Flexible Sample Size Considerations Using Information-Based Interim Monitoring

2 authors, including:

Some of the authors of this publication are also working on these related projects:

Adaptive Clinical Trials View project

Power for categorical data View project

FLEXIBLE SAMPLE SIZE CONSIDERATIONS USING INFORMATION-BASED INTERIM MONITORING

CYRUS R. MEHTA

Cytel Software Corporation, Cambridge, Massachusetts

ANASTASIOS A. TSIATIS

North Carolina State University, Raleigh, North Carolina

At the design phase of a clinical trial the total number of participants needed to detect a clinically important treatment difference with sufficient precision frequently depends on nuisance parameters such as variance, baseline response rate, or regression coefficients other than the main effect. In practical applications, nuisance parameter values are often unreliable guesses founded on little or no available past history. Sample size calculations based on these initial guesses may, therefore, lead to under- or over-powered studies. In this paper, we argue that the precision with which a treatment effect is estimated is directly related to the statistical information in the data. In general, statistical information is a complicated function of sample size and nuisance parameters. However, the amount of information necessary to answer the scientific question concerning treatment difference is easily calculated a priori and applies to almost any statistical model for a large variety of endpoints. It is thus possible to be flexible on sample size but rather continue collecting data until we have achieved the desired information. Such a strategy is well suited to being adopted in conjunction with a group sequential clinical trial where the data are monitored routinely anyway. We present several scenarios and examples of how group sequential information-based design and monitoring can be carried out and demonstrate through simulations that this type of strategy will indeed give us the desired operating characteristics.

Key Words: Sample size reestimation; Information-based design and monitoring; Adaptive design

INTRODUCTION

AT THE DESIGN PHASE of a randomized clinical trial the total number of participants needed to achieve a certain level of significance and power frequently depends on nuisance parameters such as variance, baseline response rate, or regression coefficients other than the main effect. In practical applications, nuisance parameter values are often unreliable guesses founded on little or no available past history. As a result, if the initial guesses for

the nuisance parameters are far from the truth, then the study may be under- or over-powered to detect the desired treatment difference.

For example, in a clinical trial comparing dichotomous responses, suppose the investigators want to design the study to detect a 10% difference in the response rate between a new treatment and a control treatment with 90% power using a test at the 0.05 (two-sided) level of significance. The traditional approach would necessitate that an initial guess be made for the response rate in the control group. Say this is given to be 20%, using the best available data. It is then a simple exercise to show that a total of 778 participants, randomized with equal probability to the two treatments, would provide the required 90% power for a two-sided level 0.05 test to detect a difference from 20% to 30%. In truth, however, suppose the initial guess was wrong and the response rate for the control group was 40%. Then with a sample size of 778, the power of the binomial test at the .05 level of significance to detect a 10% difference, or a difference from 40% to 50%, would have been diminished to 80%. To achieve 90% power one would need 1030 participants.

For ethical as well as practical considerations most large Phase III randomized clinical trials are designed with formal sequential stopping rules. The data from such a clinical trial are monitored at interim time points, usually by an external data and safety monitoring board, and if the observed treatment difference is sufficiently large at one of these interim analyses, the study may be stopped. Formal sequential boundaries have been derived that dictate how large the treatment difference must be at different interim analyses before a study is stopped. These boundaries are constructed so that the overall test has the desired operating characteristics, that is, the resulting sequential test will have the desired level of significance and power to detect a clinically important treatment difference. Issues regarding the effect on power of misspecifying the nuisance parameter during the design stage are similar to those discussed earlier for the fixed sample procedure. However, when data are monitored periodically, the nuisance parameters in the model can be estimated with the available data and, if these estimates are sufficiently far from the values used at the design stage, the investigators have the ability to alter the design, that is, to increase or decrease the sample size adaptively. It is the feasibility of this strategy that we will investigate in this paper.

INFORMATION-BASED DESIGN AND INTERIM MONITORING

In this section we will develop a single unified nuisance-parameter-free, information-based approach for designing and monitoring clinical trials. This approach is applicable to studies involving dichotomous response, continuous response, time-to-event response, and longitudinal response, in the presence of one or more nuisance parameters, including covariates. In all of these settings we assume that the data are generated from some multiparameter probability model in which a single unknown parameter, δ, characterizes the benefit conferred by the new treatment under investigation while the remaining unknown parameters are denoted as nuisance parameters whose values play no role in determining the clinical importance of the new treatment, but are important in properly describing the probability distribution that generates the data. Interest focuses on testing the null hypothesis

$$H_0: \delta = 0 \tag{1}$$

at the α level of significance. Suppose a two-sided test is to be conducted and is required to have power equal to $1 - \beta$ against the alternative hypothesis

$$H_a: \delta = \delta_a. \tag{2}$$

Designing Fixed-Information Studies

A typical way to carry out a two-sided fixed-information test of the null hypothesis (1) is to fit the underlying model to all the data available at the end of the study, compute the Wald test statistic

$$T = \frac{\hat{\delta}}{\text{se}(\hat{\delta})}, \tag{3}$$

and reject H_0 if

$$|T| \geq z_{\alpha/2}, \tag{4}$$

where z_u is the $(1 - u)$th quantile of the standard normal distribution and $\text{se}(\hat{\delta})$ is an estimate of the standard error of $\hat{\delta}$. A crucial question to ask is, how much data should we collect before we perform the hypothesis test (4)? In other words, how much "information" should we gather? The term information (or Fisher information) has a strict technical meaning, which is related to the variance of $\hat{\delta}$. As one might expect intuitively, the smaller the variance, the more the information. Now the **true** variance of $\hat{\delta}$ is usually not known. For all practical purposes, however, the Fisher information available at the end of the study can be well approximated by the inverse of the **estimated** variance of $\hat{\delta}$, or $[\text{se}(\hat{\delta})]^{-2}$. Throughout this development we shall use this approximation as though it were the actual Fisher information. The amount of Fisher information, I, needed in order for the test (4) to achieve a power of $1 - \beta$ can be derived by standard statistical methods as

$$I = \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta_a} \right]^2 \tag{5}$$

where I is approximated by

$$I \approx \left[\text{se}(\hat{\delta}) \right]^{-2}. \tag{6}$$

Thus, operationally, in a fixed-information study one would gather data until the inverse square of the standard error of the estimate of δ equaled the right hand side of equation (5), and then perform the hypothesis test (4).

Example 1: Normal Response with Unknown Variance. Let X_A be the response of a subject who receives treatment A and X_B be the response of a subject who receives treatment B. Let μ_A and μ_B be the expected values of the random variables X_A and X_B, respectively, both normally distributed with a common unknown variance σ^2. We are interested in testing the null hypothesis (1) in which

$$\delta = \mu_A - \mu_B. \tag{7}$$

In a fixed-information study we do not fix the sample size in advance. Rather, we continue to enroll subjects into the clinical trial until we have gathered a sufficient number, say n_A on treatment A and n_B on treatment B, so as to satisfy the fixed information requirement

$$\left[\text{se}(\hat{\delta}) \right]^{-2} \equiv \left[\frac{\hat{\sigma}^2}{n_A} + \frac{\hat{\sigma}^2}{n_B} \right]^{-1} \geq \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta_a} \right]^2. \tag{8}$$

We then perform the hypothesis test (4) using the test statistic (3) in which $\hat{\delta} = \bar{X}_A - \bar{X}_B$. As long as (8) is satisfied, the study will have the desired $1 - \beta$ power regardless of the true value of the nuisance parameter σ^2.

Example 2: Dichotomous Response with Unknown Baseline Probability. Let π_A and π_B be the response probabilities for treatment A and B, respectively. We are interested in testing the null hypothesis (1) in which

$$\delta = \pi_A - \pi_B, \tag{9}$$

and the baseline response probability π_B is unknown. In a fixed-information study we enroll subjects into the clinical trial until the sample size, say n_A on treatment A and n_B on treatment B, is large enough to satisfy

$$\left[\text{se}(\hat{\delta}) \right]^{-2} \equiv \left[\frac{\hat{\pi}_A(1 - \hat{\pi}_A)}{n_A} + \frac{\hat{\pi}_B(1 - \hat{\pi}_B)}{n_B} \right]^{-1} \geq \left[\frac{z_{\alpha/2} + z_\beta}{\delta_a} \right]^2. \tag{10}$$

We then perform the hypothesis test (4) using the test statistic (3) in which $\hat{\delta} = \hat{\pi}_A - \hat{\pi}_B$. With this approach the study will have the desired $1 - \beta$ power regardless of the true value of the nuisance parameter π_B.

Example 3: Time-to-Event Response from the Proportional Hazards Model. Suppose the data are generated from the Cox proportional hazards model

$$\lambda(t \,|\, Z_0, \; Z_1) = \lambda_0(t) e^{\delta Z_0 + \eta Z_1}. \tag{11}$$

We are interested in testing the null hypothesis that $\delta = 0$ in the presence of possibly several nuisance parameters η. In a fixed-information study we enroll subjects into the clinical trial and follow them until we have satisfied the condition

$$\left[\text{se}(\hat{\delta}) \right]^{-2} \geq \left[\frac{z_{\alpha/2} + z_\beta}{\delta_a} \right]^2, \tag{12}$$

where $\hat{\delta}$ is the maximum partial likelihood estimate of δ, and $\text{se}(\hat{\delta})$ is its standard error. Once we have gathered sufficient information, as evidenced by satisfying (12), we perform the hypothesis test (4). With this approach the study will have the desired $1 - \beta$ power regardless of the true values of the nuisance parameters η. Note that many combinations of subject accrual and follow-up time will satisfy (12). For details concerning the trade-off between additional accrual and additional follow-up time, refer to Kim and Tsiatis (1), and Scharfstein and Tsiatis (2).

Example 4: Longitudinal Response from the Random Effects Model. Suppose the data are generated from the random effects model

$$Y_{ijk} = \alpha_{ik} + \gamma_{ik} t_{ijk} + \varepsilon_{ijk}, \tag{13}$$

where

$$\begin{bmatrix} \alpha_{ik} \\ \gamma_{ik} \end{bmatrix} \sim N\left(\begin{bmatrix} A_k \\ G_k \end{bmatrix}, \begin{bmatrix} \sigma^2_{A_k} & \sigma_{A_k,G_k} \\ \sigma_{A_k,G_k} & \sigma^2_{G_k} \end{bmatrix} \right)$$

and

$$\varepsilon_{ijk} \sim N(0, \sigma^2_{\varepsilon_k}),$$

$$k = 1, 2, j = 1, \ldots, n_{ik}, i = 1, \ldots, m_k.$$

We are interested in testing the null hypothesis

$$H_0: \delta = G_1 = G_2 = 0$$

in the presence of possibly unknown variance components. In a fixed-information study we enroll subjects into the clinical trial and gather longitudinal information on them until we have satisfied the condition

$$\left[se\left(\hat{\delta} \right) \right]^{-2} \geq \left[\frac{z_{\alpha/2} + z_\beta}{\delta_a} \right]^2, \tag{14}$$

where $\hat{\delta}$ is the maximum likelihood estimate of δ, and $se(\hat{\delta})$ is its standard error. Once we have gathered sufficient information, as evidenced by satisfying (14), we perform the hypothesis test (4). With this approach the study will have the desired $1 - \beta$ power regardless of the true values of the nuisance parameters η. Note that many combinations of subject accrual and number of longitudinal observations per subject will satisfy (14). For a detailed implementation of this rather complex design refer to Scharfstein, Tsiatis, and Robins (3).

Designing Maximum Information Studies within the Group Sequential Framework

The fixed-information strategy outlined in the previous section has two limitations:

1. Unless the accruing data are monitored at interim time points one has no means of determining if the desired information has been attained, and
2. Since information is an abstract quantity, knowing how much information to gather over the course of a study is not of much practical use in planning the study. For planning purposes one needs to translate the desired information into a physical resource such as sample size. For this purpose, one must use some initial estimates or guesses of the nuisance parameters.

Group sequential designs provide a natural way to overcome both limitations. In a group sequential study one is already intending to monitor the accruing data at administratively convenient interim monitoring time points, with a view to early stopping if a stopping boundary is crossed. One can, therefore, take advantage of the interim looks to compute the current information as well as to update prior estimates of nuisance parameters with the help of the latest available data.

Suppose that instead of performing only one significance test at the end of the study we decide to perform up to K repeated significance tests at interim monitoring times τ_1, τ_2, $\ldots \tau_K$, respectively, and to terminate the study at the first test that rejects the null hypothesis.

The flexibility to monitor the data in this fashion and possibly terminate the study early comes at a cost:

1. Repeated application of the same significance test (4) will result in an elevated Type 1 error. In order to maintain the Type 1 error at level α, the criterion for rejecting the null hypothesis must be made more stringent. Therefore, if we intend to monitor the trial K times, the strategy would be to stop the trial and reject the null hypothesis at the first time point τ_j, $j \leq K$, for which the corresponding test statistic $|T_j|$, calculated using all the available data up to time τ_j, exceeds the stopping boundary c_j, where c_j must exceed $z_{\alpha/2}$. If the stopping boundary is not crossed at any of the K time points, then the study is terminated at the Kth and final analysis, and

2. The maximum information to be committed up-front in order to achieve a power of $1 - \beta$ gets inflated. We must now commit to keeping the study open until the total information equals

$$I_{\max} = \left[\frac{z_{\alpha/2} + z_\beta}{\delta_a} \right]^2 \times \text{IF}(\Delta, \alpha, \beta, K) \tag{15}$$

where IF(.) is an inflation factor that depends on α, β, K, and Δ is a "shape parameter" to be defined shortly that determines the shape of the stopping boundary over the K repeated significance tests. We must keep in mind that although the maximum information is inflated, the average time to stopping the trial, using a group sequential design, is decreased, and, in some cases, this decrease is substantial.

Derivation of the Stopping Boundaries. We now show how to obtain the stopping boundaries, c_j, c_2, ... c_K. For ease of exposition we will only discuss two-sided repeated significance tests and early stopping to reject H_0. It is convenient to define the information fraction

$$t_j = \frac{I(\tau_j)}{I_{\max}} \approx \frac{\left[\text{se}\!\left(\hat{\delta}(\tau_j) \right) \right]^{-2}}{\left[\text{se}\!\left(\hat{\delta}(\tau_K) \right) \right]^{-2}} \tag{16}$$

where $I(\tau_j)$ is the information about δ available at interim monitoring time τ_j as estimated by $[\text{se}(\hat{\delta}(\tau_j))]^{-2}$. At the jth interim monitoring time point we compute the test statistic

$$T(t_j) = \frac{\hat{\delta}(\tau_j)}{\text{se}\!\left(\hat{\delta}(\tau_j) \right)} \tag{17}$$

and reject the null hypothesis if $|T(t_j)| \geq c_j$. In this case, unlike the fixed-information test, we cannot set $c_j = z_{\alpha/2}$ for each j. Unless the stopping boundaries c_1, c_2, ... c_K are suitably adjusted, the above multiple testing strategy will inflate the overall Type 1 error of the testing procedure. A popular family of stopping boundaries (adopted, for example, in the EaSt [4] software package) is the Wang and Tsiatis (5) family of the form

$$c_j = \frac{C(\Delta, \alpha, K)}{t_j^{1/2-\Delta}}, \tag{18}$$

in which Δ is a prespecified boundary shape parameter in the range $0 \leq \Delta \leq 0.5$, and the coefficient $C(\Delta, \alpha, K)$ is so chosen that the probability of crossing a stopping boundary under the null hypothesis is α. In other words, $C(\Delta, \alpha, K)$ is computed so as to satisfy the relationship

$$1 - P_{H_0}\left\{ \bigcap_{j=1}^{K} |T(t_j)| < c_j \right\} = \alpha. \tag{19}$$

By choosing appropriate coefficients $C(\Delta, \alpha, K)$ so as to satisfy Equation (19) with c_j substituted from equation (18), we ensure that the Type 1 error of the sequential testing procedure will be preserved at level α. Notice that if $\Delta = 0$, the boundaries decrease in proportion to the inverse square root of t_j or equivalently, in proportion to the inverse square root of the information at time τ_j. These are the so called O'Brien-Fleming (6) boundaries. On the other hand, if $\Delta = 0.5$, the boundaries remain constant at each interim look. These are the so called Pocock boundaries (7). Boundary shapes in between these two extremes are obtained by specifying values of Δ between 0 and 1. Figure 1 displays the O'Brien-Fleming and Pocock boundaries for $K = 5$ equally spaced interim monitoring time points.

All of these boundaries can be adjusted appropriately by the Lan and DeMets (8) α-spending function methodology if the actual times at which interim monitoring occurs differ from the times specified at the design stage.

In order to obtain the appropriate values for the constants, $C(\Delta, \alpha, K)$, we must find a way to evaluate the left hand side of equation (19) for any specific choice of the c_js. For this we need to know the joint distribution of $\{T(t_1), T(t_2), \ldots T(t_K)\}$. We rely on a very general and powerful theorem by Scharfstein, Tsiatis, and Robins (3), which may be stated as follows:

5-Look O'Brien-Fleming Boundaries

5-Look Pocock Boundaries
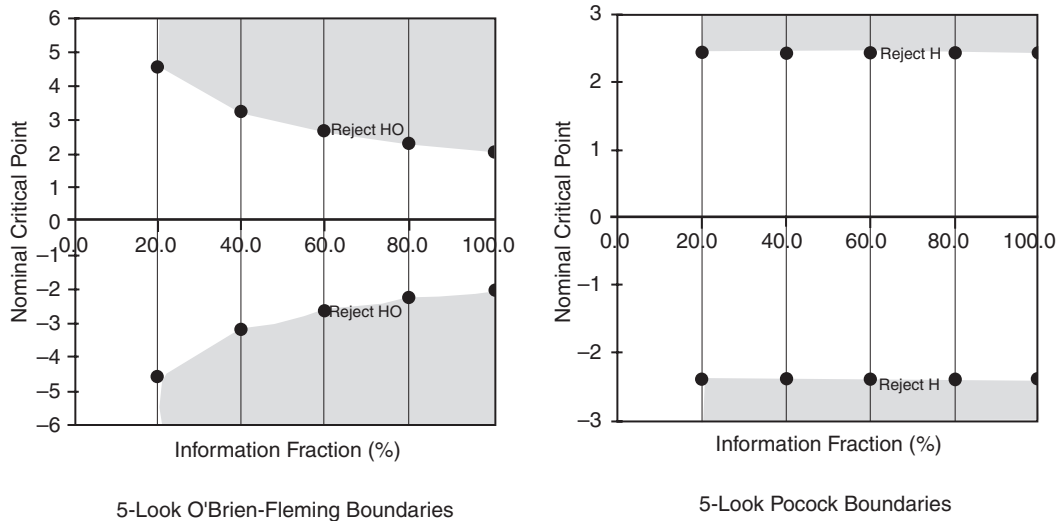
FIGURE 1. O'Brien-Fleming stopping and Pocock stopping boundaries in the EaSt software.

Theorem 1. If $\hat{\delta}$ is an efficient estimate of δ then, regardless of the model generating the data, the asymptotic joint distribution of the sequentially computed K-dimensional test statistic $\{T(t_1), T(t_2), \ldots T(t_K)\}$ is multivariate normal with mean vector

$$\mu_T = \delta\sqrt{I_{max}} \left[\sqrt{t_1}, \sqrt{t_2}, \ldots \sqrt{t_K} \right] \tag{20}$$

and variance-covariance matrix

$$V_T = \begin{bmatrix} 1 & \left[\dfrac{t_1}{t_2}\right]^{1/2} & \cdots & \left[\dfrac{t_1}{t_K}\right]^{1/2} \\[2mm] \left[\dfrac{t_1}{t_2}\right]^{1/2} & 1 & \cdots & \left[\dfrac{t_2}{t_K}\right]^{1/2} \\[2mm] \vdots & \vdots & \ddots & \vdots \\[2mm] \left[\dfrac{t_1}{t_K}\right]^{1/2} & \left[\dfrac{t_2}{t_K}\right]^{1/2} & \cdots & 1 \end{bmatrix}. \tag{21}$$

Under the null hypothesis H_0, the mean of this multivariate distribution is $\mu_T = 0$. The distribution is, thus, completely determined and we can solve equation (19) in terms of $C(\Delta, \alpha, K)$ by means of numerical integration. The numerical integration is considerably simplified because the special structure of V_T implies that $\sqrt{t_j}T(t_j)$ is a process of independent increments. This enables us to apply the recursive integration techniques of Armitage, McPherson, and Rowe (9) and evaluate equation (19) very rapidly for a large family of stopping boundaries parameterized by Δ. It is important to note that Theorem 1 applies to any probability model generating the data as long as $\hat{\delta}$ is an efficient estimate of δ. Test statistics based on likelihood methods such as Wald tests, score tests, or likelihood ratio tests are all efficient and fit into this general framework. Thus, whether we have dichotomous data, normally distributed data, censored survival data, or longitudinal data, the same unified distribution theory applies in each case.

Derivation of the Inflation Factor. Having determined the shape of the stopping boundary, our next task is to evaluate the value of I_{max} that will ensure that the study has the desired amount of power $1 - \beta$. This time we must ensure that for a specified alternative hypothesis $H_a : \delta = \delta_a$,

$$P_{H_a}\left\{ \bigcap_{j=1}^{K} |T(t_j)| < c_j \right\} = \beta. \tag{22}$$

Since, under H_a, $\mu_T = \delta_a\sqrt{I_{max}} \left[\sqrt{t_1}, \sqrt{t_2}, \ldots \sqrt{t_K} \right]$, the joint distribution of $\{T(t_1), T(t_2), \ldots T(t_K)\}$ is fully determined by the value of

$$\eta = \delta_a\sqrt{I_{max}}, \tag{23}$$

also known as the drift parameter. We may thus solve equation (22) by numerical integration in terms of the drift parameter, η. Again, the independent increments structure of V_T facilitates the numerical integration. The value of η that solves equation (22) will depend on the

coefficient $C(\Delta, \alpha, K)$ associated with the stopping boundary and on β. Denote the solution by $\eta(\Delta, \alpha, \beta, K)$. Then, by substitution into equation (23),

$$I_{\max} = \left[\frac{\eta(\Delta, \alpha, \beta, K)}{\delta_a} \right]^2. \tag{24}$$

Substituting the value of I from the fixed sample equation (5) we have

$$I_{\max} = \left[\frac{z_{\alpha/2} + z_\beta}{\delta_a} \right]^2 \times \text{IF}(\Delta, \alpha, \beta, K) \tag{25}$$

where IF $(\Delta, \alpha, \beta, K)$, the inflation factor, is given by

$$\text{IF}(\Delta, \alpha, \beta, K) = \left[\frac{\eta(\Delta, \alpha, \beta, K)}{z_{\alpha/2} + z_\beta} \right]^2. \tag{26}$$

These results apply to any model in which a single parameter, δ, captures the primary endpoint of interest. We may, thus, use equation (25) to convert any fixed sample study into a corresponding group sequential study. A selection of inflation factors for the classical Pocock and O'Brien-Fleming α-spending functions, for various choices of Δ, α, β, and K, are displayed in Table 1.

Calculators for Translating Fisher Information into Sample Size. It is convenient to express the patient resources needed to achieve the goals of a study in terms of Fisher information. The calculation is straightforward, requires no knowledge of unknown nuisance parameters, and facilitates the tabulation of inflation factors that can be utilized to convert fixed sample studies into group sequential ones. But Fisher information is difficult to interpret in physical terms. For planning purposes, it is necessary to express the patient resources in a more concrete form such as sample size. However, while a single formula (5) suffices for obtaining the Fisher information regardless of the model generating the data, the same is not true for sample size calculations. Each model requires its own sample size calculator into which

TABLE 1
Inflation Factors for Pocock and O'Brien-Fleming (O-F)
Alpha Spending Functions

| | | $\alpha = 0.05$ (two-sided) | | | | | $\alpha = 0.01$ (two-sided) | | |
| | | Power $(1 - \beta)$ | | | | | Power $(1 - \beta)$ | | |
K	Spending Function	0.80	0.90	0.95	K	Spending Function	0.80	0.90	0.95
2	Pocock	1.11	1.10	1.09	2	Pocock	1.09	1.08	1.08
2	O-F	1.01	1.01	1.01	2	O-F	1.00	1.00	1.00
3	Pocock	1.17	1.15	1.14	3	Pocock	1.14	1.12	1.12
3	O-F	1.02	1.02	1.02	3	O-F	1.01	1.01	1.01
4	Pocock	1.20	1.18	1.17	4	Pocock	1.17	1.15	1.14
4	O-F	1.02	1.02	1.02	4	O-F	1.01	1.01	1.01
5	Pocock	1.23	1.21	1.19	5	Pocock	1.19	1.17	1.16
5	O-F	1.03	1.03	1.02	5	O-F	1.02	1.01	1.01

one must input not only the effect size δ_a, desired power $1 - \beta$, and significance level α, but also our best estimates of unknown nuisance parameters. The EaSt (4) software provides such calculators for two-sample tests involving normal, binomial, and survival endpoints, with an inflation factor built-in to accommodate interim monitoring. It is possible to extend these calculators to the family of generalized linear models so as to accommodate covariates. Many important models including linear regression, logistic regression, Poisson regression, and the Cox proportional hazards model belong to this family. For logistic and Poisson regression EaSt (4) provides suitable calculators on the basis of the theory developed by Self, Mauritsen, and Ohara (10). For very complex survival and longitudinal designs, sample size is not the only design parameter. Decisions must also be made concerning the total study duration, the patient accrual period, the number of repeated measures for the primary endpoint, and the spacing between the repeated measures. In these complex settings the conversion of Fisher information into a physical resource can be achieved most effectively through simulation.

Example 1: A Dichotomous Outcome Maximum Information Trial

We wish to design a study to detect a difference $\delta = \pi_A - \pi_B$ between two binomial populations with unknown response probabilities π_A and π_B, respectively. The study is to be designed for up to four looks at the data with early stopping based on O'Brien-Fleming two-sided level-0.05 stopping boundaries and 90% power to detect a difference $\delta = 0.15$. Substituting these design parameters into equation (15) and reading off the appropriate inflation factor from Table 1 we obtain $I_{\max} = 477$. Thus, the monitoring strategy calls for accruing subjects into the study until the total information, as measured by $[\text{se}(\hat{\delta})]^{-2}$ equals or exceeds 477 units or until a stopping boundary is crossed, whichever comes first. Notice that in this design, we did not have to specify the individual values of π_A and π_B. It was sufficient to specify the clinically meaningful difference δ for which 90% power was desired.

It is, however, difficult to know how long to accrue subjects when the accrual goals are expressed in units of square inverse standard error instead of being expressed in terms of a physical quantity such as sample size. We need to translate units of information into sample size units. This is easy to do since the variance of $\hat{\delta}$ is a simple function of π_B, δ, and the total sample size, N. Thus

$$\left[\text{se}\left(\hat{\delta}\right)\right]^{-2} = \left[\frac{(\pi_B)(1 - \pi_B)}{N/2} + \frac{(\pi_B + \delta)(1 - \pi_B - \delta)}{N/2} \right]^{-1} = 477, \qquad (27)$$

whereupon

$$N = 2 \times 477 \times [(\pi_B + \delta)\, (1 - (\pi_B + \delta_1)) + (\pi_B)\, (1 - \pi_B)]. \qquad (28)$$

It is clear from equation (28) that, for the same clinically meaningful difference $\delta = 0.15$, different values of the baseline probability π_B will lead to different sample sizes. Based on historical data we assume that $\pi_B = 0.15$. Upon substituting this number into equation (28), the required $I_{\max} = 477$ units of information is translated into $N_{\max} = 323$ subjects. Of course, if the assumption that the baseline response rate is 0.15 is incorrect, then 323 subjects will not produce the desired operating characteristics. Depending on the actual value of the baseline response rate we might have either an underpowered or overpowered study. We now show that one of the major advantages of the information-based approach is that we can use all the data accrued at any interim monitoring time point to reestimate the baseline

response rate and, if it differs from what was assumed initially, recalculate the sample-size.

Results at the First Interim Monitoring Time Point. Suppose that at the first interim monitoring time point, τ_1, we observe 15/60 responders on Treatment A and 14/60 responders on Treatment B. Then $\hat{\delta}(\tau_1) = 0.0167$, and $\text{se}(\hat{\delta}(\tau_1)) = 0.0781$. The information, $I(\tau_1)$, at the first interim monitoring time point is estimated by $\text{se}(\hat{\delta}(\tau_1))]^{-2} = 163.8$. Therefore, the information fraction at calendar time τ_1 is $t_1 = I(\tau_1)/477 = 0.343$. The EaSt (4) software may now be used to apply the Lan and DeMets (8) methodology to the O'Brien-Fleming error spending function at the information fraction $t_1 = 0.343$. We thereby obtain the corresponding stopping boundary as 3.47. The observed value of the test statistic is $T(t_1) = \hat{\delta}(\tau_1)/\text{se}(\hat{\delta}(\tau_1)) = 0.21$. Since this value is less than 3.47, the study continues on to the next look.

We have accrued 120 subjects out of the 323 required under the design assumption that the nuisance parameter is $\pi_B = 0.15$. The information fraction under this design assumption is thus $120/323 = 0.37$, while the actual information fraction is $t_1 = 0.34$. Thus, the information appears to be coming in a little slower than anticipated, but this difference does not seem serious enough to alter the sample size requirements of the study.

Results at the Second Interim Monitoring Time Point. Suppose that at the second interim monitoring time point, τ_2, we observed 41/120 responders on Treatment A and 29/120 responders on Treatment B. The information accrued at this time point is estimated as $\text{se}(\hat{\delta}(\tau_2))]^{-2} = 293.978$ and the information fraction is $t_2 = I(\tau_2)/477 = 0.616$. The stopping boundary at this information fraction is obtained from EaSt (4) as 2.605. The observed value of the test statistic is computed as $T(t_2) = \hat{\delta}(\tau_2)/\text{se}(\hat{\delta}(\tau_2)) = 1.715$. Since $1.715 < 2.605$, the stopping boundary is not crossed and the study continues on to the next look.

This time the anticipated information fraction under the assumption that $\pi_B = 0.15$ is $240/323 = 0.74$, which is considerably larger than the actual information fraction $t_2 = 0.616$. Thus, there is considerable evidence that the information is coming in slower than anticipated. In fact, the data suggest that the value of π_B is close to 0.25. It might, therefore, be prudent to reestimate the sample size of the study. The new maximum sample size can be obtained by the relationship

$$\frac{N(\tau_2)}{N_{\max}} = \frac{I(\tau_2)}{I_{\max}} .$$

Thus, the maximum sample size (rounded up to the nearest integer) is

$$N_{\max} = N(\tau_2) \times \frac{I_{\max}}{I(\tau_2)} = 240 \times \frac{477}{293.978} = 390.$$

Therefore, we need to commit 390 subjects to the study, not 323 as originally estimated. We see that the original design with 323 subjects would have led to a seriously underpowered study.

Results at the Third Interim Monitoring Time Point. We continue to accrue subjects beyond the 323 in the original design, and reach the third interim monitoring time point at time τ_3 with 61/180 responders on Treatment A and 41/180 responders on Treatment B. At this look the total information accrued is $\text{se}(\hat{\delta}(\tau_3))]^{-2} = 450.07$ and the information fraction is $t(\tau_3) = I(\tau_3)/477 = 0.943$. The observed value of the test statistic is $T(t_3) = 2.357$. The stopping boundary at information fraction $t_3 = 0.943$ is obtained by EaSt (4) to be 2.062. Since

the observed value of the test statistic, 2.357, exceeds the corresponding stopping boundary, 2.062, the stopping boundary is crossed and the study terminates with a statistically significant outcome.

This example highlights the fundamental difference between a maximum information study and a maximum sample size study in a group sequential setting. Had the study been monitored by the conventional method, the maximum sample size would have been fixed from the start at 323 subjects and there would have been no flexibility to change the level of this physical resource over the course of the study. But in an information-based approach the maximum information is fixed, not the maximum amount of a physical resource. Thus, the maximum sample size could be altered over the course of the study from 323 to 390 subjects, while the maximum information stayed constant at 477 units. Without this flexibility, the power of the study would be severely compromised.

Example 2: Normal Outcome Maximum Information Trial

Facey (11) describes a placebo controlled efficacy trial investigating a new treatment for hypercholesterolemia where the primary response, reduction in total serum cholesterol over a four-week period, is assumed to be normally distributed. Suppose the goal is to design a four-look group sequential study having 90% power to detect a reduction in serum cholesterol of 0.4 mmol/litre with a two-sided level 0.05 test. The variance in the cholesterol levels among the subjects in the study is unknown. However, we do not need to know the value of this nuisance parameter to determine the maximum information needed to attain 90% power. Based on equation (15) we can compute the maximum information as $I_{\max} = 67$.

Although the variance is unknown, we need to make an initial guess at this nuisance parameter to come up with a preliminary estimate of the maximum sample size. If we assume a balanced design then

$$\left[\mathrm{se}\left(\hat{\delta}\right) \right]^{-2} = \left[\frac{4\sigma^2}{N_{\max}} \right]^{-1} = 67, \tag{29}$$

whereupon

$$N_{\max} = 4 \times \sigma^2 \times 67. \tag{30}$$

Thus, different values of σ^2 give different maximum sample sizes but the same maximum information. In fact, it is believed that the variance in serum cholesterol levels across patients on this clinical trial is 0.5. Thus, our initial estimate of the maximum sample size is $N_{\max} = 4 \times 0.5 \times 67 = 134$ patients. We will, however, monitor this trial on the information scale rather than the sample size scale, so as to avoid dependence on the assumption that $\sigma^2 = 0.5$.

Results at the First Interim Monitoring Time Point. Suppose that at the first interim monitoring time point, there were 35 subjects on the control arm (arm C), 34 subjects on the treatment arm (arm T), $\bar{x}_C = 4.8$, $\bar{x}_T = 4.58$, $s_C = 0.88$, and $s_T = 0.9$. From these observed values we obtain the current information as 21.7775 and the current value of the test statistic as -0.0267. The information fraction is $21.7775/67 = 0.33$. The O'Brien-Fleming lower stopping boundary at this information fraction is obtained from EaSt (4) as -3.571. Since the test statistic has not crossed this lower boundary, the study continues to the next look. It appears that the standard deviation of the serum cholesterol levels is somewhat larger

than the value $\sqrt{0.5} = 0.707$ estimated prior to activating the study. As this is only the first look at the data, however, it might be preferable to wait for additional data to accrue before making a decision to reestimate the sample size.

Results at the Second Interim Monitoring Time Point. Suppose that at the second interim monitoring time point, there were 55 subjects on the control arm (arm C), 58 subjects on the treatment arm (arm T), $\bar{x}_C = 4.75$, $\bar{x}_T = 4.39$, $s_C = 0.92$, and $s_T = 0.95$. Based on these numbers we compute the current value of the information as 32.25 and the current value of the test statistic as -2.0446. We might also revise the sample size requirements based on the current estimates of standard deviation. We actually need a maximum sample size of 236 patients, not 135 as estimated before the study was activated. This calculation is based on preserving the ratio

$$\frac{n(\tau_2)}{n_{\max}} = \frac{I(\tau_2)}{I_{\max}}.$$

Thus, the maximum sample size (rounded up to the nearest integer) is

$$n_{\max} = n(\tau_2) \times \frac{I_{\max}}{I(\tau_2)} = 112 \times \frac{67.126}{32.2553} = 236.$$

Notice that although we have accrued 113 patients to the study the cumulative information fraction is only 0.481. Thus, it is clear that unless we increase patient accrual from the initial specification of 135, we will have a seriously underpowered study. Let us assume then that the investigators agree at this stage to increase the sample size to 236 patients.

Results at the Third Interim Monitoring Time Point. Suppose that at the third interim monitoring time point $n_C = 90$, $n_T = 91$, $\bar{x}_C = 4.76$, $\bar{x}_T = 4.29$, $s_C = 0.91$, and $s_T = 0.92$. Based on these numbers the current information is 54.04 and the current value of the test statistic is -3.4551. This time the lower stopping boundary is -2.249 and so the study terminates with the conclusion that the new treatment does indeed lower the serum cholesterol level significantly relative to the control.

SIMULATION RESULTS

The joint distribution of the sequentially computed test statistics, the derivation of the stopping boundaries, and the computation of the information fraction at each interim monitoring time point all depend for their validity on large-sample theory. Therefore, in this section we verify through simulations that the information-based design and monitoring approach discussed in the previous section does indeed preserve the Type 1 error and maintain the power of a study. We present simulation results for clinical trials with dichotomous outcomes and clinical trials with continuous, normally distributed outcomes. In both cases we compare the operating characteristics of maximum information trials with those of maximum sample size trials. All the trials were designed for 90% power to detect a specific effect size using five-look two-sided O'Brien-Fleming stopping boundaries with a Type 1 error equal to 5%.

Results for Dichotomous Endpoints

Tables 2 and 3 display the power and Type 1 error, respectively, for simulations of two-arm clinical trials with binomially distributed outcomes. Aside from the column headers,

each table contains three rows of simulation results, corresponding to different choices of response probabilities for generating the data. Each table contains six columns. Columns 1, 2, and 3 deal with trials that were designed for an effect size $\delta = 0.05$. The maximum information for such studies is computed by equation (15) to be $I_{max} = 4316$. Since $I_{max} \approx [se(\hat{\delta})]^{-2}$, it follows that

$$N_{max} = 2I_{max}[\pi_A(1 - \pi_A) + \pi_B(1 - \pi_B)]. \tag{31}$$

Therefore, under the design assumption that $\pi_A = 0.1$ and $\pi_B = 0.05$, the maximum information translates into a maximum sample size of $N_{max} = 1187$ subjects (both arms combined). Similarly, Columns 4, 5, and 6 deal with trials that were designed for an effect size $\delta = 0.1$ with $\pi_A = 0.15$ and $\pi_B = 0.05$. For trials with these input parameters, $I_{max} = 1080$ and $N_{max} = 378$.

The simulation results in Table 2 show that maximum information studies do indeed achieve the desired 90% power to detect a prespecified difference of δ, regardless of the value of the nuisance parameter, π_B. On the other hand, the actual power of a maximum sample size study designed for 90% power to detect a prespecified difference of δ declines if the true value of the nuisance parameter π_B is larger than the value assumed for design purposes, even though $\pi_A - \pi_B = \delta$.

Let us consider in detail the simulation results displayed in Columns 1, 2, and 3 of Table 2. Column 1 lists three pairs of response probabilities, ($\pi_A = 0.10$, $\pi_B = 0.05$), ($\pi_A = 0.15$, $\pi_B = 0.10$) and ($\pi_A = 0.20$, $\pi_B = 0.15$), that were used in the simulations. Each pair satisfies the condition $\pi_A - \pi_B = 0.05 = \delta$, which is substituted into equation (15) to obtain $I_{max} = 4316$. However, only the first of the three pairs also matches the response probabilities assumed at the design stage and substituted into equation (31) to compute $N_{max} = 1187$. Each row of Column 2 lists the proportion of times in 5000 simulated five-look group sequential clinical trials, *monitored on the information scale*, that the null hypothesis $\delta = 0$ is rejected, given that the response probabilities (π_A, π_B) for the simulations are taken from Column 1 of the same row. For example, if the simulations are performed with response probabilities ($\pi_A = 0.20$, $\pi_B = 0.15$), then 90.4% of 5000 simulated maximum information trials reject the null hypothesis. Each row of Column 3 lists the proportion of times in 20000 simulated five-look group sequential clinical trials, *monitored on the sample size scale*, that the null hypothesis $\delta = 0$ is rejected, given that the response probabilities (π_A, π_B) for the simulations are taken from Column 1 of the same row. For example, if the simulations are performed

TABLE 2
Power Comparison of Maximum Information and Maximum Sample Size Dichotomous Outcome Clinical Trials: Simulation Results for Five-look Group Sequential Trials, Using Two-sided $\alpha = 5\%$ O'Brien-Fleming Boundaries, Designed for 90% Power to Detect the Specified $\pi_A - \pi_B$

	Design Parameters $\pi_A = 0.10$, $\pi_B = 0.05$			Design Parameters $\pi_A = 0.15$, $\pi_B = 0.05$	
Simulation Parameters (π_A, π_B) with $\delta = 0.05$	Maximum Information Study $I_{max} = 4316$	Maximum Sample Size Study $N_{max} = 1187$	Simulation Parameters (π_A, π_B) with $\delta = 0.10$	Maximum Information Study $I_{max} = 1080$	Maximum Sample Size Study $N_{max} = 378$
(0.10, 0.05)	90.5%	90.5%	(0.15, 0.05)	88.7%	90.8%
(0.15, 0.10)	90.8%	73.4%	(0.20, 0.10)	90.5%	77.8%
(0.20, 0.15)	90.4%	61.3%	(0.30, 0.20)	89.8%	60.4%

with response probabilities ($\pi_A = 0.20$, $\pi_B = 0.15$), only 61.3% of 20000 simulated maximum sample size trials reject the null hypothesis.

We now describe how the simulations displayed in Row 3, Columns 2 and Row 3, Column 3 of Table 2 were carried out. First consider Row 3, Column 3. Each trial starts out with an up-front sample size commitment of $N_{max} = 1187$ subjects. This sample size ensures that the study will have 90% power to detect a difference $\delta = 0.05$, provided the baseline response probability is $\pi_B = 0.05$. Since, however, the baseline response probability used for the simulations is $\pi_B = 0.15$, the variance of $\hat{\delta}$ is much larger than was assumed at the design stage, and $N_{max} = 1187$ subjects are not sufficient for 90% power. This is seen clearly in the results displayed in Row 3, Column 3 of Table 2 where each simulation is run on the basis of $N_{max} = 1187$ with no provision to increase this maximum at any of the interim looks. Thus, Row 3, Column 3 shows that only 61.3% of the 20000 simulations are able to reject H_0 rather than 90% as desired.

Next consider the simulation results displayed in Row 3, Column 2. In these simulated trials the goal is to accrue as many subjects as are needed to reach $I_{max} = 4316$ units of information. We adopt the following monitoring strategy to achieve this goal. We begin with the initial assumption that a maximum sample size $N_{max} = 1187$ will suffice to attain $I_{max} = 4316$. This assumption is based on our initial assessment, made at the time of study design, that the nuisance parameter $\pi_B = 0.05$. The maximum sample size is, however, subject to revision as data from the clinical trial become available at each interim monitoring time point and provide a more accurate estimate of the nuisance parameter. We take the first interim look at a time point τ_1, after generating $N(\tau_1) = 1187/5 \approx 238$ binomial responses, 119 from $\pi_B = 0.15$, and 119 from $\pi_A = 0.20$. We then revise the maximum sample size, N_{max}, so as to equate the sample size fraction with the information fraction:

$$\frac{N(\tau_1)}{N_{max}} = \frac{I(\tau_1)}{I_{max}} .$$

We now have a more realistic estimate of the sample size that will be needed to attain $I_{max} = 4316$. The same procedure is adopted at each subsequent interim look and the value of N_{max} is thus altered adaptively until the maximum information is achieved or a stopping boundary is crossed.

The simulation results in Table 3 show that the Type 1 error of both maximum information

TABLE 3
Type 1 Error of Maximum Information and Maximum Sample Size Dichotomous Outcome Clinical Trials: Simulation Results for Five-look Group Sequential Trials, Using Two-sided $\alpha = 5\%$ O'Brien-Fleming Boundaries, Designed for 90% Power to Detect the Specified $\pi_A - \pi_B$

	Design Parameters $\pi_A = 0.10$, $\pi_B = 0.05$			Design Parameters $\pi_A = 0.15$, $\pi_B = 0.05$	
Simulation Parameters (π_A, π_B) with $\delta = 0$	Maximum Information Study $I_{max} = 4316$	Maximum Sample Size Study $N_{max} = 1187$	Simulation Parameters (π_A, π_B) with $\delta = 0$	Maximum Information Study $I_{max} = 1080$	Maximum Sample Size Study $N_{max} = 378$
(0.05, 0.05)	5.9%	4.8%	(0.05, 0.05)	4.9%	5.1%
(0.10, 0.10)	5.8%	5.1%	(0.10, 0.10)	6.1%	4.6%
(0.20, 0.20)	4.8%	5.0%	(0.20, 0.20)	5.1%	4.9%

and maximum sample size studies is preserved regardless of the value of the nuisance parameter, π_B. These simulations were carried out in the same manner as described above, but under the null hypothesis $\pi_A = \pi_B$.

Results for Continuous Normally Distributed Endpoints

Tables 4 and 5 display the power and Type 1 error, respectively, for simulations of two-arm clinical trials with normally distributed outcomes. Aside from the column headers, each table contains three rows of simulation results, corresponding to different choices of σ for generating the data. Each table contains six columns. Columns 1, 2, and 3 deal with trials that were designed for data generated from a normal distribution with an effect size $\delta = 0.2$ and a standard deviation $\sigma = 1$. The maximum information for such studies is computed by equation (15) to be $I_{max} = 270$. Since $I_{max} \approx [se(\hat{\delta})]^{-2}$, it follows that

$$N_{max} = 4\sigma^2 I_{max}. \tag{32}$$

Therefore, under the design assumption that $\sigma = 1$, the maximum information translates into a maximum sample size of $N_{max} = 1080$ subjects (both arms combined). Similarly, Columns 4, 5, and 6 deal with trials that were designed for an effect size $\delta = 0.4$ with $\sigma = 1$. For trials with these input parameters, $I_{max} = 68$ and $N_{max} = 272$.

The simulation results in Table 4 show that maximum information studies do indeed achieve the desired 90% power to detect a prespecified difference of δ, regardless of the value of the nuisance parameter, σ. On the other hand, the actual power of a maximum sample size study designed for 90% power to detect a prespecified difference of δ declines if the true value of the nuisance parameter σ is larger than the value assumed for design purposes.

Let us consider in detail the simulation results displayed in Columns 1, 2, and 3 of Table 4. Column 1 lists three values of the nuisance parameter, $\sigma = 1$, $\sigma = 1.25$, and $\sigma = 1.5$, that were used in the simulations. Only the first of these values corresponds to the value of σ that was actually used to design the study. Each row of Column 2 lists the proportion of times in 5000 simulated five-look group sequential clinical trials, *monitored on the information scale*, that the null hypothesis $\delta = 0$ is rejected, given that the standard deviation σ is taken from Column 1 of the same row. For example, if the simulations are performed with a standard deviation $\sigma = 1.5$, then 90.2% of 5000 simulated maximum information trials reject the null hypothesis. Each row of Column 3 lists the proportion of times in 20000

TABLE 4
Power Comparison of Maximum Information and Maximum Sample Size Normal Outcome Clinical Trials: Simulation Results for Five-look Group Sequential Trials, Using Two-sided $\alpha = 5\%$ O'Brien-Fleming Boundaries, Designed for 90% Power to Detect the Specified δ given $\sigma = 1$

	Design Parameters ($\delta = 0.2$), ($\sigma = 1$)			Design Parameters ($\delta = 0.4$), ($\sigma = 1$)	
Simulation Parameters ($\delta = 0.2$)	Maximum Information Study $I_{max} = 270$	Maximum Sample Size Study $N_{max} = 1080$	Simulation Parameters ($\delta = 0.4$)	Maximum Information Study $I_{max} = 68$	Maximum Sample Size Study $N_{max} = 272$
($\sigma = 1.00$)	90.7%	90.1%	($\sigma = 1.00$)	90.3%	89.7%
($\sigma = 1.25$)	90.6%	73.2%	($\sigma = 1.25$)	90.2%	73.8%
($\sigma = 1.50$)	90.2%	57.7%	($\sigma = 1.50$)	90.0%	58.1%

simulated five-look group sequential clinical trials, *monitored on the sample size scale*, that the null hypothesis $\delta = 0$ is rejected, given that the standard deviation σ is taken from Column 1 of the same row. For example, if the simulations are performed with a standard deviation $\sigma = 1.5$, only 57.7% of 20000 simulated maximum sample size trials reject the null hypothesis.

The simulations in Tables 4 and 5 were implemented using the same basic approach discussed in the section on dichotomous endpoints. The details are therefore omitted.

CONCLUSIONS

When trying to assess treatment difference in clinical trials, the scientific objective of the study can be formulated as the ability to detect clinically important differences with sufficient precision. We have shown that this precision is directly related to the statistical information in the data as measured by the magnitude of the standard error of the estimator of treatment difference. In general, statistical information is a complicated function of sample size and nuisance parameters (ie, parameters in the probability model other than treatment difference). However, the amount of information needed to answer the scientific question is easily calculated and applies generally to almost any statistical model for a large variety of endpoints. The general paradigm, currently practiced, is to posit some initial guesses for the nuisance parameters and then, using the complicated function that relates information to sample size and nuisance parameters, work out the sample size that leads to the desired information.

We argue in this article that we should instead be flexible concerning the sample size requirements, but should focus more on collecting data until we have achieved the desired information, as estimated by the inverse of the square of the standard error. Such a strategy would guarantee that we meet the scientific objective of the study. The difficulty with this approach, is, of course, the logistics of carrying it out. Investigators will not want to launch a clinical trial without some sense of how many resources will be needed to carry out the study. We suggest that the usual strategy of sample size calculations, based on initial best guesses of the nuisance parameters, be carried out to get some sense of the feasibility of the study. But the data should be monitored periodically to assess whether the desired information can be achieved with the proposed sample size. If, during the monitoring process, it looks as if the initial guesses were incorrect, then the study might be altered appropriately to meet the information goal necessary to meet the scientific objective.

Such a strategy is well suited for use in conjunction with a group sequential approach

TABLE 5
Type 1 Error of Maximum Information and Maximum Sample Size Normal Outcome Clinical Trials: Simulation Results for Five-look Group Sequential Trials, Using Two-sided $\alpha = 5\%$ O'Brien-Fleming Boundaries, Designed for 90% Power to Detect the Specified δ given $\sigma = 1$

Simulation Parameters ($\delta = 0$)	Design Parameters ($\delta = 0.2$), ($\sigma = 1$)		Simulation Parameters ($\delta = 0$)	Design Parameters ($\delta = 0.4$), ($\sigma = 1$)	
	Maximum Information Study $I_{max} = 270$	Maximum Sample Size Study $N_{max} = 1080$		Maximum Information Study $I_{max} = 68$	Maximum Sample Size Study $N_{max} = 272$
($\sigma = 1.00$)	5.3%	5.1%	($\sigma = 1.00$)	5.6%	5.3%
($\sigma = 1.25$)	4.8%	5.2%	($\sigma = 1.25$)	5.7%	5.2%
($\sigma = 1.50$)	5.0%	4.9%	($\sigma = 1.50$)	5.9%	5.3%

where the data are routinely monitored anyway. In this context we have described the general large sample theory of Scharfstein, Tsiatis, and Robins (3) in which sequentially computed test statistics of treatment effect have a common distributional structure that depends only on the treatment effect parameter and on the statistical information available at the interim monitoring time points. This result leads to a general approach to sequential monitoring that applies to virtually all clinical trials. Consequently, not only are we able to construct stopping boundaries based on the statistical information estimated at each interim analysis in a simple and unified manner, but we can also assess whether the amount of information is what we would expect based on our initial guesses. In addition, we have the ability to estimate the nuisance parameters during the interim analysis and, if it seems that the original design will not meet the goals of the study, we may extend the study, increase the sample size, or make other adjustments that are necessary to obtain the desired information. As long as we are monitoring the data and deriving the stopping boundaries on the basis of statistical information, these adaptive changes will have no effect on the level and power of the group sequential test. If we are able to work out the logistics of keeping the clinical trial open until we have accumulated the desired information, then the power of the test to detect the clinically important difference will be achieved regardless of the underlying nuisance parameters. This would not be the case for strategies using the commonly practiced paradigm in which a study is terminated once the sample size of the original design is achieved. Although in this approach one can control the level of the test, the subsequent power of the test may be greatly affected if the values of the nuisance parameters are guessed incorrectly.

It is also reasonable to adopt information-based monitoring for sample size adjustment without any intention of early stopping. Since regulatory concerns preclude interim monitoring without spending some Type 1 error, one could make this approach operational by adopting extremely conservative stopping boundaries such as those of Haybittle (12). In this way there would be almost no likelihood of stopping early under the null hypothesis, but one would still have the opportunity to adjust sample size based on the values of the nuisance parameters and the amount of information actually accrued. In the limit, taking the interim boundaries to infinity would be equivalent to monitoring only for sample size adjustment.

REFERENCES

1. Kim K, Tsiatis AA. Study duration for clinical trials with survival response and early stopping rule. *Biometrics*. 1990;46:81–92.
2. Scharfstein DO, Tsiatis AA. The use of simulation and bootstrap in information-based group sequential studies. *Stat Med*. 1998;17:75–87.
3. Scharfstein DO, Tsiatis AA, Robins JM. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *J Am Stat Assoc*. 1997;92:1342–1350.
4. *EaSt: Software for design and interim monitoring of group sequential clinical trials*. Cambridge, MA: Cytel Software Corporation; 2000.
5. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*. 1987;43:193–199.
6. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35:549–556.
7. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977;64:191–199.
8. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70:659–663.
9. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *J R Stat Soc A*. 1969;132:232–244.
10. Self SG, Mauritsen RH, Ohara J. Power calculation for likelihood ratio tests in generalized linear models. *Biometrics*. 1992;48:31–39.
11. Facey KM. A sequential procedure for a Phase II efficacy trial in hypercholesterolemia. *Control Clinical Trials*. 1992;13:122–133.
12. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiology*. 1971;44: 793–797.