Research Article

Statistics in Medicine

Received 27 February 2013, Accepted 9 April 2014 Published online 4 May 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6192

Information-based sample size re-estimation in group sequential design for longitudinal trials

Jing Zhou,[a*†] Adeniyi Adewale,[b] Yue Shentu,[b] Jiajun Liu[b] and Keaven Anderson[c]

Group sequential design has become more popular in clinical trials because it allows for trials to stop early for futility or efficacy to save time and resources. However, this approach is less well-known for longitudinal analysis. We have observed repeated cases of studies with longitudinal data where there is an interest in early stopping for a lack of treatment effect or in adapting sample size to correct for inappropriate variance assumptions. We propose an information-based group sequential design as a method to deal with both of these issues. Updating the sample size at each interim analysis makes it possible to maintain the target power while controlling the type I error rate. We will illustrate our strategy with examples and simulations and compare the results with those obtained using fixed design and group sequential design without sample size re-estimation. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: sample size re-estimation; information; group sequential design; adaptive design; longitudinal data analysis

1. Introduction

Clinical trials with longitudinal endpoints are very common. A key issue in designing such a trial is to determine how large of a study is necessary to detect a clinically important difference with a desired power. A traditional approach of sample size calculation for fixed design requires the investigator to specify a clinically meaningful difference to be detected, the significance level, a desired level of power, and any additional nuisance parameters (e.g. the error variance for continuous data and the control group response rate for binary data). As for repeated measure endpoints, Lu *et al.* [1,2] generalized a formula for calculating the sample size with nuisance parameters containing (i) correlation among longitudinal visits; (ii) standard deviation within longitudinal measurements for each subject; and (iii) retention rates in both treatment groups. For planning purposes, best guesses are made for the value of the nuisance parameters. However, there is a great concern that these assumptions of nuisance parameters based on previous studies are often unreliable because of differences in the study population, changes in medical practice, or the measurement techniques. Because incorrect assumptions can lead to substantial underpowering or overpowering to detect the clinically important difference, it may be prudent to check the validity of those assumptions using interim data from the study. There is a rich literature [3,4] discussing the sample size re-estimation methods to rescue the power. Wittes and Brittain [5] introduced the concept of an internal pilot design, which re-estimates the sample size in the midcourse of the study with no interim testing involved. Internal pilot designs have also been extended to different settings, besides normally distributed outcomes, such as repeated measures. Shih and Gould [6] described a method to re-estimate sample size in the repeated measure framework. However, it is only for a simplified setting, where the

[a]*Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.*
[b]*BARDS, Merck Research Laboratories, Rahway, NJ 07065, U.S.A.*
[c]*BARDS, Merck Research Laboratories, Upper Gwynedd, PA 19454, U.S.A.*
**Correspondence to: Jing Zhou, Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.*
[†]*E-mail: jingzhou@live.unc.edu*

parameter of interest is the rate of change (slope) of a continuous measurement. Zucker and Denne [7] extended Shih and Gould's model to a general setting in which missing and dropout are allowed and a linear combination of treatment effect over time can be set as the meaningful difference to be detected.

Group sequential design [8] promises to be more efficient because we are given an opportunity to terminate the study before the planned completion if there is strong evidence that the treatment effect is meaningfully large or the treatment is unlikely to be better than the control group. This design can benefit plenty of longitudinal trials. For instance, suppose we are doing a trial of weight loss, and the primary endpoint is weight loss at one year, with other measures at 3, 6, and 9 months. At the time of an interim analysis, some patients will have less than full follow-up but will have some follow-up measurements indicating a trend in their weight. If no weight loss is seen early in follow-up, it may be reasonable to stop a trial for futility. On the other hand, if substantial weight loss is observed and maintained, a convincing efficacy finding may result prior to the final planned analysis. Another example could be a trial of Alzheimer's disease in which an endpoint indicating cognitive decline such as Alzheimer's Disease Assessment Scale-Cognitive Subscale may be used. While the primary timepoint of interest may be after 18 months of treatment, intermediate measures may be taken at 6 and 12 months of follow-up. At the time of an interim analysis, some patients will have less than full follow-up but will have some follow-up measures indicating a trend in cognitive function that may be useful. An interim analysis may be able to stop a trial for futility or, if done later in the trial, may provide convincing results prior to the planned final analysis. While bounds may be set to be broad in order to avoid a premature stop, having bounds may provide useful guidance to a Data Monitoring Committee to help them avoid an early stop that may be due to a spurious finding. Both Galbraith and Marschner [9] and Kittelson *et al.* [10] discussed sequential methods when monitoring trials with longitudinal endpoints as well as making use of people who have not completed the study. Kittelson *et al.* [10] also provided a nice discussion when the outcomes are not measured according to the pretrial schedule. However, to adjust for the sequential monitoring stopping rules, both of them used the estimated information at the end of the study in computing the information-timing rather than the fixed maximal information from the pretrial design. Moreover, the two papers did not address the potential problem of insufficient power because of the incorrect initial sample size calculation if the variance assumption is incorrect. Burington and Emerson [11] focused on making flexible group sequential stopping rules when the actual interim analyses deviate from the design with respect to the number and timing. One can either choose to maintain the power or maintain the maximal sample size. But it did not cover the case where the primary endpoint of interest is longitudinal. Thus, with the goal of designing and analyzing a longitudinal trial using group sequential design along with the concern of insufficient power, it is natural to combine internal pilot designs into group sequential design in the longitudinal framework. Mehta and Tsiatis [12] and Tsiatis [13] initiated the use of information-based monitoring for implementing internal pilot designs in conjunction with group sequential methods but only for normal and binary endpoints. The counterpart for the longitudinal setting is missing, yet not trivial. Section 2 gives the background information on how to determine the sample size for fixed design and group sequential design, respectively. In Section 3, we introduce the information-based sample size re-estimation method in group sequential design to be utilized in longitudinal trials. Adaptation rules for updating sample size have been developed that will be described and illustrated by examples. Section 4 provides the simulation results for our method compared with fixed design and group sequential design without sample size re-estimation. A simple data analysis example is presented in Section 5. Finally, Section 6 contains a discussion and summary of the results.

2. Sample size determination for longitudinal analysis

2.1. *Model notation*

We model the longitudinal data as in Liang and Zeger [14] that include the baseline value as part of the response vector. The marginal mean model is given as

$$E(Y_{ijt}) = \mu_0 + \gamma_{jt} I(treatment = j) I(time = t; t > 0), t = 0, 1, \ldots, T, \quad (1)$$

where i indexes the subject, j indexes the treatment group ($j = 1$ for the control group and $j = 2$ for the active treatment group), and t indexes the time point ($t = 0$ for the baseline and $t = 1, \ldots, T$ for the post baseline time points). In addition, μ_0 is the mean response at $t = 0$, which is constrained to be the same for both treatment groups because of randomization. As a result, the baseline measurement is not considered

as an 'outcome' to treatment although it is included in the response vector together with the post baseline measurements. Hereafter, we will refer to (1) as the constrained longitudinal data analysis (cLDA) model. The parameter γ_{jt} is the effect of change from baseline at time t for treatment j; hence, $\mu_{jt} = \mu_0 + \gamma_{jt}$ is the mean at time t for treatment j. Let $\theta_j = (\mu_{j1}, \ldots, \mu_{jT})'$ denote the mean vector for post baseline measurements for treatment j. The mean parameters for model (1) can be written as $\psi = (\mu_0, \theta_1', \theta_2')'$.

The cLDA model assumes that baseline and post baseline values are jointly multivariate normal with $\Sigma = \{\sigma_{st} : s, t = 0, 1, \ldots, T\}$. This matrix can be represented as a correlation matrix sandwiched by the diagonal matrix of standard deviations where the correlation matrix is given by $R = \{\rho_{st} : s, t = 0, 1, \ldots, T\}$, and the standard deviations are with respect to the pure error within each longitudinal measurement. Let us denote n_{jt} as the subjects retained at time t in treatment j with the assumption of a monotone missing data pattern and $n_j = n_{j0}$ as the total number of subjects in treatment j at baseline. Define $r_{jt} = n_{jt}/n_j$ as the proportion of enrolled subjects retained at time t in treatment j. Note that the retained people include those that are still under active follow-up but exclude those who drop out. The drop-out rate, the proportion of enrolled subjects dropped out between time t and $t + 1$, is $p_{jt} = (n_{jt} - n_{j,t+1})/n_j$. It follows immediately that $p_{jt} = r_{jt} - r_{j,t+1}$.

2.2. *Fixed design*

Suppose we are interested in a linear contrast of the treatment means across time,

$$\delta = c'(\theta_2 - \theta_1), \tag{2}$$

where c is a contrast vector of length T corresponding to the T post baseline assessment time points. For instance, $c = (0, \ldots, 0, 1)$ is for treatment comparison at the last time point. If we want to detect the treatment effect δ_a, the Fisher information, I, to be needed based on a two-sided Z-test for $H_0 : \delta = 0$ versus $H_a : \delta = \delta_a$ with power $1 - \beta$ at significant level, α can be derived as

$$I = \left(\frac{Z_{\alpha/2} + Z_\beta}{\delta_a} \right)^2. \tag{3}$$

Now, we can determine the sample size with the knowledge that

$$I = Var^{-1}(\hat{\delta}), \tag{4}$$

where $Var(\hat{\delta})$ is a function of sample size with some nuisance parameters. $\hat{\delta}$ denotes the estimate of δ to be calculated from data. This variance varies among different types of trials. For longitudinal trials, in particular, based on the cLDA model we defined earlier in (1), the variance inverse of $\hat{\delta}$ incorporating missingness is given by

$$Var^{-1}(\hat{\delta}) = \left(\frac{c' S \Lambda_1^{-1} S c}{n_1} + \frac{c' S \Lambda_2^{-1} S c}{n_2} \right)^{-1}, \tag{5}$$

where n_j is sample size for treatment j, $j = 1, 2$. For simplicity, we assume the randomization ratio is 1, so $n_1 = n_2 = \frac{N}{2}$ though extensions are trivial. The parameter c is denoted as the previous, $S = diag(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{TT}})$ denotes the standard deviations at post baseline time points. Λ_j is given by

$$\Lambda_j = \sum_{t=1}^{T} p_{jt} \begin{pmatrix} R_{tt\cdot 0}^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \tag{6}$$

where $R_{tt\cdot 0}^{-1} = R_{tt} - R_{0t}' R_{0t}$, $R_{tt} = \{\rho_{ij} : i, j = 1, \ldots, t\}$ and $R_{0t} = (\rho_{01}, \ldots, \rho_{0t})$. The proof of the derivation of $Var(\hat{\delta})$ for cLDA is in Lu *et al.* [2]. We note that the nuisance parameters in (5) are essentially R, S, and r_j, where R, S are the correlation matrix and standard deviations respectively, and $r_j = (r_{j1}, \ldots, r_{jT})$ is the retention rate across time for treatment j. Without loss of generality, we assume $r_1 = r_2$ hereafter. Therefore, connecting (3), (4), and (5) while assuming the nuisance parameters (R, S, and r_1), the calculation of sample size is straightforward.

2.3. *Group sequential design*

Group sequential design has the advantage to stop early for efficacy or futility. Rather than performing only one analysis at the end of the study, we perform up to K analyses at interim monitoring times $\tau_1, \tau_2, \ldots \tau_K$, respectively, and terminate the study at the first interim look that rejects the null hypothesis. However, this flexibility to possibly terminate early comes at a cost. In particular, the type I error is inflated because of the multiple interim hypothesis tests, if we keep the stopping boundaries unchanged. We thus need to adjust the stopping criteria appropriately such that the type I error is controlled. Moreover, in order to achieve a power of $1-\beta$ while controlling the type I error rate, the information has to be inflated by

$$I_{max} = \left(\frac{Z_{\alpha/2} + Z_{\beta}}{\delta_a}\right)^2 \times IF\left(\Delta, \alpha, \beta, K\right), \tag{7}$$

where I_{max} is denoted as the information at the final look, which is a counterpart of the I in (3) under the framework of group sequential design. The function IF$(\cdot)$ is an inflation factor that depends on α, β, K, and Δ. The parameter Δ is defined with respect to the shape of the stopping boundaries over the K repeated tests. Mehta and Tsiatis [12], relying on theoretical results by Scharfstein *et al.* [15], has the detailed derivation and examples of the inflation factor. Keep in mind that this maximum information does not require any knowledge of unknown nuisance parameters.

In practice, however, an estimated number of patients is required at the time of study design. The corresponding sample size determination follows the same strategy as that used in fixed design. One can similarly connect (7), (4), and (5) to solve the maximum sample size (N_{max}). In short, the required maximum sample size can be computed in the following two steps:

(1) Utilize the `'gsDesign'` R package developed by Anderson [16] to calculate I_{max} once we define the necessary parameters in (7),
(2) Convert I_{max} to N_{max} by using (5) that is essentially a function of N_{max} with some nuisance parameters R, S, and r_1.

These calculations are possible in many other group sequential design packages (e.g. RCTdesign in R, PEST, EaSt) as long as the correct variance and information timing are used. Given accurate assumptions of nuisance parameters, collecting N_{max} subjects will, in the end, result in obtaining I_{max} while achieving the desired power and maintaining the type I error. For example, if we plan for four looks with an equal-spaced information-based design, 25% of I_{max} is expected at each interim analysis. But in real world studies, the sample size may be incorrect because the nuisance parameters are unknown, and it often happens that we do not have good estimates for these parameters at hand. The power will be affected as a consequence. Thus, our method with the aim of tackling this problem is introduced in the next section. It is noted that the inflation factor only allows us to maintain the power and control the type I error provided that the assumptions of R, S, and r_1 are correct. Instead of maintaining power by adjusting the sample size, it is also possible to fix the total sample size at N_{fix} (sample size for a fixed-sample trial) and then evaluate the effect of the monitoring plan on trial power, which is however not a focus of this paper. A design with analyses spaced by equal amounts of information is assumed from now just for simplicity in illustration; it is easy to extend to an unequally spaced information.

3. Information-based sample size re-estimation

3.1. *Sample size re-estimation*

Define $I(\tau_k)$ as the information at the k^{th} interim analysis. Assuming complete follow-up for observations, it can be easily shown in (5) that,

$$\frac{I(\tau_k)}{I_{max}} = \frac{N(\tau_k)}{N_{max}}. \tag{8}$$

Thus, one can re-estimate N_{max} at each interim using

$$N^*_{max} = N(\tau_k) / \frac{I(\tau_k)}{I_{max}}, \tag{9}$$

where $N(\tau_k)$ is the sample size of subjects with completed longitudinal visits at the k^{th} look, $I(\tau_k) = Var^{-1}(\hat{\delta}(\tau_k))$ is estimated from the data, and I_{max} is fixed under the design in (7). Note the denominator on the right hand side is an evaluated information fraction that is to be compared with the anticipated information fraction (e.g. 25% if currently at first interim with four looks in total planned). If it is larger than planned, N^*_{max} will be smaller than original N_{max} and vice versa. Updating the maximum sample size at each look can correct inaccurate assumptions for nuisance parameters and maintain the power while controlling the type I error rate.

This approach is fairly easy to understand and implement without unblinding the trial; however, it has the drawback that only completed subjects are contributing to the interim analysis. An alternative way to re-estimate the maximum sample size is to first estimate the nuisance parameters using all available data at the current interim, and then use them as input in the calculation of N_{max} as discussed in Section 2.3. It can make use of all data including ongoing patients, but it loses the simplicity of the previous method as we need to estimate the many nuisance parameters at each interim and have to use a statistical package to obtain the updated sample size. When the enrollment is slow, the method in (9) is more attractive in practice because the additional information provided by the ongoing patients may be neglectable. Otherwise, the latter method is recommended. The power analysis results based on the second approach are provided later in Section 4.

3.2. *Adaptation*

We develop a sample size adaptation rule in the following based on the practical characteristics of clinical trials. Note that the 'overrun of patients' later in (b) stands for all the subjects enrolled so far including not only the completed/discontinued ones but also those still continuing, while the current sample size is only with regard to completed/discontinued patients.

(a) If current sample size is enough, meaning, I_{max} is reached, stop the trial regardless of whether efficacy or futility is detected.
(b) If overrun of patients is enough to provide sufficient information, stop enrolling more patients but keep collecting data for enrolled patients.
(c) If next planned sample size is enough, stop the enrollment when the updated maximum sample size is reached.
(d) If next planned sample size is not enough but original maximum sample size is, continue enrollment to the next planned interim.
(e) If original maximum sample size is not enough, use the updated maximum sample size but with an upper limit depending on practical aspects of certain trials (e.g. two times the previous maximum sample size) and continue enrollment to the interim analysis.

For illustration purposes, the following is an example of adaptation. Suppose the longitudinal study is to be designed for up to four interim monitoring looks including a final analysis; each subject is expected to have four longitudinal visits to the clinic after the baseline measurement and the corresponding N_{max} is 800 with all the necessary parameters assigned. Hence, $K = 4$, $T = 4$, and $N_{max} = 800$. Let $k = 2$, $N(\tau_2) = 400$, 480 patients have been enrolled and the current plan for $N(\tau_3) = 600$, the consequent adaptation rule at second interim can be represented as follows:

- If $N^*_{max} \leqslant 400$, stop the trial regardless.
- If $400 < N^*_{max} \leqslant 480$, stop enrolling more patients but keep collecting data for enrolled patients and do one final analysis.
- If $480 < N^*_{max} \leqslant 600$, stop the enrollment when N^*_{max} is reached and perform one final analysis.
- If $600 < N^*_{max} \leqslant 800$, continue the next planned interim.
- If $N^*_{max} > 800$, use min (2×800, N^*_{max}) as our new maximum sample size and continue the interim analysis with three-fourths of the new maximum sample size.

In actuality, the total sample size only needs to be increased when either the planned sample size has been reached (with incomplete follow-up), when the last interim analysis is performed, or when planning for increased patient enrollment and clinical supplies is needed.

Next, we use Figure 1 to illustrate the sample size adaptation for an entire trial if we use the method in (9). Once again, we plan four looks at the design stage and analyze 200, 400, 600, and 800 patients at a time. Different colors correspond to each interim look. At the first interim, we see that instead of anticipated 25% of I_{max}, 30% of I_{max} has been observed. The following re-estimation of N_{max} tells us 667

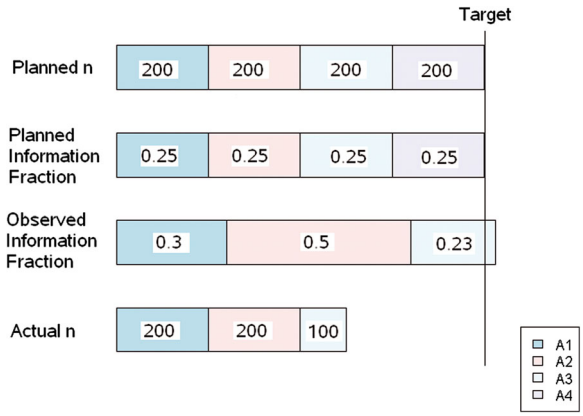

Figure 1. Adaptation example.

patients are required to obtain the maximum information in the end rather than the original 800 patients. Because 667 falls into the fourth bullet of the previous adaptation rule, we need to collect 200 more patients as planned and perform the second analysis. At that time, another 50% of I_{max} has been gathered. A total of 80% is considerably larger than what we anticipated (50%), which suggests that the assumption of nuisance parameters is very conservative. Therefore, the next analysis is planned as the final analysis at $N^*_{max} = 500$ and enrollment stops at that point. The final analysis was conducted with 1.03 times the I_{max} observed that is also evidence that our approach is useful to save time and resources while maintaining all of the good statistical characteristics. In addition to the previous sample size planning rules, the trial may stop if an efficacy or futility bound is crossed.

3.3. *Interim analysis procedure*

Using what we supposed in the previous subsection ($K = 4$, $N_{max} = 800$), we plan to enroll subjects continuously and conduct each interim analysis cumulatively for every 200 subjects with complete visits that have been gathered. Our interest is to test whether the treatment difference at the final look (δ_a) is 0.25. Using all the completed/drop-out data (for method 1) or the available data (for method 2) to fit the constrained longitudinal model (1) assuming unstructured covariance structure [17], the testing procedure at interim k is given in the following: $k = 1 \ldots K$,

(1) Estimate $I(\tau_k) = Var^{-1}(\hat{\delta}(\tau_k)) = s.e.^{-2}(\hat{\delta}(\tau_k))$.
(2) Estimate $T(\tau_k) = \frac{\hat{\delta}(\tau_k)}{s.e.(\hat{\delta}(\tau_k))}$.
(3) Update the actual information fraction vector up to current k^{th} interim:

$$\left(\frac{I(\tau_1)}{I_{max}}, \frac{I(\tau_2)}{I_{max}}, \ldots, \frac{I(\tau_k)}{I_{max}}, \frac{k+1}{K}, \ldots, 1\right), \tag{10}$$

where $\frac{I(\tau_1)}{I_{max}}, \frac{I(\tau_2)}{I_{max}}, \ldots, \frac{I(\tau_k)}{I_{max}}$ are observed information fraction up to k^{th} interim look, and $\frac{k+1}{K}, \ldots, 1$ are planned information fraction after k^{th} interim. Then, one can calculate the corresponding stopping upper and lower boundaries by updating bounds using the methods of Lan and DeMets [18]; this can be done, for example using the `'gsDesign'` R package.

(4)
$$\begin{cases} \text{stop for efficacy,} & \text{if } T(\tau_j) \geqslant \text{upper bound} \\ \text{continue,} & \text{if lower bound } < T(\tau_j) < \text{upper bound} \\ \text{stop for futility,} & \text{if } T(\tau_j) \leqslant \text{lower bound.} \end{cases}$$

One extra step is required here if we do not stop at the current analysis, that is, to re-estimate N_{max} as in Section 3.1 and to adapt the new maximum sample size discussed in Section 3.2. It is noted that besides futility or efficacy, the study is terminated when I_{max} is reached, or at the K^{th} final analysis.

4. Simulation study

As discussed earlier, incorrect assumptions of nuisance parameters will lead to an incorrect sample size that will affect the power. In this section, we verify through simulations that our method works as planned in the sense that the power is maintained while preserving the type I error rate. The expected sample size ($E(n)$) is another characteristic of interest. We define n, for group sequential design, as the number of enrolled when stopping early, but as the number of analyzed when stopping at the final look. For the fixed design, however, n is always N_{fix}. In the meantime, we compare performance of our approach with that of fixed design and that of group sequential design without the sample size recalculation. The procedure is composed of four parts: design, data generation, testing, and results comparison. All the trials are designed for 90% power to detect a treatment difference at the last (fourth post baseline) measurement of 0.25 using four-look one-sided O'Brien-Fleming stopping boundaries with a type I error 0.025. The assumptions for nuisance parameters ($R, S,$ and r_1) are that R_0 = compound symmetry with correlation coefficient 0.579, $S_0 = 0.8$, and $r_{10} = (0.91, 0.84, 0.77, 0.70)$. It is then straightforward to obtain the sample size for a fixed design (N_{fix}) and a group sequential design (N_{max}). Because the sample size only depends on $R, S,$ and r_1 through information, and the true values of these nuisance parameters (R, S, r_1) could be different from what we planned (R_0, S_0, r_{10}). Thus, we generate 1000 datasets under each of 18 different combinations of $(R, S,$ and $r_1)$ to see how the power and type I error behave; we expect that the one scenario with the assumed values will result in good power with type I error controlled. True R is chosen to be among compound symmetry (cs), toplitz, and AR(1); S is either 0.8 or 0.925, whereas the three options of r_1 are (0.84, 0.71, 0.60, 0.5), (0.91, 0.84, 0.77, 0.7), and (0.97, 0.95, 0.92, 0.9). We simulate data by considering the mean in the control group as (3.0, 2.8, 2.6, 2.4, 2.0). The true treatment effect is (0, 0.13, 0.17, 0.19, 0.25) if under the alternative or (0, 0, 0, 0, 0) if under the null. Because the treatment difference at the last measurement is of interest in the simulation, the contrast vector is (0, 0, 0, 1) excluding baseline. Cases with greater treatment effects would tend to stop early due to crossing the efficacy bound.

Figures 2 and 3 (left) show the power curves under 18 different combinations of true nuisance parameters. Each circle corresponds to one of the 18 scenarios while x axis is for the three different retention rates, different line color stands for the three correlation structures, and the line type denotes different standard deviations. The assumed nuisance parameters in both figures are the same. The dot with a cross symbol denotes that the nuisance parameters share the same values in design and in true data, whereas other 17 circles employed various true values of R, S, and r_1. As is visible in Figure 2, the power using our approach is well maintained around 90%, while the power under fixed design is not satisfactory under some scenarios where the nuisance parameters assumption deviates much from the true values. This is also seen in left plot in Figure 3 for group sequential design without sample size re-estimation. When checking the expected sample size in the right plot in Figure 3, we noticed that for about half of the cases

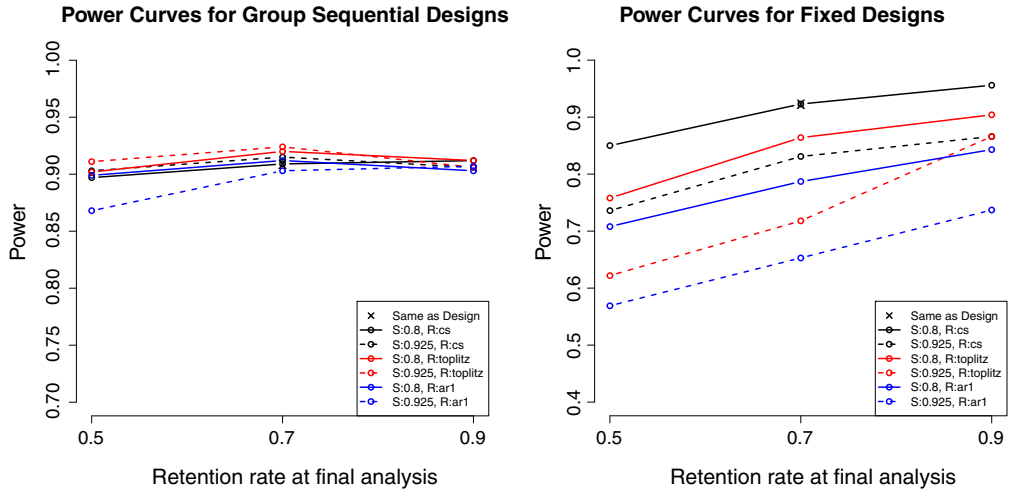

Figure 2. Power curves using our re-estimation method (left) versus fixed design (right).

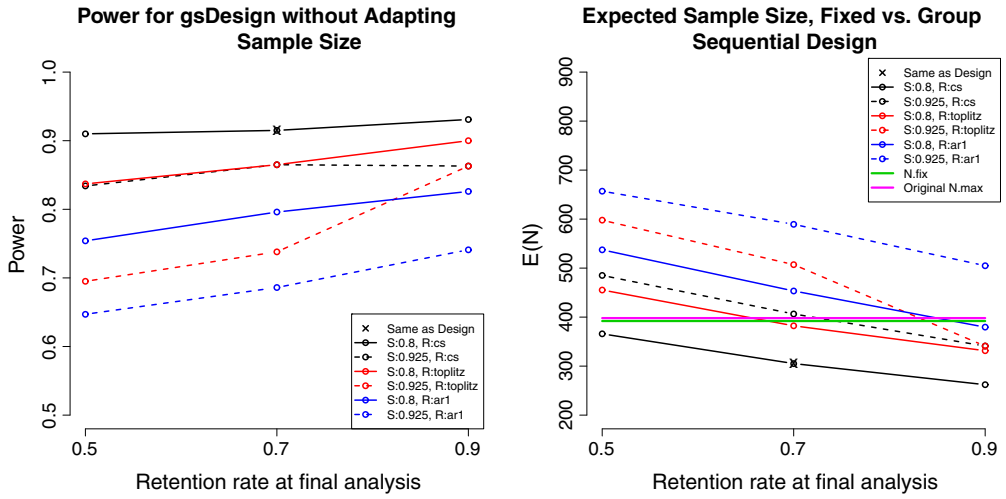

Figure 3. Left: power curves using group sequential design without re-estimation; right: expected sample size using our method versus fixed design.

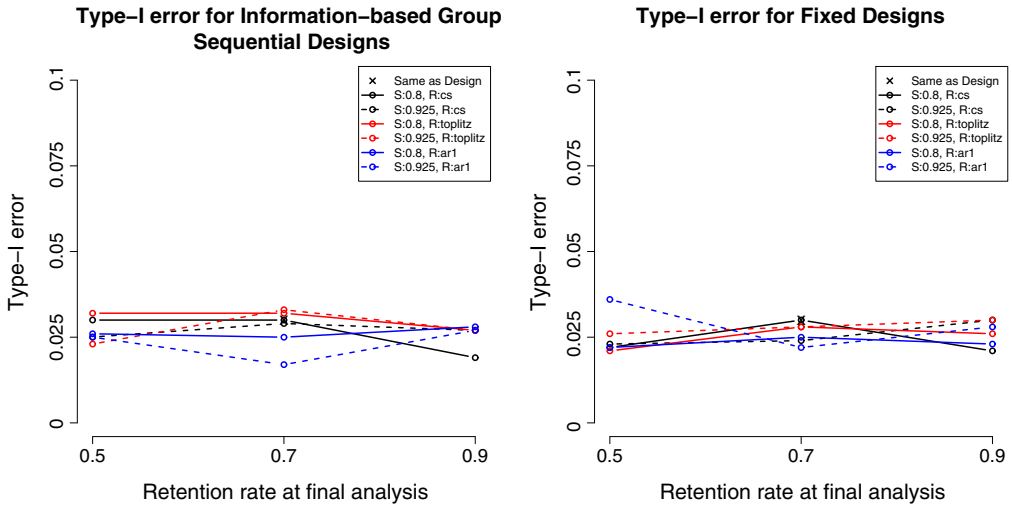

Figure 4. Type I error using our re-estimation method (left) versus fixed design (right).

for which we did not assume nuisance parameters accurately enough, it requires more patients for our method in group sequential design than that in fixed design.

On the other hand, under the null, Figures 4 and 5 (left) show that all three methods can control the type I error. Furthermore, the expected sample size in Figure 5 (right) is similar to what we have under the alternative. Table I provides the standard error for the power and type I error based on 1000 simulations. Symmetric and asymmetric two-sided tests were both examined as well to assess the performance, and they turn out to have very similar results to that for one-sided test; hence, they are omitted here. All the simulations are based on the second method in Section 3.1, although it is noticed that all the results look very similar to when we instead use the first method (i.e. information fraction method as in (9)) with completed and drop-out data only (results are omitted). The reason is that there is not much information gained by having around 10 more ongoing patients at each interim given the assumed slow enrollment.

The correlation coefficient in the previous simulations was set as 0.579. To check performance when varying the correlation coefficient, we keep the design parameters (R_0, S_0, and r_{10}) the same as the previous, let the true nuisance parameters $S = S_0$ and $r_1 = r_{10}$, but vary the true correlation structure using a correlation coefficient 0.3 or 0.8 under compound symmetry, toplitz, and AR(1). In Table II, targeted power is observed under our re-estimation approach with various correlation coefficients 0.3 and 0.8. In contrast, the fixed design and the group sequential design without sample size re-estimation lead to

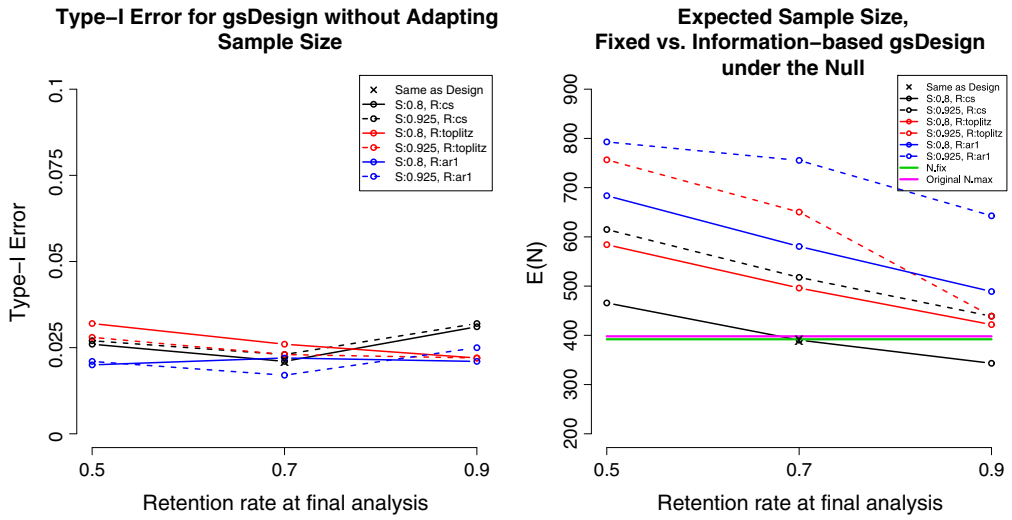

Figure 5. Left: type I error using group sequential design without re-estimation; right: expected sample size using our method versus fixed design.

Table I. Simulation error.

	S.E. of power			S.E. of type I error		
	Retention rate at final interim			Retention rate at final interim		
	0.5	0.7	0.9	0.5	0.7	0.9
R = cs, S = 0.8	0.010	0.009	0.009	0.005	0.005	0.004
R = cs, S = 0.925	0.009	0.009	0.009	0.005	0.005	0.005
R = toplitz, S = 0.8	0.009	0.009	0.009	0.006	0.006	0.005
R = toplitz, S = 0.925	0.009	0.008	0.009	0.005	0.006	0.005
R = ar1, S = 0.8	0.010	0.009	0.009	0.005	0.005	0.005
R = ar1, S = 0.925	0.011	0.009	0.009	0.005	0.004	0.005

Table II. Power based on varying correlation coefficients.

	Compound symmetry		Toplitz		AR(1)	
	0.3	0.8	0.3	0.8	0.3	0.8
Group sequential design	0.9	0.93	0.92	0.91	0.91	0.9
Fixed design	0.8	1	0.8	0.94	0.77	0.84
gsDesign without adapting sample size	0.82	0.98	0.83	0.95	0.81	0.87

Table III. Type I error based on varying correlation coefficients.

	Compound symmetry		Toplitz		AR(1)	
	0.3	0.8	0.3	0.8	0.3	0.8
Group sequential design	0.025	0.039	0.023	0.029	0.028	0.017
Fixed Design	0.021	0.033	0.023	0.027	0.02	0.02
gsDesign without adapting sample size	0.025	0.024	0.026	0.041	0.02	0.03

power ranging from 0.77 to 1, and from 0.81 to 0.98, respectively. As indicated in Table III, there is no significant problem of controlling type I error because of the simulation error for the three methods. The corresponding expected sample size are (422, 201, 459, 275, 468, 373) for the six scenarios given $N_{fix} = 392$ and the original $N_{max} = 398$.

Lastly, because the simulation is designed to detect a treatment difference at the last (fourth post baseline) measurement, it would be interesting to implement the fixed design and the group sequential design with and without sample size redetermination to analyze only the last time point as normal data using method by Mehta and Tsiatis [12]. Keeping the correlation coefficient at 0.579 and following the same simulation setup, as expected, in Figures 6 and 7 (left), different correlation structures do not make a difference for all three designs because we only use the last measurement for all patients. It also makes sense that the information-based method by Mehta and Tsiatis [12] is able to maintain the power when the retention rate is not too low. Similar feature is observed for the expected sample size as shown in Figure 7 (right). Figures 8 and 9 are the corresponding type I error and expected sample size under the null.

5. Example

We build functions in *R* to formalize our method and to perform a data analysis. Our data are motivated by clinical trials studying change in tumor size over time. Before analyzing the data, we need to know the sample size assignment for each interim by designing α, β, δ, one-sided test or two-sided test, number of planned looks, planned information fraction at each look, and nuisance parameters assumption (R_0, S_0, and r_{10}). We wish to detect 0.25 treatment difference for the last repeated measurements between two groups of patients by designing a study planning one interim look and one final analysis using a one-sided test with type I error 0.025, power 90%, and with a planned information fraction of (0.5, 1) for

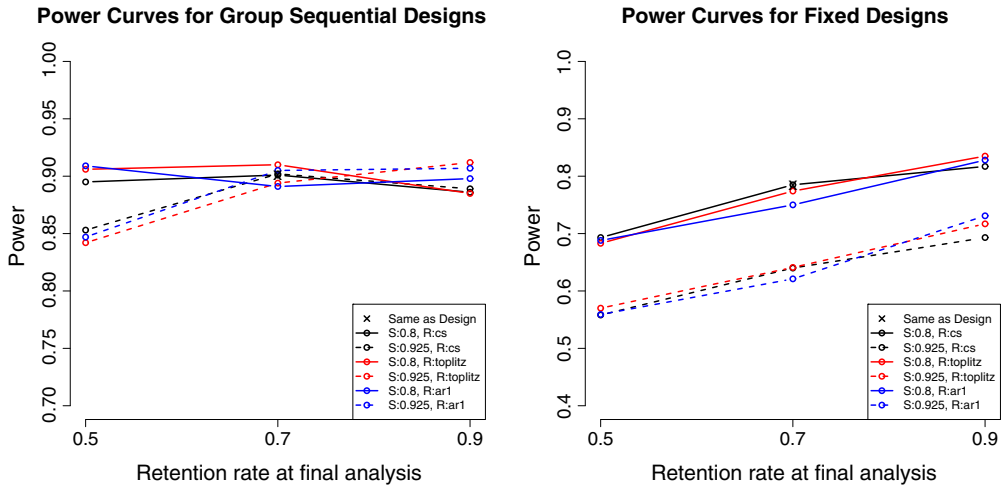

Figure 6. Simulation results based on the last time point only.

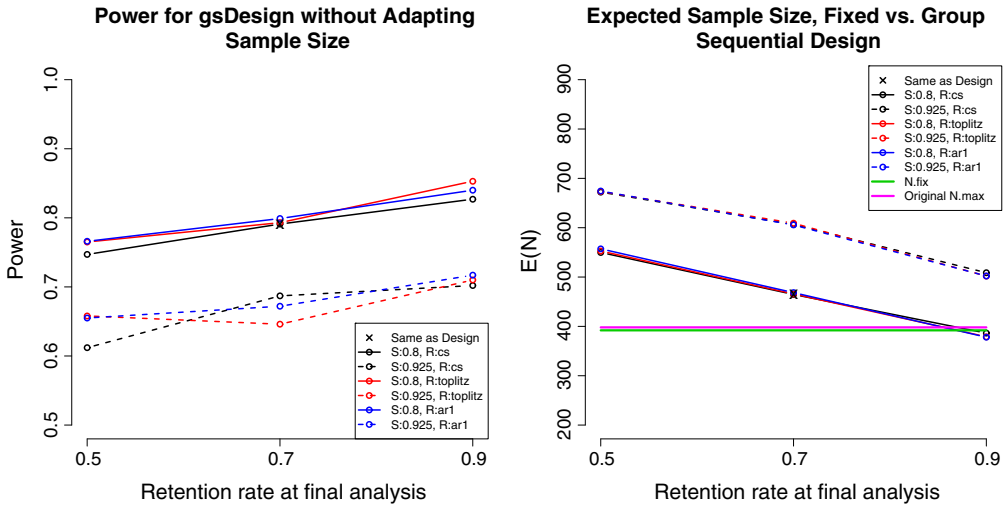

Figure 7. Simulation results based on the last time point only (continued).

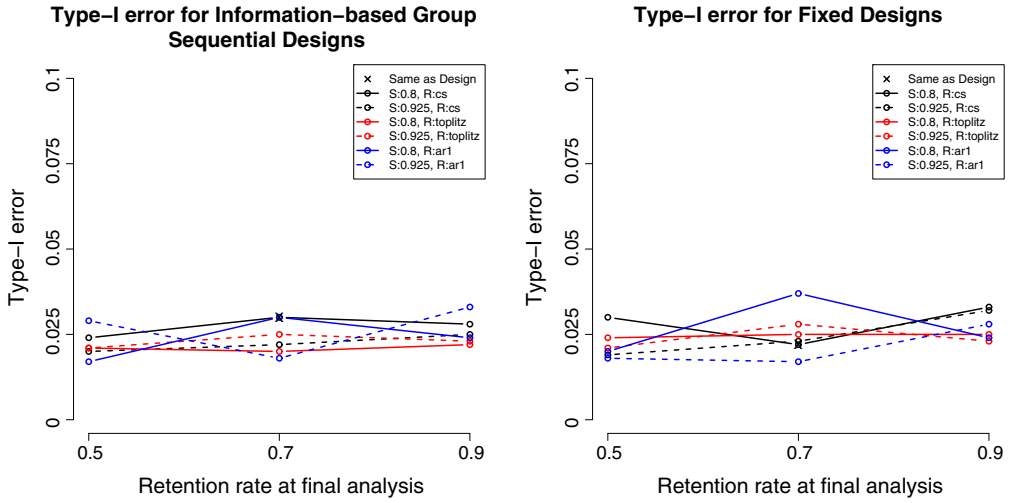

Figure 8. Simulation results based on the last time point only (continued).

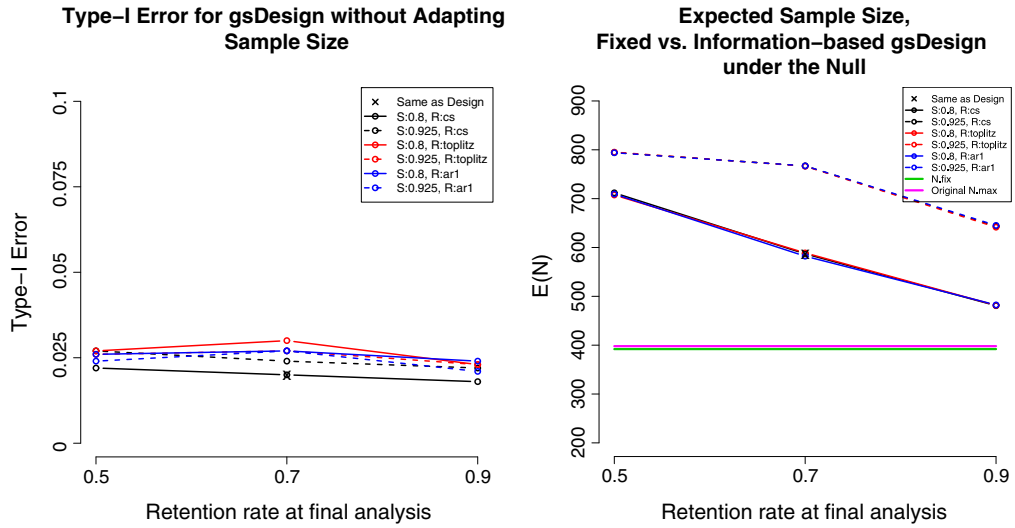

Figure 9. Simulation results based on the last time point only (continued).

clarity and simplicity. Each patient is expected to have four visits to the clinic to get the tumor measured. A monotone missing pattern is assumed for this study. The nuisance parameters assumed here are

$$R_0 = \begin{pmatrix} 1.000 & 0.579 & 0.579 & 0.579 & 0.579 \\ 0.579 & 1.000 & 0.579 & 0.579 & 0.579 \\ 0.579 & 0.579 & 1.000 & 0.579 & 0.579 \\ 0.579 & 0.579 & 0.579 & 1.000 & 0.579 \\ 0.579 & 0.579 & 0.579 & 0.579 & 1.000 \end{pmatrix},$$

$$S_0 = (0.925, 0.925, 0.925, 0.925, 0.925), r_{10} = (0.950.90, 0.85, 0.80).$$

The design and analysis are presented using our proposed information fraction approach in (9) because of the ease of implementation and illustration. The *R* program for the other re-estimation method is also available upon request. The corresponding N_{max} is then 466. The fixed design sample size for this longitudinal study calculated using the strategy in Section 3.1 is also 466 because in this case, the inflation factor in (7) is nearly 1. Hence, 233 patients (= 466/2) including completed and dropout shall be collected before analyzing the first interim analysis under group sequential design. A sample of typical clinical data is shown in Table IV.

Table IV. Longitudinal data with four visits.

Subject	Treatment	Enrollment time	Week	Y	Flag
1	1	0.0045	0	3.93	0
1	1	0.0045	1	2.95	0
1	1	0.0045	2	3.60	0
1	1	0.0045	3	2.24	0
1	1	0.0045	4	2.17	0
2	2	0.0079	0	3.53	0
2	2	0.0079	1	3.83	0
2	2	0.0079	2	2.63	0
2	2	0.0079	3	1.76	0
3	1	0.0927	0	2.01	1
3	1	0.0927	1	3.22	1
⋮	⋮	⋮	⋮	⋮	⋮

Table V. Interim analysis results.

Alpha	0.025
Power	0.9
Planned.timing1	0.5
Planned.timing2	1
Delta	0.25
First interim analysis	
ifStopTrial	Continue to the next interim
ifNextFinal	TRUE
update.act.t1	0.514
update.act.t2	1
orig.Nmax	466
new.Nmax	411
current.info	0.514
Final interim analysis	
ifStopTrial	Stop with sig. efficacy
ifNextFinal	FALSE
update.act.t1	0.514
update.act.t2	1
orig.Nmax	466
new.Nmax	411
current.info	1.01

The third column is the time at which each patient is enrolled. The column 'week' is recording the number of the longitudinal visits per person with '0' denoting baseline and 1–4 denoting post baseline visits. The column 'y' is the response of interest, and 'flag' distinguishes patients who are still continuing (1) or not (0). It is noticed that the second patient does not have all four post baseline measurements but he or she is not continuing, implying that this person dropped out at the last visit.

Once the first interim data has been collected ($N_{max} \times 1/2$), plugging in all the design parameters including α, β, δ, one-sided test or two-sided test, number of planned looks, planned information fraction at each look, and nuisance parameters assumption (R_0, S_0, and r_{10}) as well as available data, the *R* function employing the strategy introduced earlier generates the result in Table V (second section).

The first section of the table displays the parameters we defined at the design stage. The rows 'Planned.timing1' and 'Planned.timing2' are the planned information fraction at first and final interim. The number of planned looks is clearly the number of rows for these variables (two in this case). The second section of the table suggests that we should continue to do a second analysis because neither efficacy nor futility has been detected and that next analysis should be our final analysis according to the adaptation rule. The rows 'update.act.t1' and 'update.act.t2' are the actual information fraction as updated in (10), 'orig.Nmax' and 'new.Nmax' are N_{max} calculated from the design and re-estimated at

the first interim, respectively. The decrease of maximum sample size is because the anticipated information fraction (0.5) is smaller than the actual 0.514 that is produced in the bottom of the section. The underlying reason is that the nuisance parameter assumptions deviate from the truth, and as a result, the estimated covariance or information estimated by the real data does not agree with that using originally assumed nuisance parameters. Next, we collect another 178 ($= 411 - 466 \times \frac{1}{2}$) patients and analyze the final look. At the second look, we need to add the actual information fraction vector (0.514, 1) and the new N_{max} (411) into our function. The result in the bottom section of Table V shows that the study can be stopped for efficacy and current information fraction is 1.01. The function and its help files are available upon request.

6. Discussion

We presented two information-based group sequential sample size re-estimation methods that can be for longitudinal trials that adapts appropriately depending on the true value of unknown nuisance parameters. Whereas previous work by Shih and Gould [6] and Zucker and Denne [7] only evaluates a single interim and no hypothesis testing is performed until the final analysis; our approach has advantages of early termination and multiple interim looks. The simulation results confirm the method maintains power while controlling the type I error, while a fixed design or a group sequential design without adjusting for nuisance parameters cannot. The reason is that in some cases where we do not have good historical evidence of the nuisance parameters, we have the ability of correcting it during the interim. In addition, a smaller sample is expected when the assumption is reasonably accurate; however, poor assumption requires more patients to maintain the statistical power. In conclusion, our method will help to both limit investment in treatments that do not work and ensure an appropriate investment to power trials for drugs that do work. For drugs that provide more than a minimally interesting treatment effect, the group sequential efficacy bounds provide a method to bring very effective drugs to market quickly.

We assume equally spaced information-based design and equal retention rate for control and active treatment group just for simplicity in explanation. It is, however, fairly easy to extend it to a general case. To perform a real data analysis by our methods, we have built functions to calculate the necessary sample size before starting the trial enrollment and to re-estimate the sample size with testing if stopping early at the same time. Although this is being carried out in an unblinded fashion, our method can certainly be used to recalculate the sample size and testing as long as the estimated parameter of interest and its variance are provided from the third party. Moreover, our methods work well for a small sample size as long as there are sufficient data to be analyzed in the random effect model at each interim look. However, given the complexity of the problem, it would be difficult to back-calculate the interim treatment effect based on the sample size adaptation as can be carried out in cases that are simpler than longitudinal data analysis.

Our methods presume that all subjects have measurement at their predefined measurement times. It is possible to introduce bias at the interim analyses if measurements occur at times other than the predefined follow-up times. Methods such as using a piece-wise linear approximation proposed by Kittelson *et al.* [10] may be incorporated for the future work to handle departure from the protocol-defined measurement times. Another extension of our work could be to loosen the assumption of monotone missingness to missing at random. Moreover, subjects who are still under active follow-up may be different from those who drop out. Methods for evaluation of sensitivity to informative dropouts is another potential topic. Gao *et al.* [19], Emerson and Fleming [20], and Kim [21] introduced methods for unbiased estimation following sequential testing, and these methods could be incorporated when reporting results from any group sequential trial.

References

1. Lu K, Luo X, Chen PY. Sample size estimation for repeated measures analysis in randomized clinical trials with missing data. *The International Journal of Biostatistics* 2008; **4**(1):1–16.
2. Lu K, Mehrotra DV, Liu G. Sample size determination for constrained longitudinal data analysis. *Statistics in Medicine* 2009; **28**(4):679–699.
3. Coffey CS, Kairalla JA. Adaptive clinical trials: progress and challenges. *Drugs in R&D* 2008; **9**(4):220–242.
4. Chuang-Stein C, Anderson K, Gallo P, Collins S. Sample size reestimation: a review and recommendations. *Drug Information Journal* 2006; **40**(4):475–484.
5. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* 1990; **9**(1-2):65–72.

6. Shih WJ, Gould AL. Re-evaluating design specifications of longitudinal clinical trials without unblinding when the key response is rate of change. *Statistics in Medicine* 1995; **14**(20):2239–2248.
7. Zucker DM, Denne J. Sample–size redetermination for repeated measures studies. *Biometrics* 2002; **58**(3):548–559.
8. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. CRC Press: Boca Raton, FL, 2000.
9. Galbraith S, Marschner IC. Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. *Statistics in Medicine* 2003; **22**(11):1787–1805.
10. Kittelson JM, Sharples K, Emerson SS. Group sequential clinical trials for longitudinal data with analyses using summary statistics. *Statistics in Medicine* 2005; **24**(16):2457–2475.
11. Burington BE, Emerson SS. Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics* 2003; **59**:770–777.
12. Mehta CR, Tsiatis AA. Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal* 2001; **35**(4):1095–1112.
13. Tsiatis AA. Information-based monitoring of clinical trials. *Statistics in Medicine* 2006; **25**(19):3236–3244.
14. Liang KY, Zeger SL. Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhyā: The Indian Journal of Statistics, Series B* 2000; **62**:134–148.
15. Scharfstein DO, Tsiatis AA, Robins JM. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association* 1997; **92**(440):1342–1350.
16. Zhu L, Ni L, Yao B. Group sequential methods and software applications. *The American Statistician* 2011; **65**(2):127–135.
17. Diggle P. *Analysis of Longitudinal Data*, Vol. 25. Oxford University Press: USA, 2002.
18. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**(3):659–663.
19. Gao P, Liu L, Mehta C. Exact inference for adaptive group sequential designs. *Statistics in Medicine* 2013; **32**(23): 3991–4005.
20. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika* 1990; **77**(4):875–892.
21. Kim K. Point estimation following group sequential tests. *Biometrics* 1989; **45**(2):613–617.