.64	0.79	0.95
n_M =1000	$\Delta_\gamma T_{end} = 0.2$	$n = 6$	$k_E = 18$	0.31	0.41	0.61	$k_E = 13$	0.27	0.36	0.54
		$n = 8$	$k_E = 14$	0.31	0.41	0.61	$k_E = 10$	0.27	0.35	0.53
		$n = 10$	$kE = 11$	0.30	0.40	0.60	$k_E = 8$	0.26	0.35	0.52
	$\Delta_\gamma T_{end} = 0.5$	$n = 6$	$k_E = 18$	0.96	> 0.99	> 0.99	$k_E = 13$	0.92	0.98	> 0.99
		$n = 8$	$k_E = 14$	0.96	> 0.99	> 0.99	$k_E = 10$	0.92	0.98	> 0.99
		$n = 10$	$k_E = 11$	0.95	0.99	> 0.99	$k_E = 8$	0.91	0.97	> 0.99
$n_M = 2000$	$\Delta_\gamma T_{end} = 0.2$	$n = 6$	$k_E = 37$	0.55	0.70	0.90	$k_E = 26$	0.48	0.61	0.83
		$n = 8$	$k_E = 28$	0.55	0.69	0.89	$k_E = 20$	0.47	0.61	0.82
		$n = 10$	$k_E = 23$	0.54	0.69	0.89	$k_E = 16$	0.46	0.60	0.81
	$\Delta_\gamma T_{end} = 0.5$	$n = 6$	$k_E = 37$	> 0.99	> 0.99	> 0.99	$k_E = 26$	> 0.99	> 0.99	> 0.99
		$n = 8$	$k_E = 28$	> 0.99	> 0.99	> 0.99	$k_E = 20$	> 0.99	> 0.99	> 0.99
		$n = 10$	$k_E = 23$	> 0.99	> 0.99	> 0.99	$k_E = 16$	> 0.99	> 0.99	> 0.99

Stat Med. Author manuscript; available in PMC 2014 May 30.

Table 5

Number of Visits and Subgroups Required to Achieve at Least 80% Power for the Example with $n_M = 500$, $\rho_1 = 0.3$ and $\rho_2 = 0.05$

	n_T	k_E
$n = 6$	4	11
$n = 8$	3	11
$n = 10$	3	9