Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities

Sandro Pampallona[1], Anastasios A. Tsiatis[2], KyungMann Kim[3]

[1] ForMed, Statistics for Medicine
Les Chales
1983 Evolène
Switzerland

[2] Department of Statistics
North Carolina State University
Patterson Hall
Raleigh, NC 27695
USA

[3] Department of Biostatistics & Medical Informatics
University of Wisconsin
600 Highland Ave, K6/438 CSC
Madison, WI 53792-4675
USA

Address for correspondence

Sandro Pampallona
ForMed, Statsitics for Medicine
Les Chales
1983 Evolène
Switzerland
Tel+Fax (+41-27) 2833014
spampallona@atge.automail.com

Keywords

Group sequential trials, information time, error spending function.

Acknowledgements

This work has been supported in part by grants from the National Institutes of Health, DHHS, and the National Institute of Allergy And Infectious Diseases, grant number AI31789. Cytel Software Corporation is gratefully acknowledged for making available the EaSt 2000 Software which performs all of the calculations presented in this work.

Summary

Lan and DeMets (1983) introduced a flexible procedure for the analysis of sequential trials based on the discretization of Brownian motion. In this paper we consider an extension of this strategy that preserves both the desired significance level and the power of any group sequential trial. This entails the derivation of boundaries at monitoring stage by means of two spending functions, one for the type I and one for the type II error probabilities, as well as the adjustment of the target maximum information as the trial progresses. The general solution to the problem is provided together with a discussion of implementation strategies. The procedure is intended for group sequential designs that allow early stopping under both the null and the alternative hypotheses, and an example is presented for this case. However, its application is easily extended also for designs where there is no early stopping under the null.

Introduction

The application of the work by Armitage, McPherson and Rowe (1969) on repeated significance testing, led to the development of group sequential methods, intended to monitor accumulating data at regular intervals. The designs proposed by Pocock (1977), O'Brien and Fleming (1979), Wang and Tsiatis (1987), Emerson and Fleming (1989) and Pampallona and Tsiatis (1994), among others, are based on this approach. The standard strategy consists in deriving appropriate critical values that will guarantee the desired Type I and II error probabilities under repeated significance testing. Common to these plans is that they assume that the maximum number of analyses, K, be fixed in advance and also that interim analyses be equally spaced on the information scale. In most applications strict enforcement of either restriction may prove impractical.

Lan and DeMets (1983) have proposed a monitoring strategy that allows for any number and frequency of looks at the accumulating data. To each analysis a fraction of the pre-specified overall significance level is allocated according to a given spending function for the type I error probability. The aim of this contribution is to extend the Lan and DeMets strategy in order to guarantee control over the type II error probability as well. The simple approach proposed here does not modify the usual steps required when designing a group sequential study that adopts any of the standard families of boundaries. Rather, given that a study has been planned to have a fixed number of equally spaced analyses it allows relaxing these constraints at the monitoring stage.

Design of a standard group sequential study

The natural application of the monitoring strategy to be presented below is in the context of group sequential studies with boundaries that allow for early stopping either in favor of the null or of the alternative, as proposed by Emerson and Fleming (1989) or Pampallona and Tsiatis (1994). We shall introduce the minimum required notation without making reference to a specific family of boundaries. Suppose that we had actually designed a study, allowing for a maximum of K equally spaced analyses, based on a one sided test with overall significance level α. A unique set of (upper) standardized boundary values $\{u_j\}$, j=1,…, K, for early stopping in favor of the alternative, $H_1 : \eta = \eta_1$, and a corresponding set of (lower) standardized boundary values $\{l_j\}$ for early stopping in favor of the null , $H_0 : \eta = \eta_0$, would have been found such that at the last analysis $u_k = l_k$. This condition is required in order to reach a decision at the end of the study should none of the interim analyses have yielded a significant result. In order to detect the alternative of interest with power $1-\beta$, a given projected maximum information, say V_K, would be required.

Standard monitoring of a group sequential study

The analyses would have to be performed after every additional constant increment of information, $\frac{1}{K}V_K$, for a maximum of K analyses. If we let $Q(t_j)$ be the value of the normally distributed standardized statistic at the j[th] analysis, where $t_j = \frac{j}{K}V_K$, then the study would stop in favor of the alternative the first time $Q(t_j) > u_j$ or in favor of the null the first time $Q(t_j) < l_j$. If the standardized statistic fell in the continuation region, that is if $l_j \le Q(t_j) \le u_j$, j=1, …, K-1, a further analysis would be required.

An alternative monitoring strategy

If analyses were not performed according to the schedule specified at design then clearly the boundaries found at design would not apply to interim monitoring. For one-sided tests, when the monitoring schedule departs from the assumed K equally spaced analyses, we therefore propose the following testing strategy.

Derivation of boundary values. Assume that the first analysis was performed at $t_1^* = \frac{V_1}{V_K}$. Appropriate upper and lower boundary values would have to be found such that:

$$P_{\eta_0}\left(Q(t_1^*) \ge u_1^*\right) = \alpha(t_1^*)$$

$$P_{\eta_1}\left(Q(t_1^*) \le l_1^*\right) = \beta(t_1^*)$$

where $\alpha(t)$ and $\beta(t)$ are error probability spending functions, in the sense introduced by Lan and DeMets (1983). The spending functions describe the rate at which the error probabilities have to be portioned over successive analyses. In particular, at t=0 both functions are zero and at t=1 their value is α and β respectively. We shall see later how such spending functions can be empirically derived in order to respect the desired characteristics of the chosen design. The boundary values at subsequent analyses performed at $t_j^* = \frac{V_j}{V_K}$ will have to satisfy:

$$P_{\eta_0}\left(l_1^* < Q(t_1^*) < u_1^*, \ldots, l_{j-1}^* < Q(t_{j-1}^*) < u_{j-1}^*, Q(t_j^*) \ge u_j^*\right) = \alpha(t_j^*) - \alpha(t_{j-1}^*)$$

$$P_{\eta_1}\left(l_1^* < Q(t_1^*) < u_1^*, \ldots, l_{j-1}^* < Q(t_{j-1}^*) < u_{j-1}^*, Q(t_j^*) \le l_j^*\right) = \beta(t_j^*) - \beta(t_{j-1}^*)$$

Except for equations and the derivation of the boundary values has to be made using numerical recursive integration as described by Armitage, McPherson and Rowe (1969).

Derivation of empirical spending functions. The choice of the spending function can be dictated by how conservative the boundaries should be at the early versus the late analyses. Lan and DeMets, beside an spending function for the type I error which mimics the behavior of the O'Brien and Fleming designs, propose a function in the spirit of the Pocock designs and one that represents a way of spending the error probability uniformly over time. Kim and DeMets (1987) have studied the properties of other functions. We suggest to use empirical spending functions that retain the spirit of the boundaries selected at design. These can be derived as follows. At design stage, the following computations provide the amount of type I and type II error probabilities associated with each boundary value for each of the planned K equally spaced analyses occurring by design at $t_j = \frac{j}{K} V_K$, for j=1, …, K:

$$\alpha_j = P_{\eta_0}\left(l_1 < Q(t_1) < u_1, \ldots, l_{j-1} < Q(t_{j-1}) < u_{j-1}, Q(t_j) \ge u_j\right)$$

$$\beta_j = P_{\eta_1}\left(l_1 < Q(t_1) < u_1, \ldots, l_{j-1} < Q(t_{j-1}) < u_{j-1}, Q(t_j) \le l_j\right)$$

The cumulative error probabilities will thus be given by $\alpha_j^C = \sum_{i=1}^{j} \alpha_i$ and $\beta_j^C = \sum_{i=1}^{j} \beta_i$. A continuous function can be fitted to these cumulative errors, or even simple linear interpolation can be used between successive points. The fitted curves can be used to provide the type I and II error probabilities to be spent at the actual arbitrary

monitoring times, $t_j^* = \frac{V_j}{V_K}$, and thus to generate the boundaries according to equations to above. The resulting boundaries will enjoy approximately the same properties of the chosen family of designs. As a special case, if the actual sequence of analyses was exactly as anticipated at design, that is if $\{t_j^*\} \equiv \{t_j\}$ and if the empirical spending functions gave a perfect fit, then the boundaries obtained at monitoring stage would replicate precisely those considered at design. It should be noted that in practice group sequential studies are rarely designed to have more than 5 looks. If the design has been set up with a little K, then the same little number of points would be available to establish the empirical spending functions. We therefore suggest that, once the desired design has been chosen, the spending functions be established on the basis of an identical design for which the number of looks is set to 10. Curve fitting or linear interpolation would be much improved without loss of the salient features of the chosen design.

Positioning the next look. In the proposed strategy the actual monitoring schedule is arbitrary. With respect to the K equally spaced analyses considered for design purposes this entails that the monitoring schedule actually adopted may produce an underpowered or overpowered procedure. To avoid such situations we propose to adjust the projected maximum information. For a one sided test, before the study starts we suppose that the first analysis will be the last. If this was the case we would want to know with how much information it should be performed and what should be the boundary value to be used in order for the test to have the desired power and size. We would therefore solve the following equations for t_1^{Last} and $u_1^{Last} = l_1^{Last}$ respectively:

$$P_{\eta_0}\left(Q\left(t_1^{Last}\right) \geq u_1^{Last}\right) = \alpha$$

$$P_{\eta_1}\left(Q\left(t_{Last}^1\right) \leq u_1^{Last}\right) = \beta$$

It should be noted that by definition these computations will yield the solution corresponding to a fixed sample size study and in particular that $t_1^{Last} V_K$ will equal the information requirement for such a study. This is what we would expect since in a group sequential study it would not make sense to performed the first look with more resources than what would be required for a fixed sample size study. This information can be used as guidance for deciding when to perform the first analysis. For logistical reason it may nonetheless happen that $t_1^* > t_1^{Last}$, that is that $V_1 > t_1^{Last} V_K$. In this case the study will be overpowered but the size of the test can be maintained using the boundary value $u_1^* = l_1^*$ that satisfies:

$$P_{\eta_0}\left(Q\left(t_1^*\right) \geq u_1^*\right) = \alpha$$

If the first analysis will actually be performed at $t_1^* < t_1^{Last}$ then the boundary values, u_1^* and l_1^*, would be computed according to and . The procedure can continue in a similar way for any subsequent analysis. In particular, as concerns the j^{th} analysis, j>1, it would be advisable not to perform it t_j^{Last}, when the boundary

values to be used would be $u_j^{Last} = l_j^{Last}$. Both quantities can be found as solutions to:

$$P_{\eta_0}\left(l_1^* < Q(t_1^*) < u_1^*, \ldots, l_{j-1}^* < Q(t_{j-1}^*) < u_{j-1}^*, Q(t_j^{Last}) \geq u_j^{Last}\right) = \alpha - \alpha(t_j^*)$$

$$P_{\eta_1}\left(l_1^* < Q(t_1^*) < u_1^*, \ldots, l_{j-1}^* < Q(t_{j-1}^*) < u_{j-1}^*, Q(t_j^{Last}) \leq u_j^{Last}\right) = \beta - \beta(t_j^*).$$

Once again, if $t_j^* > t_j^{Last}$, that is if $V_j > t_j^{Last} V_K$, then the second analysis will be the last, the study will be overpowered but the size of the test can again be maintained using the boundary value $u_j^* = l_j^*$ that satisfies:

$$P_{\eta_0}\left(l_1^* < Q(t_1^*) < u_1^*, \ldots, l_{j-1}^* < Q(t_{j-1}^*) < u_{j-1}^*, Q(t_j^*) \geq u_j^*\right) = \alpha - \alpha(t_j^*).$$

If the j[th] analysis will actually be performed at $t_j^* < t_j^{Last}$ then the boundary values, u_j^* and l_j^*, would be computed as usual according to and . It should be noted that for consistency with the error spending functions, which are defined in the interval $t_j^* \in (0,1)$, only one analysis can be allowed at $t_j^* > 1$ and it must be the last. In this case the boundary values will again be computed using . It should also be noted that if the last analysis needs to be performed with less information that suggested by t_i^{Last} because of logistical reasons (e.g. accrual slower than initially anticipated) then the study will be underpowered but the type I error probability can still be maintained using .

Post-hoc power. When the null hypothesis is not rejected even at the last analysis, it may be of interest to know what the real power of the adopted procedure really was given the adopted monitoring schedule. Indeed, in most cases the last analysis will not be performed at exactly t_j^{Last} but either before (underpowered test) or after (overpowered test). For a one-sided test and when the last analysis is performed at t_j^*, the following computation provides the post-hoc power (PHP):

$$PHP = 1 - \left[\begin{array}{l} P_{\eta_1}\left(Q(t_1^*) < l_1^*\right) + P_{\eta_1}\left(l_1^* < Q(t_1^*) < u_1^*, Q(t_2^*) < l_2^*\right) + \ldots \\ \ldots + P_{\eta_1}\left(l_1^* < Q(t_1^*) < u_1^*, \ldots, l_{j-1}^* < Q(t_{j-1}^*) < u_{j-1}^*, Q(t_j^*) < l_j^*\right) \end{array}\right]$$

Extensions

Early stopping only in favor of the alternative. In this presentation attention has focused on designs that allow for early stopping both in favor of the alternative and of the null. However, the methods described can be applied also to the monitoring of designs allowing for early stopping only in favor of the alternative. In such cases, the whole of the type II error probability will be spent at the last look, that is $\beta(t) = 0$ for t<1 and $\beta(1) = \beta$.

Two sided tests. The methodology has been illustrated here for one-sided tests but the extension to two-sided tests is straightforward. In particular, at the first analysis, appropriate upper and lower boundary values would have to be found such that:

$$P_{\eta_0}\left(\left|Q\left(t_1^*\right)\right| \geq u_1^*\right) = \alpha\left(t_1^*\right)$$

$$P_{\eta_1}\left(\left|Q\left(t_1\right)\right| \leq l_1^*\right) = \beta\left(t_1\right)$$

The boundary values at subsequent analyses, j>1, will have to satisfy:

$$P_{\eta_0}\left(l_1^* < \left|Q\left(t_1^*\right)\right| < u_1^* \,,\, l_{j-1}^* < \left|Q\left(t_{j-1}^*\right)\right| < u_{j-1}^* \,,\, \left|Q\left(t_j^*\right)\right| \geq u_j^*\right) = \alpha\left(t_j^*\right) - \alpha\left(t_{j-1}^*\right)$$

$$P_{\eta_1}\left(l_1^* < \left|Q\left(t_1^*\right)\right| < u_1^* \,,\, l_{j-1}^* < \left|Q\left(t_{j-1}^*\right)\right| < u_{j-1}^* \,,\, \left|Q\left(t_j^*\right)\right| \leq u_j^*\right) = \beta\left(t_j^*\right) - \beta\left(t_{j-1}^*\right)$$

All other considerations would also apply to this situation with similar adaptations. We assumed above that the two-sided tests define a continuation region that is symmetric around the information axis, as it is for the Emerson and Fleming or the Pampallona and Tsiatis designs. If this is not the case the strategy would still apply though the equations defining the boundaries, beyond the scope of this presentation, would need to be further adapted.

Designing studies on the basis of spending functions. Although numerically more complex, the monitoring strategy given here can be naturally adapted to the design of group sequential studies. Suppose that the spending functions $\alpha(t)$ and $\beta(t)$ were given at design, together with a tentative arbitrary (i.e. not necessarily equally spaced) schedule of K analyses to be performed at information fractions $\{t_i^*\}$, say. Through numerical search the value of V_K could be found together with the sets $\{u_i^*\}$ and $\{l_i^*\}$ under the constraint that $u_K^* = u_K^*$. This approach would be more internally consistent since the boundary values obtained during study monitoring would not need to be based on empirical spending functions.

Example

The approach presented here can be applied to any of the common families of group sequential designs. For ease of presentation we shall refer to the boundaries proposed by Pampallona and Tsiatis (1994). These boundaries are indexed by a shape parameter which relates to the probability of stopping. In a typical randomized clinical trial comparing a control, C, to an experimental arm, E, on the basis of normally a distributed response, we might be interested in testing the null hypothesis $H_0 : \mu_E - \mu_C = 0$ versus the alternative hypothesis $H_1 : \mu_E - \mu_C = \delta$, for a known common variance of the observations, σ^2. In particular, suppose that under the alternative the

standardized difference was $\frac{\delta}{\sigma} = 0.25$. We are interested in a design with 4 looks (for design purposes to be assumed equally spaced) based on a one sided test with overall significance level α and shape parameter, $\Delta = 0$, that is in the spirit of the designs proposed by O'Brien and Fleming which require relatively large values of the test statistic to stop the trial at an early stage. The tables published in Pampallona and Tsiatis (1994) provide a maximum projected sample size of 600, that is $V_K = 600$, if the power is set to 90%. By design the monitoring schedule should be $\left\{\frac{150}{600}, \frac{300}{600}, \frac{450}{600}, \frac{600}{600}\right\}$. The tales in the paper also provide the corresponding set of upper boundary values, {3.372. 2.384, 1.947, 1.686}, and of lower boundary values, {-1.220, 0.220, 1.063, 1.686}. The design boundaries are displayed in Figure 1. We derive empirical spending functions to be used for monitoring purposes based on exactly the same design except for the number of looks which we set to 10. This produces cumulative type I and II error probabilities as displayed in Figure 2. Before performing the first analysis we would compute the fixed sample size requirement using and above, this gives $t_1^{Last} = 0.915$ or equivalently a sample size of 549. It would therefore be sensible to perform the first analysis before having recruited 549 patients onto the study. Suppose that the first analysis was actually performed with 225 patients, that is at $t_1^* = \frac{225}{600} = 0.375$, and with a standardized test statistic with value 0.6. Using and with $\alpha(t_1) = 0.00277$ and $\beta(t_1) = 0.0114$ we obtain $u_1^* = 2.767$ and $l_1^* = -0.396$. The test statistic falls in the continuation region so we can prepare for the second analysis. Application of and generates $t_2^{Last} = 0.925$ or equivalently a sample size of 555. Suppose that the second analysis was performed with 304 patients; that is, at $t_2^* = \frac{374}{600} = 0.623$, and with a standardized test statistic with value 1.2. Here $\alpha(t_2) = 0.01789$ and $\beta(t_2) = 0.0475$, corresponding to boundary values calculated according to and which are $u_2^* = 2.119$ and $l_2^* = 0.691$. The study should continue further. At this stage we have $t_3^{Last} = 0.97$ or equivalently a sample size of 582. Suppose that for logistical reason (e.g. slow accrual) the investigators together with the data monitoring committee decide to stop accrual at a time when 480 patients were on study. The last analysis is therefore performed at $t_3^* = 0.8$ with a value of the test statistic of 1.15. Since we are forcing the third look to be last and we want the procedure to have the desired significance level we shall use with the balance of type I error, that is $0.05 - \alpha(t_2^*) = 0.03211$. At this third and final look we have $u_3^* = l_3^* = 1.669$ and the procedure fails to reject the null hypothesis. Using we can compute the post hoc power which results in 0.857, below the desired power, as expected. Figure 1 also shows the actual boundaries used in this hypothetical study as well as the sample path of the observed test statistic. Is is worth noting that despite the actual schedule of analyses, $\{0.375, 0.623, 0.8\}$, the boundaries obtained through the proposed strategy are very close in spirit to the ones originally required by the adopted design.

Comment

The strategy suggested here extends the approach originally suggested by Lan and DeMets (1983) to designs that also allow for early stopping in favor of the null hypothesis. We have suggested a general strategy for the derivation of the monitoring boundaries assuming appropriate spending functions are available. Alternatively, we have proposed a method for deriving empirical spending functions that determine boundaries enjoying the same properties as the boundaries chosen for design purposes. Imposing the restriction that the upper and the lower boundaries meet at the last look removes the non-uniqueness of the solution otherwise inherent to the problem. This choice also allows the fine-tuning of the timing of the last look, in terms of the total information, in order for the desired overall error probabilities to be satisfied exactly. Such a feature is highly desirable: departures from the constraint of a fixed number of equally spaced analyses are frequent in most clinical trials; in the proposed approach assumptions made at the design stage not only do not constrain the actual analysis pattern and can be compensated for at the monitoring stage. The procedure suggested here also allows the declared type I error probability to be respected exactly and provides a useful post-hoc assessment of the power of inconclusive trials. The method has been illustrated here for one-sided hypothesis testing with boundaries allowing for early stopping either in favor of the null or of the alternative but can easily be adapted to two-sided tests and to tests allowing early stopping only in favor of the alternative. We believe that the strategy proposed here can help the practical realization of group sequential trials since the flexibility of the proposed approach does not entail any loss of rigor in the realization of clinical studies.

References

Armitage P., McPherson C.K. and Rowe B.C. (1969). Repeated significance tests on accumulating data. Journal of the Royal Statistical Society, A, 132:235-44.

Emerson S.S. and Fleming T.R. (1989). Symmetric group sequential test designs. Biometrics:45:905-23.

Kim K. and DeMets D.L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. Biometrika 74;149-54.

Lan K.K.G. and DeMets D.L. (1983). Discrete sequential boundaries for clinical trials. Biometrika 70;3:659-63.

O'Brien P.C. and Fleming T.R. (1979). A multiple testing procedure for clinical trials. Biometrics 35:459-56.

Pampallona S, Tsiatis AA. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favour of the null hypothesis. Journal of Statistical Planning and Inference 42: 19-35.

Pocock S.J. (1977). Group sequential methods in the design and analysis of clinical trials. Biometrika 64;2:191-9.

Wang S.K. and Tsiatis A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. Biometrics 43;193-199.

Figure 1. Design and monitoring boundaries for the example

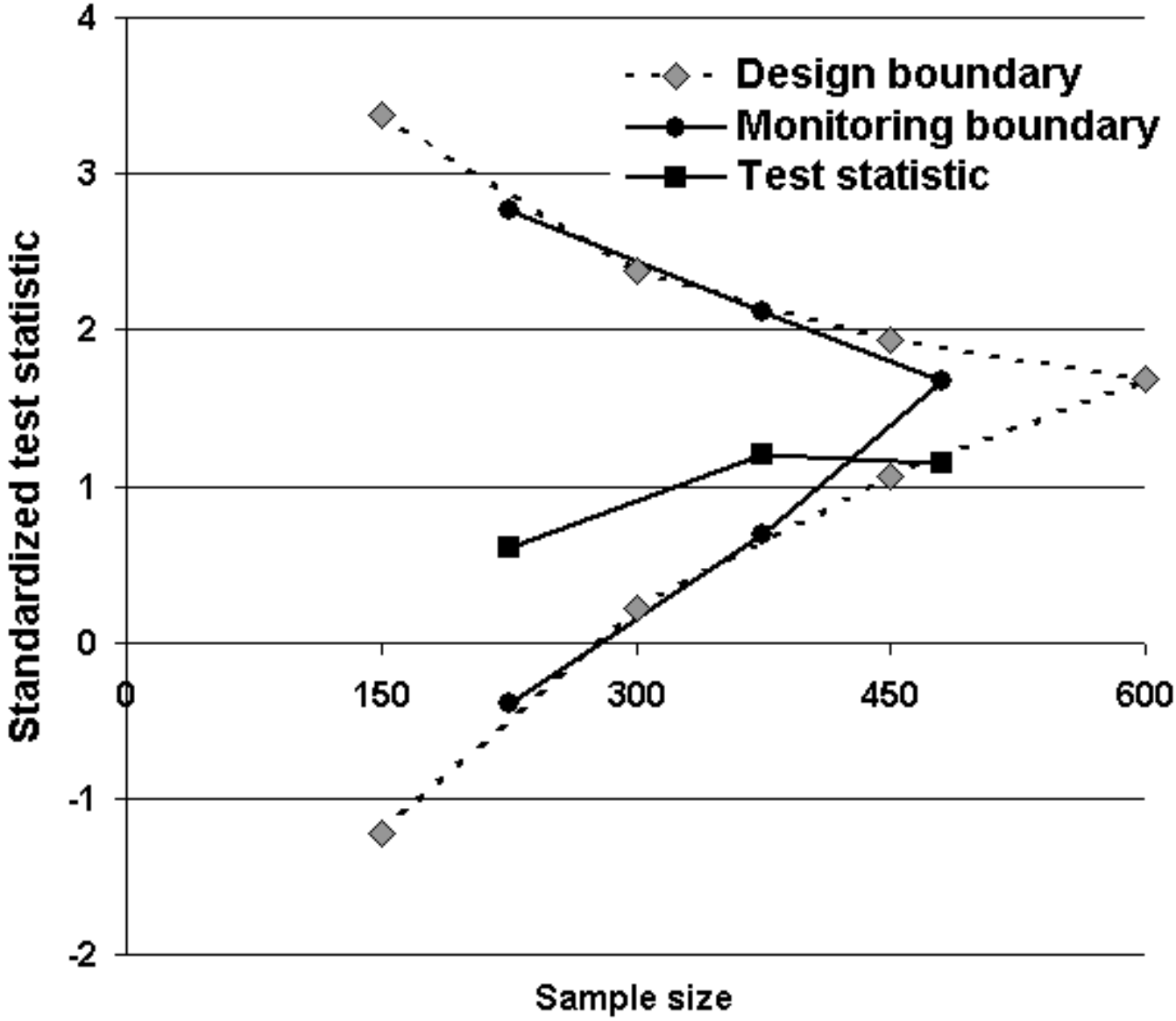

Figure 2. Error spending functions for the example.