Biometrics 61, 362–369
June 2005

DOI: 10.1111/j.1541-0420.2005.00333.x

# Adaptive Design and Estimation in Randomized Clinical Trials with Correlated Observations

**Guosheng Yin*** **and Yu Shen**

Department of Biostatistics and Applied Mathematics, M. D. Anderson Cancer Center,
The University of Texas, Houston, Texas 77030, U.S.A.
**email:* gsyin@mdanderson.org

Summary. Clinical trial designs involving correlated data often arise in biomedical research. The intracluster correlation needs to be taken into account to ensure the validity of sample size and power calculations. In contrast to the fixed-sample designs, we propose a flexible trial design with adaptive monitoring and inference procedures. The total sample size is not predetermined, but adaptively reestimated using observed data via a systematic mechanism. The final inference is based on a weighted average of the block-wise test statistics using generalized estimating equations, where the weight for each block depends on cumulated data from the ongoing trial. When there are no significant treatment effects, the devised stopping rule allows for early termination of the trial and acceptance of the null hypothesis. The proposed design updates information regarding both the effect size and within-cluster correlation based on the cumulated data in order to achieve a desired power. Estimation of the parameter of interest and its confidence interval are proposed. We conduct simulation studies to examine the operating characteristics and illustrate the proposed method with an example.

Key words: Correlated data; Generalized estimating equation; Hypothesis testing; Power; Sample size; Self-designing trial.

## 1. Introduction

In biomedical applications, we often encounter longitudinal data measured over time for each patient or clustered outcomes. For example, blood pressures measured repeatedly on one patient, and clinical trials studying the eyes, ears, or teeth naturally involve dependent data. Methodological development for analyzing correlated data has been greatly advanced (Diggle et al., 2002), and clinical trial designs involving sample size and power calculations have also been studied for correlated outcomes (e.g., Vonesh and Schork, 1986; Rochon, 1991; Liu and Liang, 1997; Liu, Shih, and Gehan, 2002), where all the design parameters including the effect size and correlation structure are assumed to be known.

For fixed-sample designs, one usually relies on previous studies or expert experience to determine the effect size and correlations among the outcomes. With uncertainty in multiple design parameters, it is often difficult to obtain a sample size that is sufficient to detect the expected treatment effect. One must consider the intracluster correlation in addition to the parameters required for independent observations when designing a trial. Misspecification of any design parameter may lead to a design with an insufficient power or an overestimation of the total sample size. For trials with correlated data, it is particularly appealing to update the initial design parameters using the ongoing trial data and reestimate the sample size at interim stages to ensure an adequate testing power. Extensive research in adaptive clinical trial designs has been recently carried out by Gould (1992), Shih (1992), Bauer and Köhne (1994), Proschan and Hunsberger (1995), Betensky and Tierney (1997), Fisher (1998), Wassmer (1998), Cui, Hung, and Wang (1999), Lehmacher and Wassmer (1999), Shen and Fisher (1999), Posch and Bauer (2000), Liu and Chi (2001), and Müller and Schäfer (2001), among others. However, the adaptive designs described in the aforementioned studies have been proposed for independent data. Limited attention has been paid to adaptive trial designs with correlated outcomes, because correlation often makes the design and adaptation much more complicated.

For the fixed-sample designs with correlated data, Liu and Liang (1997) extended the sample size and power calculations for generalized linear models (Self and Mauritsen, 1988), and developed a unified framework for experimental designs involving correlated observations based on the generalized estimating equations (GEE; Liang and Zeger, 1986). Zucker and Denne (2002) and Lake et al. (2002) studied two-stage adaptive designs with correlated data, where the sample size is reestimated at one interim point based on an internal pilot study. In contrast, without prespecifying the maximum number of interim analyses, we present a flexible design and inference procedure for correlated observations in the GEE framework. The proposed method is applicable in a wide variety of situations including discrete and continuous correlated data using regression models. A working correlation structure needs to be specified in GEE while the values in the correlation matrix are updated inherently. In an adaptive fashion, we update the sample size given the observed data in the

ongoing trial to ensure an adequate power, as well as to maintain the type I error rate under a desirable level. Throughout, the sample size refers to the number of clusters instead of total observations.

## 2. Design and Inference Strategies

### 2.1 *Generalized Estimating Equations*

Let $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ be independent vectors of response variables with means $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n)$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iL_i})'$ for the $i$th cluster, $i = 1, \ldots, n$. Let $\mu_{il} = E(y_{il})$, and $\text{var}(y_{il}) = \zeta h(\mu_{il})$, $l = 1, \ldots, L_i$, where $h(\cdot)$ is the variance function and $\zeta$ is the scale parameter. A function $g(\cdot)$ links the mean $\mu_{il}$ with the covariate vectors $\mathbf{z}_{il}$ and $\mathbf{x}_{il}$ through

$$g(\mu_{il}) = \boldsymbol{\phi}'\mathbf{z}_{il} + \boldsymbol{\psi}'\mathbf{x}_{il}, \tag{1}$$

where $\boldsymbol{\phi}$, a $p \times 1$ vector, represents the parameters of interest, and $\boldsymbol{\psi}$, a $q \times 1$ vector, is nuisance. The hypotheses are $\text{H}_0$: $\boldsymbol{\phi} = \mathbf{0}$ versus $\text{H}_1$: $\boldsymbol{\phi} = \boldsymbol{\delta}$, where $\boldsymbol{\delta} > \mathbf{0}$ is the vector of the assumed treatment effects. Inference for $\boldsymbol{\phi}$ based on a quasi-score or Wald test statistic has been well investigated and shown to have desirable properties (Diggle et al., 2002).

Let $\boldsymbol{\theta} = (\boldsymbol{\phi}', \boldsymbol{\psi}')'$, and let $\mathbf{C}_i$ denote the working correlation matrix which may not be identical to the true correlation matrix. The generalized estimating equations are given by

$$\sum_{i=1}^{n} \mathbf{D}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \tag{2}$$

where $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\theta}$, $\mathbf{V}_i = \mathbf{G}_i\mathbf{C}_i\mathbf{G}_i$, and $\mathbf{G}_i = \text{diag}\,\{[h(\boldsymbol{\mu}_i)]^{1/2}\}$. Under certain regularity conditions, the estimator $\hat{\boldsymbol{\theta}}$ obtained from (2) is consistent and asymptotically Gaussian, that is, as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \longrightarrow N(\mathbf{0}, \boldsymbol{\Sigma}_\theta), \qquad \boldsymbol{\Sigma}_\theta = \lim_{n\to\infty} n\mathbf{A}_1^{-1}\mathbf{A}_2\mathbf{A}_1^{-1}, \tag{3}$$

where

$$\mathbf{A}_1 = \sum_{i=1}^{n} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i, \quad \mathbf{A}_2 = \sum_{i=1}^{n} \mathbf{D}_i'\mathbf{V}_i^{-1}\Gamma_i\mathbf{V}_i^{-1}\mathbf{D}_i, \text{ and}$$

$$\boldsymbol{\Gamma}_i = \text{var}(\mathbf{y}_i).$$

The consistent variance estimator of $\boldsymbol{\Sigma}_\theta$ is obtained by evaluating the matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ at their empirical estimates and replacing $\boldsymbol{\Gamma}_i$ in $\mathbf{A}_2$ by $(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)'$, which is referred to as the sandwich variance estimator.

The sample size calculation for correlated data is often complicated even for the fixed-sample design (Liu and Liang, 1997). In addition to the marginal modeling structure, the working correlation matrix, and the type I ($\alpha$) and II ($\beta$) error rates, one must specify all the design parameters, including nuisances in the models under the null and alternative hypotheses. Moreover, the design relies on the assumed distributions for the configurations on discretized covariates of $(\mathbf{z}_{il}, \mathbf{x}_{il})$. Conversely, with the generalized self-designing trial, one can adaptively derive the sample size based on the specified $\alpha$ and $\beta$, the expected treatment effects (only used for futility stopping), and the marginal regression model, while the treatment effects and intracluster correlations can be updated sequentially.

### 2.2 *Adaptive Design and Test Statistics*

In a self-designing trial (Shen and Fisher, 1999), the data are reviewed periodically after observing every $B_j$ clusters of observations, where $B_j$ is a prefixed block size for the $j$th stage, and $j = 1, 2, \ldots$ As opposed to the fixed-sample design, the total sample size is adaptively determined through the cumulated data so that a desired power may be achieved.

It is important to terminate the trial early for ethical and economic reasons if the cumulated information shows that the new treatment is ineffective or inferior to the standard one. If the trial is not terminated at the $(j-1)$th stage, we continue to observe the next block of data and estimate a weight function $w_j$ using data up to the $(j-1)$th stage. The strategy will be iterated until the weight function is used up at a certain step, $m$, when $\sum_{j=1}^{m-1} w_j^2 < 1$ and $\sum_{j=1}^{m} w_j^2 \geq 1$, when the conditional power based on the cumulated data is greater than or equal to $(1-\beta)$, or when the cumulated data indicate the treatment to be ineffective based on a futility stopping rule introduced in the next section. We then exit the trial and let the weight at the final stage be $w_m = (1 - \sum_{j=1}^{m-1} w_j^2)^{1/2}$. Thus, the total number of blocks, $m$, is not prespecified but is a finite random integer depending on the data observed prior to the $m$th block.

Let $\mathcal{D}_j$ denote the cumulated data up to step $j$. The corresponding information at that time can be defined by $\sigma$-algebra, $\mathcal{F}_j = \sigma(\mathcal{D}_j)$. To construct a final test statistic, we derive a Wald-type statistic for each block of data,

$$\mathbf{U}_j = \sqrt{B_j}\hat{\boldsymbol{\Sigma}}_{\phi_j}^{-\frac{1}{2}}\hat{\boldsymbol{\phi}}_j,$$

where $\hat{\boldsymbol{\phi}}_j$ and $\hat{\boldsymbol{\Sigma}}_{\phi_j}$ are the consistent estimators of $\boldsymbol{\phi}$ and $\boldsymbol{\Sigma}_\phi$, using the data only from the $j$th block. One referee suggested a preferable alternative to estimate $\hat{\boldsymbol{\Sigma}}_{\phi_j}$ based on all the cumulated data up to stage $j$, for which the asymptotic properties can be justified by Slutsky's theorem. Under $\text{H}_0$, $\mathbf{U}_j$ asymptotically follows a standard $p$-dimensional normal distribution, $N_p(\mathbf{0}, \mathbf{I})$.

We can show that the weighted average of the blockwise statistics, $\mathbf{T}_m = \sum_{j=1}^{m} w_j\mathbf{U}_j$, asymptotically converges to $N_p(\mathbf{0}, \mathbf{I})$ under $\text{H}_0$, where $\sum_{j=1}^{m} w_j^2 = 1$ and $w_j = w_j(\mathcal{D}_{j-1})$. Under $\text{H}_1$, it is much more difficult to obtain the distribution of $\mathbf{T}_m$, which usually does not follow a multivariate normal distribution. However, the *pivot*, which is a function of $\mathbf{T}_m$,

$$\mathbf{P}_m = \sum_{j=1}^{m} w_j\mathbf{Q}_j, \quad \text{with} \quad \mathbf{Q}_j = \sqrt{B_j}\hat{\boldsymbol{\Sigma}}_{\phi_j}^{-\frac{1}{2}}(\hat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}),$$

asymptotically follows a normal distribution under both the null and the alternative hypotheses (Cheng and Shen, 2004). The proof is outlined in the Appendix.

Testing for efficacy is only performed at the final stage, based on the fact that $\mathbf{T}_m'\,\mathbf{T}_m$ asymptotically follows a central chi-square distribution with $p$ degrees of freedom, $\chi_p^2$, under the null hypothesis. The validity of the test statistic $\mathbf{T}_m'\mathbf{T}_m$ depends on the condition of $\sum_{j=1}^{m} w_j^2 = 1$. If $\boldsymbol{\phi}$ is scalar (i.e., $p = 1$), the final test statistic, $T_m$, reduces to a statistic having a standard normal distribution under $\text{H}_0$.

Following rejection or acceptance of the null hypothesis, it is of interest to estimate the unknown parameter $\boldsymbol{\phi}$. Estimation of $\boldsymbol{\phi}$ and its confidence interval can be derived from the

asymptotic property of $\mathbf{P}_m$. Let $c_{p,1-2\alpha}$ denote the $(1-2\alpha)$ percentile of $\chi^2_p$. The elliptical confidence set is then defined by $\Pr(\mathbf{P}'_m\mathbf{P}_m \le c_{p,1-2\alpha}) = 1-2\alpha$, which may not have an explicit closed form. For ease of exposition, let $p=1$ (i.e., we consider the special case with one overall treatment effect). The $100(1-2\alpha)\%$ confidence interval of $\phi$ is constructed as follows:

$$
\begin{aligned}
1-2\alpha &= \Pr\left\{-z_{1-\alpha} \le \sum_{j=1}^{m} w_j\sqrt{B_j}\hat{\Sigma}_{\phi_j}^{-\frac{1}{2}}(\hat{\phi}_j-\phi) \le z_{1-\alpha}\right\} \\
&= \Pr\left(\frac{\sum_{j=1}^{m} w_j\sqrt{B_j}\hat{\Sigma}_{\phi_j}^{-\frac{1}{2}}\hat{\phi}_j - z_{1-\alpha}}{\sum_{j=1}^{m} w_j\sqrt{B_j}\hat{\Sigma}_{\phi_j}^{-\frac{1}{2}}} \le \phi \right. \\
&\qquad \left. \le \frac{\sum_{j=1}^{m} w_j\sqrt{B_j}\hat{\Sigma}_{\phi_j}^{-\frac{1}{2}}\hat{\phi}_j + z_{1-\alpha}}{\sum_{j=1}^{m} w_j\sqrt{B_j}\hat{\Sigma}_{\phi_j}^{-\frac{1}{2}}}\right),
\end{aligned}
$$

where $z_{1-\alpha}$ is the $(1-\alpha)$ percentile of the standard normal distribution. It is known that the naive estimator of $\phi$, obtained by treating the self-designing trial as a fixed-sample design, is often biased. Based on the moment estimation, a consistent estimator of $\phi$ with reduced bias is given by

$$
\hat{\phi} = \frac{\sum_{j=1}^{m} w_j\sqrt{B_j}\hat{\Sigma}_{\phi_j}^{-\frac{1}{2}}\hat{\phi}_j}{\sum_{j=1}^{m} w_j\sqrt{B_j}\hat{\Sigma}_{\phi_j}^{-\frac{1}{2}}}. \tag{4}
$$

To ensure the validity of the estimation procedure, one more block of data should be observed after the cumulated data indicate that the trial should terminate due to futility, or that the conditional power exceeds the specified level. If we allow the trial to stop at the $j$th step and accept $H_0$ due to futility, the sum of the weights up to step $j$ is most likely less than 1. In this case, the pivot $\mathbf{P}_m(m=j)$ does not have a standard multivariate normal distribution, and then the construction of the confidence interval for $\boldsymbol{\phi}$ via $\mathbf{P}_m$ is invalid. Therefore, it is important to take one more block of data to spend the remaining weight. Moreover, it is worth emphasizing that the decision to accept $H_0$ should not be altered when the trial is stopped early due to futility, even though another block of data is observed at the last stage for the purpose of estimation.

### 2.3 *Futility Stopping Rule*

The futility stopping rule is outlined as follows. We apply the GEE method to data cumulated up to the $j$th step to obtain the consistent estimators of $\boldsymbol{\phi}$ and $\boldsymbol{\Sigma}_\phi$, denoted by $\hat{\boldsymbol{\phi}}^{(j)}$ and $\hat{\boldsymbol{\Sigma}}_\phi^{(j)}$, where $\boldsymbol{\Sigma}_\phi$ is the submatrix corresponding to $\boldsymbol{\phi}$ of the covariance matrix $\boldsymbol{\Sigma}_\theta$ in (3). We can estimate the confidence ellipsoid of vector $\boldsymbol{\phi}$ at stage $j, \mathcal{E}_j$, having the limiting confidence coefficient $(1-\alpha_f)$ based on the following asymptotic distribution,

$$
n_j\big(\hat{\boldsymbol{\phi}}^{(j)}-\boldsymbol{\phi}\big)'\big(\hat{\boldsymbol{\Sigma}}_\phi^{(j)}\big)^{-1}\big(\hat{\boldsymbol{\phi}}^{(j)}-\boldsymbol{\phi}\big) \xrightarrow{d} \chi^2_p,
$$

where $n_j$ is the cumulated sample size up to the $j$th step. Specifically,

$$
\begin{aligned}
\Pr(\boldsymbol{\phi}\in\mathcal{E}_j) = \Pr\Big\{ & n_j\big(\hat{\boldsymbol{\phi}}^{(j)}-\boldsymbol{\phi}\big)'\big(\hat{\boldsymbol{\Sigma}}_\phi^{(j)}\big)^{-1}\big(\hat{\boldsymbol{\phi}}^{(j)}-\boldsymbol{\phi}\big) \\
& \le c_{p,1-\alpha_f}\Big\} = 1-\alpha_f.
\end{aligned} \tag{5}
$$

At the $j$th step, if the assumed design parameter vector, $\boldsymbol{\delta}$, is not contained in $\mathcal{E}_j$, we then step down to estimate the confidence interval with the limiting confidence coefficient $(1-\alpha_f)$ for each $\phi_k(k=1,\ldots,p)$ marginally, $[\hat{\phi}^{(j)}_{L,k}, \hat{\phi}^{(j)}_{U,k}]$. If there exists a $k$ such that $\hat{\phi}^{(j)}_{U,k} < \delta_k$, we stop the trial at step $j$ for insufficient beneficial treatment effects. As long as the design parameter corresponding to one endpoint is out of the range of the confidence region marginally, we terminate the trial due to futility. Note that this futility stopping rule is a conservative one to stop a futile trial with multiple endpoints. An alternative rule may be proposed to terminate a trial for futility when all components of $\boldsymbol{\delta}$ lie outside of their confidence intervals marginally. Indeed, the futility stopping rule is not unique and can be specified according to the actual problem in practice and elicited from the study investigators. For this type of adaptive designs, it is not only desirable but necessary to have a futility stopping rule for ethical and regulatory reasons.

### 2.4 *Constructing Weight Functions*

An efficient self-designing trial relies on a sensitive weighting scheme, which plays a critical role in determining when to extend or terminate the trial. At the design stage, we would like to use the observed data from the ongoing trial to adaptively determine the weight for the next block. An inverse function of the conditional sample size can be a sensible choice for the weight function so that the trial can be terminated soon after a strong treatment effect is revealed from the cumulated data. Based on the conditional power derivation (Lan, Simon, and Halperin, 1982), we obtain the additional sample size $N^*_j$ for step $j$ when assuming that $j$ is the last step, to ensure the power of $(1-\beta)$ given the observed data up to step $(j-1)$. The purpose of estimating $N^*_j$ is to calculate the weight for the next block of data, as opposed to using it directly for determining the sample size needed for the next step. In fact, the actual block size of each step is pre-fixed.

The conditional sample size $N^*_j$ can be obtained from the proposed chi-square test statistic,

$$
\begin{aligned}
\Pr\Bigg\{ & \left(\sum_{i=1}^{j-1} w_i\mathbf{U}_i + w_j\mathbf{U}^*_j\right)' \left(\sum_{i=1}^{j-1} w_i\mathbf{U}_i + w_j\mathbf{U}^*_j\right) \\
& > c_{p,1-2\alpha}\Big|\mathcal{F}_{j-1}\Bigg\} = 1-\beta,
\end{aligned} \tag{6}
$$

where $\mathbf{U}^*_j = \sqrt{N^*_j}\boldsymbol{\Sigma}_\phi^{-\frac{1}{2}}\hat{\boldsymbol{\phi}}^*_j$ and $\hat{\boldsymbol{\phi}}^*_j$ is the statistic based on $N^*_j$ observations. Note that $w_j = (1-\sum_{i=1}^{j-1} w_i^2)^{1/2}$ in the derivation, when we assume that the $j$th step is the last step of the trial. Given the data observed up to the $(j-1)$th step,

$\mathbf{U}_j^* + w_j^{-1}\sum_{i=1}^{j-1} w_i\mathbf{U}_i$ approximates to a $p$-dimensional normal distribution with mean

$$\boldsymbol{\gamma}_j = \sqrt{N_j^*}\boldsymbol{\Sigma}_\phi^{-\frac{1}{2}}\boldsymbol{\phi} + w_j^{-1}\sum_{i=1}^{j-1} w_i\mathbf{U}_i,$$

and an identity covariance matrix, thus the quadratic form $(\mathbf{U}_j^* + w_j^{-1}\sum_{i=1}^{j-1} w_i\mathbf{U}_i)'(\mathbf{U}_j^* + w_j^{-1}\sum_{i=1}^{j-1} w_i\mathbf{U}_i)$ has a noncentral chi-square distribution with $p$ degrees of freedom. The noncentrality parameter is $\boldsymbol{\gamma}_j{}'\,\boldsymbol{\gamma}_j$, which is a function of $N_j^*$. To estimate the conditional sample size $N_j^*$, we can solve the following equation numerically:

$$c_{p,1-2\alpha} = \left(1 - \sum_{i=1}^{j-1} w_i^2\right) c_{p,1-\beta}(\boldsymbol{\gamma}_j'\boldsymbol{\gamma}_j), \tag{7}$$

where $c_{p,1-\beta}(\boldsymbol{\gamma}_j'\ \boldsymbol{\gamma}_j)$ is the $(1-\beta)$ percentile for a noncentral chi-square distribution with $p$ degrees of freedom. The unknown parameters of $\boldsymbol{\phi}$ and $\boldsymbol{\Sigma}_\phi$ in $\boldsymbol{\gamma}_j$ can be replaced by their consistent estimators using the cumulated observations up to the $(j-1)$th step.

To illustrate the procedure, we consider an important special case in which the repeated measurements follow a generalized linear regression model,

$$g(\mu_{il}) = \phi z_{il} + \psi x_{il}, \qquad i = 1,\ldots,n; \quad l = 1,\ldots,L_i,$$

where the scalar $\phi$ represents the treatment effect. For a typical two-sample problem with repeated measurements, it is of interest to test the hypothesis that $H_0$: $\phi = 0$, for an overall treatment effect. In contrast to equation (7), an explicit conditional sample size formula can be obtained for this case. Specifically, we can solve the conditional power equation using the Gaussian approximation for $U_j^*$,

$$N_j^* = \left(\frac{z_{1-\alpha} - \sum_{i=1}^{j-1} w_i U_i}{\sqrt{1 - \sum_{i=1}^{j-1} w_i^2}} + z_{1-\beta}\right)^2 \frac{\hat{\Sigma}_\phi^{(j-1)}}{(\hat{\phi}^{(j-1)})^2}.$$

It is clear that the conditional sample size $N_j^*$ will be small, if the cumulated data indicate a strong treatment effect. In this case, it is reasonable to assign a relatively large weight to the next block of data, because we want the trial to terminate soon. Toward this goal, the weight function for the $j$th block should be inversely associated with $N_j^*$, and thus a natural form is given by

$$w_j = \left\{\frac{B_j}{N_j^*}\left(1 - \sum_{i=1}^{j-1} w_i^2\right)\right\}^{1/2}, \quad j = 2,\ldots,m-1,$$

and for the last step, $w_m = (1 - \sum_{i=1}^{m-1} w_i^2)^{1/2}$.

## 3. Simulation Studies

We carried out simulations to compare the proposed adaptive design and the fixed-sample design with respect to the type I error rate, statistical power, average sample number (ASN), and estimation of the parameter of interest. We considered linear models with multivariate Gaussian errors, and logistic regression with correlated binary data, and also conducted a sensitivity analysis on the misspecification of the working correlation matrix in the GEE. We replicated 5000 trials for each setup.

### 3.1 *Repeated Measurements with Gaussian Errors*

The linear regression model with repeated measurements is common in longitudinal studies, where a typical example is to compare two treatment groups with repeated measurements. For clarity and ease of exposition, suppose that all the clusters are of the same size, that is, $L_i \equiv L$. Let $p = q = 1$ with $x_{il} = 1$ corresponding to the intercept, and a treatment indicator $z_{il} \equiv z_i = 1$ or 0 with probability 0.5. The linear regression model for repeated measurements is given by

$$y_{il} = \psi + \phi z_i + \epsilon_{il}, \quad i = 1,\ldots,n; \quad l = 1,\ldots,L, \tag{8}$$

where $\boldsymbol{\epsilon}_i = (\epsilon_{i1},\ldots,\epsilon_{iL})'$ is assumed to follow a multivariate normal distribution with mean 0 and covariance $\sigma^2\mathbf{R}$. With an exchangeable correlation matrix $\mathbf{R} = (1-\tilde{\rho})\mathbf{I} + \tilde{\rho}\mathbf{1}\mathbf{1}'$, a type I error rate of $\alpha$ and a power of $(1-\beta)$, the sample size formula for the fixed-sample design (Liu and Liang, 1997) is

$$K = \frac{4(z_{1-\alpha} + z_{1-\beta})^2\sigma^2\{1 + (L-1)\tilde{\rho}\}}{\delta^2 L}.$$

Note that $\delta$ and $\tilde{\rho}$ are the design parameters from previous studies or expert opinions, which may be misspecified.

We set $\alpha = 0.025$, $\beta = 0.1$, the cluster size $L = 2$, underlying model parameters $\sigma^2 = 1$, $\psi = 1$, $\rho = 0.3$, $H_0$: $\phi = 0$, and $H_1$: $\phi = 0.5$. We took a constant block size with $B = 15$ or 20. The size of the first block is usually relatively large in order to obtain a more reliable result at the initial step, for example, $B_1 = 2B$. The corresponding weight at the first block should not depend on the data; we took $w_1 = 0.4$. In the GEE procedure, we chose the working correlation matrix to be exchangeable in Tables 1 and 2. We used $\alpha_f = 0.01$ for the futility stopping rule in (5).

Three different scenarios were examined: $\delta$ taking the true value ($\delta = \phi$) and varying $\tilde{\rho} = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$; fixing $\tilde{\rho}$ at the true value ($\tilde{\rho} = \rho$) and varying $\delta$ = (0.4, 0.5, 0.6, 0.7); and varying both $\delta$ and $\tilde{\rho}$. These various configurations gave us an opportunity to examine how sample size, type I error rate, and power changed with respect to one design parameter while the other was fixed, and also when all the design parameters were misspecified.

The evaluation of the test size based on data generated under $H_0$ is given in Table 1. We see that the proposed self-designing trial is able to preserve the type I error rate in all the cases, while the fixed-sample design shows slight inflation in some scenarios. Because of the futility stopping rule, the type I error rate is in fact deflated for the proposed method. This stopping rule is somewhat conservative, and may be improved in the future. No substantial difference is observed in terms of the average sample number (ASN) between the self-designing trial and the fixed-sample design. Apparently, the estimator $\hat{\phi}$ using (4) is less biased compared to $\hat{\phi}_{\text{naive}}$ when naively treating the adaptive design as the fixed-sample design. The 95% confidence interval coverage rates are slightly lower than the nominal level. This undercoverage phenomenon could be caused by the variation due to the relatively small sample size in each block.

Table 2 summarizes the comparisons of the ASN and power between the fixed-sample design and the self-designing trials.

**Table 1**
*Simulation results under* $H_0$: $\phi = 0, \rho = 0.3$, *with one-sided* $\alpha = 0.025, \beta = 0.1$, *and the Gaussian error*

| | | Fixed design | | Self-designing ($B = 15$) | | | Self-designing ($B = 20$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\delta$ | $\tilde{\rho}$ | Size | $K$ | Size | ASN | n.b. | Size | ASN | n.b. | $\hat{\phi}_{\text{naive}}$ | $\hat{\phi}$ | c.r. |
| 0.4 | 0.3 | 0.0288 | 171 | 0.0158 | 126.1 | 7.4 | 0.0098 | 138.1 | 5.9 | −0.054 | −0.017 | 0.921 |
| 0.5 | 0.3 | 0.0264 | 110 | 0.0164 | 91.5 | 5.1 | 0.0088 | 104.0 | 4.2 | −0.046 | −0.012 | 0.914 |
| 0.6 | 0.3 | 0.0322 | 76 | 0.0138 | 73.0 | 3.9 | 0.0090 | 84.6 | 3.2 | −0.036 | −0.010 | 0.918 |
| 0.5 | 0 | 0.0234 | 85 | 0.0174 | 91.6 | 5.1 | 0.0120 | 103.0 | 4.2 | −0.045 | −0.013 | 0.915 |
| 0.5 | 0.1 | 0.0296 | 93 | 0.0150 | 91.4 | 5.1 | 0.0112 | 103.7 | 4.2 | −0.045 | −0.012 | 0.917 |
| 0.5 | 0.2 | 0.0254 | 101 | 0.0116 | 91.4 | 5.1 | 0.0136 | 104.4 | 4.2 | −0.042 | −0.008 | 0.913 |
| 0.5 | 0.4 | 0.0248 | 118 | 0.0108 | 91.7 | 5.1 | 0.0110 | 103.1 | 4.2 | −0.046 | −0.011 | 0.915 |
| 0.5 | 0.5 | 0.0306 | 127 | 0.0148 | 91.2 | 5.1 | 0.0110 | 102.6 | 4.1 | −0.049 | −0.015 | 0.918 |

$B$ is the block size, size corresponds to the type I error rate, $K$ is the fixed-sample size, ASN is the average sample number, n.b. is the average number of blocks, and c.r. is 95% confidence interval coverage rate.

**Table 2**
*Simulation results under* $H_1$: $\phi = 0.5$, *and* $\rho = 0.3$ *with one-sided* $\alpha = 0.025, \beta = 0.1$, *and the Gaussian error*

| | | Fixed design | | Self-designing ($B = 15$) | | | Self-designing ($B = 20$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\delta$ | $\tilde{\rho}$ | Power | $K$ | Power | ASN | n.b. | Power | ASN | n.b. | $\hat{\phi}_{\text{naive}}$ | $\hat{\phi}$ | c.r. |
| 0.5 | 0.3 | 0.8976 | 110 | 0.9032 | 105.8 | 6.0 | 0.9138 | 117.7 | 4.9 | 0.542 | 0.518 | 0.922 |
| 0.6 | 0.3 | 0.7772 | 76 | 0.8376 | 97.6 | 5.5 | 0.8570 | 109.0 | 4.4 | 0.536 | 0.522 | 0.917 |
| 0.7 | 0.3 | 0.6434 | 56 | 0.7246 | 87.0 | 4.8 | 0.7484 | 97.5 | 3.9 | 0.527 | 0.517 | 0.915 |
| 0.5 | 0 | 0.8146 | 85 | 0.8918 | 105.5 | 6.0 | 0.9094 | 115.9 | 4.8 | 0.541 | 0.519 | 0.923 |
| 0.5 | 0.1 | 0.8504 | 93 | 0.8954 | 104.7 | 6.0 | 0.9054 | 117.8 | 4.9 | 0.544 | 0.520 | 0.918 |
| 0.5 | 0.2 | 0.8764 | 101 | 0.8920 | 104.8 | 6.0 | 0.9058 | 114.9 | 4.7 | 0.545 | 0.521 | 0.924 |
| 0.5 | 0.4 | 0.9206 | 118 | 0.9028 | 104.9 | 6.0 | 0.9110 | 117.8 | 4.9 | 0.543 | 0.522 | 0.911 |
| 0.5 | 0.5 | 0.9394 | 127 | 0.8862 | 105.6 | 6.0 | 0.9148 | 116.5 | 4.8 | 0.544 | 0.521 | 0.916 |
| 0.6 | 0 | 0.6772 | 59 | 0.8342 | 97.4 | 5.5 | 0.8484 | 108.0 | 4.4 | 0.537 | 0.520 | 0.916 |
| 0.6 | 0.1 | 0.7074 | 65 | 0.8272 | 96.9 | 5.5 | 0.8428 | 109.1 | 4.5 | 0.533 | 0.516 | 0.915 |
| 0.6 | 0.2 | 0.7496 | 71 | 0.8294 | 96.6 | 5.4 | 0.8528 | 109.4 | 4.5 | 0.533 | 0.515 | 0.923 |
| 0.7 | 0 | 0.5330 | 43 | 0.7272 | 85.8 | 4.7 | 0.7458 | 97.4 | 3.9 | 0.522 | 0.512 | 0.913 |
| 0.7 | 0.1 | 0.5776 | 48 | 0.7238 | 86.2 | 4.7 | 0.7472 | 98.3 | 3.9 | 0.524 | 0.516 | 0.921 |
| 0.7 | 0.2 | 0.6270 | 52 | 0.7244 | 86.6 | 4.8 | 0.7542 | 98.6 | 3.9 | 0.527 | 0.514 | 0.915 |

$B$ is the block size, $K$ is the fixed-sample size, ASN is the average sample number, n.b. is the average number of blocks, and c.r. is 95% confidence interval coverage rate.

When the design parameters $\delta$ and $\tilde{\rho}$ are specified correctly (the first row), both the fixed and the adaptive designs can achieve 90% power, and the corresponding ASNs are very close. However, when $\delta$ overestimates $\phi$, or $\tilde{\rho}$ underestimates $\rho$, the powers based on the self-designing procedure are much higher than those of the fixed-sample design. As expected, the estimated sample size for the fixed-sample design decreases as $\delta$ increasingly deviates away from $\phi$, and increases as $\tilde{\rho}$ becomes greater than $\rho$. Figure 1 shows the pattern of the weight function assigned to each step for 100 randomly replicated trials and the histogram of the number of blocks, using $B = 20$ under model (8). It is clear that most of the trials are terminated after the second or third interim analysis, and only a few trials continue beyond the eighth step.

The working correlation structure in the GEE might be specified incorrectly in the design stage. Following this route, we studied the sensitivity of the proposed method by specifying the "independence" or "unstructured" working correlation matrix while the true correlation was exchangeable. Table 3 shows that the adaptive procedure is very robust when the working matrix is not the same as the true one, under either $H_0$ or $H_1$. There are no notable differences in terms of the ASN and power between using the true and the misspecified working correlation matrices.

### 3.2 *Logistic Regression with Clustered Data*

Correlated binary responses are often encountered in biomedical research. Considering a two-group comparison with $L = 2$ and a treatment covariate $z_{il} \equiv z_i = 0$ or 1 with probability 0.5, via a logit link function,

$$\text{logit}(\mu_{il}) = \psi + \phi z_i.$$

The response probabilities corresponding to $z_i = 0$ and $z_i = 1$ are given by

$$p_0 = \frac{\exp(\psi)}{1 + \exp(\psi)}, \qquad p_1 = \frac{\exp(\psi + \phi)}{1 + \exp(\psi + \phi)}.$$

Testing $p_0 = p_1$ is equivalent to testing $H_0$: $\phi = 0$. With an exchangeable correlation matrix, the required sample size for

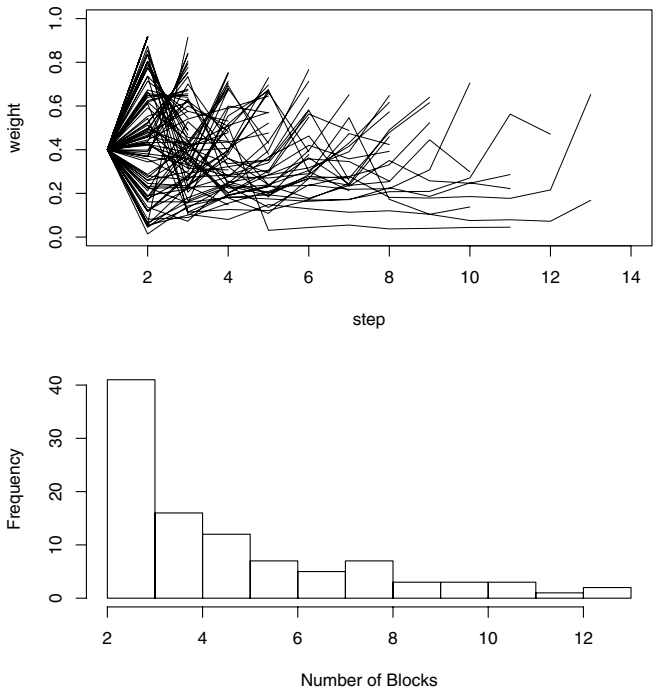


**Figure 1.** Weight functions varying over steps, and the histogram of the number of blocks under $H_1$: $\phi = 0.5$, $\rho = 0.3$, $\delta = 0.5$ with the Gaussian errors in 100 simulations.

the fixed-sample design is

$$K = \frac{4(z_{1-\alpha} + z_{1-\beta})^2\{p_0(1-p_0)/2 + p_1(1-p_1)/2\}\{1 + (L-1)\tilde{\rho}\}}{(p_1 - p_0)^2 L}. \quad (9)$$

It is not trivial to generate correlated binary data. Much research has been conducted in efficiently simulating dependent binary variates (Qaqish, 2003). Based on the fact that any Poisson random variable can be expressed as a convolution of other independent Poisson random variables, we generated dependent binary data from correlated Poisson variables (Park, Park, and Shin, 1996). In the logistic regression model, we assumed an independence working correlation, and set $\rho = 0.3$, $\psi = -0.2$, $H_0$: $\phi = 0$, and $H_1$: $\phi = 1$. We examined the following scenarios under $H_0$: fixing $\tilde{\rho} = \rho = 0.3$ and taking $\delta =$ (0.8, 1, 1.2); and fixing $\delta = 1$ and taking $\tilde{\rho} = (0, 0.1, 0.2, 0.4)$. As shown in Table 4, the test sizes under the self-designing trials are well preserved under the nominal level of $\alpha = 0.025$.

Under $H_1$, we similarly examined different configurations by varying $\delta$ or/and $\tilde{\rho}$, as indicated in Table 4. When design parameters moderately deviate from the true values, the adaptive design still maintains a power above 85%, whereas the fixed-sample design only achieves a power of 65–70%. As

**Table 4**

*Simulation results under $H_0$: $\phi = 0$ versus $H_1 : \phi = 1$, and $\rho = 0.3$ with one-sided $\alpha = 0.025, \beta = 0.1$, and $B = 20$ with logistic regression*

| | | Fixed design | | Self-designing | | |
|---|---|---|---|---|---|---|
| $\delta$ | $\tilde{\rho}$ | Size/power | $K$ | Size/power | ASN | n.b. |
| | | Size under $H_0$ | | | | |
| 0.8 | 0.3 | 0.0248 | 171 | 0.0122 | 147.1 | 6.4 |
| 1 | 0.3 | 0.0226 | 110 | 0.0112 | 108.4 | 4.4 |
| 1.2 | 0.3 | 0.0248 | 77 | 0.0078 | 87.6 | 3.4 |
| 1 | 0 | 0.0248 | 85 | 0.0138 | 109.2 | 4.5 |
| 1 | 0.1 | 0.0282 | 93 | 0.0108 | 108.7 | 4.4 |
| 1 | 0.2 | 0.0220 | 102 | 0.0112 | 108.2 | 4.4 |
| 1 | 0.4 | 0.0246 | 119 | 0.0118 | 109.0 | 4.4 |
| | | Power under $H_1$ | | | | |
| 1 | 0.3 | 0.8842 | 110 | 0.9214 | 125.3 | 5.3 |
| 1.2 | 0.3 | 0.7446 | 77 | 0.8504 | 117.7 | 4.9 |
| 1.5 | 0.3 | 0.5640 | 51 | 0.7034 | 100.3 | 4.0 |
| 1 | 0 | 0.8008 | 85 | 0.9150 | 124.8 | 5.2 |
| 1 | 0.1 | 0.8376 | 93 | 0.9208 | 125.8 | 5.3 |
| 1 | 0.2 | 0.8600 | 102 | 0.9116 | 124.4 | 5.2 |
| 1.2 | 0 | 0.6442 | 60 | 0.8630 | 117.4 | 4.9 |
| 1.2 | 0.1 | 0.6760 | 66 | 0.8578 | 117.3 | 4.9 |
| 1.2 | 0.2 | 0.7122 | 71 | 0.8552 | 117.1 | 4.8 |
| 1.5 | 0 | 0.4700 | 39 | 0.7058 | 99.5 | 4.0 |
| 1.5 | 0.1 | 0.5108 | 43 | 0.7146 | 99.8 | 4.0 |
| 1.5 | 0.2 | 0.5418 | 47 | 0.7012 | 100.1 | 4.0 |

$K$ is the fixed-sample size, ASN is the average sample number, and n.b. is the average number of blocks.

**Table 3**

*Sensitivity analysis for misspecification of the correlation matrix under $H_0 : \phi = 0$ versus $H_1 : \phi = 0.5$, when the true exchangeable correlation is $\rho = 0.3$, with one-sided $\alpha = 0.025, \beta = 0.1$, $B = 20$, and the Gaussian error*

| | | | Fixed design | | Self-designing | | |
|---|---|---|---|---|---|---|---|
| $\delta$ | $\tilde{\rho}$ | Working correlation | Size/power | $K$ | Size/power | ASN | n.b. |
| | | | | Size under $H_0$ | | | |
| 0.5 | 0.3 | Exchangeable | 0.0264 | 110 | 0.0088 | 104.0 | 4.2 |
| | | Independence | 0.0258 | 110 | 0.0102 | 102.9 | 4.1 |
| | | Unstructured | 0.0254 | 110 | 0.0130 | 103.5 | 4.2 |
| | | | | Power under $H_1$ | | | |
| 0.5 | 0.3 | Exchangeable | 0.8976 | 110 | 0.9138 | 117.7 | 4.9 |
| | | Independence | 0.9026 | 110 | 0.9136 | 116.1 | 4.8 |
| | | Unstructured | 0.9032 | 110 | 0.9100 | 114.9 | 4.7 |

$K$ is the fixed-sample size, ASN is the average sample number, and n.b. is the average number of blocks.

an extreme case with $\delta = 1.5$ and $\tilde{\rho} = 0$, the fixed-sample design results in a power of 47%. Through updating the sample size, the proposed adaptive design shows great improvement in power, with an increase that is close to 25%.

## 4. Rodent Teratology Data

As an illustration, we applied the proposed method to a randomized rodent teratology experiment with correlated binary data (Lefkopoulou and Ryan, 1993). Hartsfield (1986) conducted a study to investigate the effects of in utero exposure to the anticonvulsant, phenytoin. The degree of skeletal maturity was the outcome of interest. Because no particular site was more important than the other sites, we chose to base our design and analysis on the binary outcome (the presence of ossification or not) in the forepaws of the examined offspring. If the fetus had at least one digit ossified, we recorded a response of 0, otherwise the response was 1. The litter size varied from 1 to 10 with an average size of 7. Nineteen litters (clusters) were under dosage-exposure while 17 served as the control. We randomized the order of the data, and then partitioned the data into five blocks. The first block contained 12 clusters with 6 from the exposed arm and 6 from the control. After the first block, we used the block of size $B = 6$ with three clusters for each arm. Because the observed data were not exactly evenly distributed between the two arms, the last block of data had four litters in the treatment arm and two litters in the control.

The hypothesis of interest was the following: $H_0$: $\phi = 0$ versus $H_1$: $\phi = \delta$, where $\phi$ was the treatment effect of the anticonvulsant phenytoin on the rate of ossification. The design parameters were specified as $\delta = 1.3$, for the one-sided $\alpha = 0.025$ and $\beta = 0.1$. With a total of 36 litters of average litter size 7, the study would have a power of 90% to detect the treatment difference $\delta$ at the significance level of 0.025, assuming an exchangeable correlation matrix with a common intralitter correlation of 0.32.

Applying the proposed adaptive design and analysis strategy, the interim analysis at each step did not direct the trial to stop early for futility. After observing the data up to the fourth block, the estimated weight function for the fifth block indicated that the trial should stop at the fifth block ($m = 5$), so that the rest of the weight should be used up. The Wald test statistics ($U_j$) for the five blocks of data were (0.88, 2.57, 1.57, 5.08, and $-0.61$), and the corresponding weights ($w_j$) were (0.40, 0.19, 0.33, 0.46, and 0.70). The final weighted test statistic was $T_5 = 3.24$, while that from the fixed-sample design was 3.02. Both test statistics asymptotically followed the standard normal distribution under $H_0$ and showed a significant dosage-exposure effect. Employing an independent working matrix in the adaptive GEE procedure also indicated that the trial should be terminated at the fifth step, and the resulting final test statistic had a value of 3.39.

## 5. Discussion

We have extended the self-designing method of Fisher (1998) and Shen and Fisher (1999) to multivariate cases. The investigated adaptive design and analysis procedure for correlated observations are flexible and general within the GEE framework. The total sample size is not pre-fixed but adaptively determined based on the accrued data while the trial is ongoing, in order to achieve a desirable power. The final test statistic is a weighted sum of the block-wise Wald test statistics, where the weight for each block is calculated based on the conditional sample size. We have proposed an estimator for the parameter of interest with reduced bias and have constructed its confidence interval. The comparisons between the adaptive and the conventional fixed-sample designs suggest that adaptation be necessary when the initial estimates of design parameters are not reliable or are misspecified.

The proposed adaptive methodology can be particularly valuable in a clinical trial design with correlated data, because it is often difficult to obtain reliable estimates for the intracluster correlations. In the adaptive design, the parameters for the correlation are not required to be prespecified. All the design parameters, such as the treatment difference, variance, and correlation, are updated inherently at each step in the procedure.

In the analysis of correlated data, we often encounter multiple outcomes and need to test composite hypotheses, for example, when different treatment effects are expected for several correlated outcomes. One-sided hypothesis testing with multiple endpoints has been well studied (Bloch, Lai, and Tuber-Bitter, 2001). In the scenarios that we investigated, the futility boundary would cause the trial to terminate early and to accept $H_0$, if the multivariate test of the treatment effects did not show sufficient overall superiority. Thus, the global test statistic for the usual two-sided test is virtually equivalent to the one-sided efficacy test. In the situation of forced termination of a trial due to financial or administrative reasons, we can simply assign the remaining weight to the last block and then conduct the testing procedure, which remains to be validated.

As pointed out by the referees, we often need an estimate of the sample size to make an assessment of the necessary resources and logistics of the trial during the planning stage. This initial sample size estimate can be based on the conventional fixed-sample design with initial estimates of the design parameters. By using the interim data, we can then apply the adaptation rules to extend the sample size to restore power, or to terminate the study early for futility. Because it is often more difficult to implement an adaptive design in clinical trials involving fast recruitment but time-lagged responses, the adaptive design is particularly useful for group sequential trials where the recruitment can be controlled by the investigators.

## Acknowledgements

We would like to thank the editor, an associate editor, and two referees for their critical reading and constructive comments which greatly improved the earlier version of this manuscript. We thank Jianwen Cai for helpful discussions. This research was partially supported by National Institute of Health grant CA 79466 and bladder SPORE CA 091846.

## References

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50,** 1029–1041 (Correction (1996). *Biometrics* **52,** 380).

Betensky, R. A. and Tierney, C. (1997). An examination of methods for sample size recalculation during an experiment. *Statistics in Medicine* **16,** 2587–2598.

Bloch, D. A., Lai, T. L., and Tuber-Bitter, P. (2001). One-sided tests in clinical trials with multiple endpoints. *Biometrics* **57,** 1039–1047.

Cheng, Y. and Shen, Y. (2004). Estimation of a parameter and its exact confidence interval following sequential sample size re-estimation trials. *Biometrics* **60,** 910–918.

Cui, L., Hung, H. M. J., and Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55,** 853–857.

Diggle, J. P., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd edition. New York: Oxford University Press.

Fisher, L. (1998). Self-designing clinical trials. *Statistics in Medicine* **17,** 1551–1562.

Gould, A. L. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine* **11,** 55–66.

Hartsfield, J. K. (1986). Phenytoin embryopathy: The effect of epoxide hydrolase inhibitor on chronic phenytoin exposure in utero in M57BL/6J mice. M.Sc. Thesis, Harvard University, Department of Orthodontics, Cambridge, Massachusetts.

Lake, S., Kammann, E., Klar, N., and Betensky, R. (2002). Sample size re-estimation in cluster randomization trials. *Statistics in Medicine* **21,** 1337–1350.

Lan, K. K. G., Simon, R., and Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Sequential Analysis* **1,** 207–219.

Lefkopoulou, M. and Ryan, L. (1993). Global tests for multiple binary outcomes. *Biometrics* **49,** 975–988.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55,** 1286–1290.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Liu, Q. and Chi, G. Y. H. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics* **57,** 172–177.

Liu, G. and Liang, K. Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics* **53,** 937–947.

Liu, A., Shih, W. J., and Gehan, E. (2002). Sample size and power determination for clustered repeated measurements. *Statistics in Medicine* **21,** 1787–1801.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57,** 886–891.

Park, C. G., Park, T., and Shin, D. W. (1996). A simple method for generating correlated binary variates. *American Statistician* **50,** 306–310.

Posch, M. and Bauer, P. (2000). Interim analysis and sample size reassessment. *Biometrics* **56,** 1170–1176.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51,** 1315–1324.

Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* **90,** 455–463.

Rochon, J. (1991). Sample size calculations for two-group repeated-measures experiments. *Biometrics* **47,** 1383–1398.

Self, S. G. and Mauritsen, R. H. (1988). Power/sample size calculations for generalized linear models. *Biometrics* **44,** 79–86.

Shen, Y. and Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55,** 190–197.

Shih, W. J. (1992). Sample size reestimation in clinical trials. In *Biopharmaceutical Sequential Statistical Applications*, K. Peace (ed), 285–301. New York: M. Dekker.

Vonesh, E. F. and Schork, M. A. (1986). Sample sizes in the multivariate analysis of repeated measurements. *Biometrics* **42,** 601–610.

Wassmer, G. (1998). A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics* **54,** 696–705.

Zucker, D. M. and Denne, J. (2002). Sample-size redetermination for repeated measures studies. *Biometrics* **58,** 548–559.

*Received January* 2004. *Revised September* 2004. *Accepted October* 2004.

## Appendix

### *Asymptotic Normality of* $\mathbf{P}_m$

Recall that $\mathcal{F}_j$ is $\sigma$-algebra induced by the observed data up to the $j$th block. Note that $m = \inf\{k : \sum_{j=1}^{k} w_j^2 = 1\}$ is a stopping time and $\mathcal{F}_{m-1}$ measurable, where $w_j = w_j(\mathcal{D}_{j-1})$. Given $\mathcal{F}_{k-1}, w_k(\mathbf{U}_k - (B_k)^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-\frac{1}{2}}\boldsymbol{\phi})|\mathcal{F}_{k-1} \sim N_p(\mathbf{0}, w_k^2\mathbf{I})$, and the characteristic function is

$$E\{\exp(i\mathbf{t}'w_k\mathbf{U}_k)|\mathcal{F}_{k-1}\} = \exp\left(i\mathbf{t}'w_k\sqrt{B_k}\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-\frac{1}{2}}\boldsymbol{\phi} - w_k^2\mathbf{t}'\mathbf{t}/2\right).$$

Define

$$S_k = \frac{\exp\left(i\mathbf{t}'\sum_{j=1}^{k} w_j\mathbf{U}_j\right)}{\exp\left(i\mathbf{t}'\sum_{j=1}^{k} w_j\sqrt{B_j}\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-\frac{1}{2}}\boldsymbol{\phi} - \sum_{j=1}^{k} w_j^2\mathbf{t}'\mathbf{t}/2\right)}.$$

Through some algebraic manipulations, we have $E(S_k \,|\, \mathcal{F}_{k-1}) = S_{k-1}$, and thus $(S_k, \mathcal{F}_k; k = 1, \ldots, m)$ is a bounded martingale with $E(S_1) = E(S_m) = 1$. By using the fact that $\hat{\boldsymbol{\Sigma}}_{\phi_j}$ is a consistent estimator of $\boldsymbol{\Sigma}_{\phi}$, we can show the asymptotic normality of $\mathbf{P}_m$.