Interim monitoring of clinical trials: decision theory, dynamic programming and optimal stopping

CHRISTOPHER JENNISON[1] AND BRUCE W. TURNBULL[2]

[1]*Department of Mathematical Sciences, University of Bath, Bath, U.K.*

[2]*Department of Operations Research and Information Engineering, Cornell University, Ithaca, U.S.A*

ABSTRACT

It is standard practice to monitor clinical trials with a view to stopping early if results are sufficiently compelling. We explain how the properties of stopping boundaries can be calculated numerically and how to optimise boundaries to minimise expected sample size while controlling type I and II error probabilities. Our optimisation method involves the use of dynamic programming to solve Bayes decision problems with no constraint on error rates. This conversion to an unconstrained problem is equivalent to using Lagrange multipliers. Applications of these methods in clinical trial design include the derivation of optimal adaptive designs in which future group sizes are allowed to depend on previously observed responses; designs which test both for superiority and non-inferiority; and group sequential tests which allow for a delay between treatment and response.

Keywords: Clinical trial; group sequential test; Bayes decision problem; dynamic programming; optimal stopping.

INTRODUCTION

It is natural to wish to examine data as they accumulate during the course of a long-term clinical trial. However, with frequent looks at the data, there is greater opportunity to make an erroneous decision. Armitage *et al*. (1969) report the overall type I error rate when applying repeated two-sided significance tests at $\alpha = 0.05$ to accumulating data and show this rises to 0.11 with 3 analyses and 0.14 with 5 analyses. Thus, special statistical methods are required to avoid inflation of the type I error rate due to over-interpretation of interim results.

Group sequential designs which require data to be analysed on a small number of occasions during the course of a study are well suited to clinical trials (Pocock, 1977). DeMets *et al*. (1984) report an early application of a group sequential clinical trial design in the Beta-Blocker Heart Attack Trial which compared propanolol with placebo. A stopping boundary of the form proposed by O'Brien & Fleming (1979) was employed and the trial stopped after the sixth of seven planned analyses. This stopping rule permitted early termination for a positive conclusion. In a retrospective analysis of 72 cancer studies conducted by the U.S. Eastern Co-operative Oncology Group, Rosner & Tsiatis (1989) found that, if group sequential stopping rules had been applied, the major benefit would have come from stopping early for a negative

outcome, with this occurring in around 80% of studies. Thus, a good clinical trial design should allow early termination for either positive or negative results.

Our interest is, therefore, in group sequential designs which achieve specified type I error rate and power and stop early, on average, under both null and alternative parameter values. In addition, it is desirable that optimised designs can be applied to a variety of response distributions to give flexibility of use in different types of study.

SEQUENTIAL DISTRIBUTION THEORY

The properties of a group sequential design depend on the joint distribution of the test statistics being monitored at each interim analysis. We consider first the simple example of a balanced two-sample problem with normal response. Here, responses X_{A1}, X_{A2}, ... from Treatment A and X_{B1}, X_{B2}, ... from Treatment B are observed sequentially. Suppose the $\{X_{Ai}\}$ and $\{X_{Bi}\}$ are independent and normally distributed with common variance σ^2 and means μ_A and μ_B, respectively. Then the "treatment effect" $\theta = \mu_A - \mu_B$ is the parameter of primary interest.

At interim analysis k $(k = 1, \ldots, K)$, the first n_k responses from each treatment arm are observed. The maximum likelihood estimate of θ at this analysis is

$$\widehat{\theta}_k = \sum_{i=1}^{n_k} (X_{Ai} - X_{Bi})/n_k$$

and this has the marginal distribution

$$\widehat{\theta}_k \sim N(\theta,\, I_k^{-1}),$$

where $I_k = n_k/(2\sigma^2)$ is the *Fisher information* for θ at analysis k.

The standardized test statistic based on the responses available at analysis k is

$$Z_k = \sum_{i=1}^{n_k} (X_{Ai} - X_{Bi})/(\sigma\sqrt{2n_k}) = \widehat{\theta}_k \sqrt{I_k}.$$

It is easy to check that the joint distribution of $Z_1, \ldots, Z_K$ has the defining properties

$$\begin{aligned} &\text{(i)} && (Z_1, \ldots, Z_K) \text{ is multivariate normal}, \\ &\text{(ii)} && Var(Z_k) = 1 \quad \text{and} \quad E(Z_k) = \theta\sqrt{I_k}, \quad k = 1, \ldots, K, \\ &\text{(iii)} && Cov(Z_{k_1}, Z_{k_2}) = \sqrt{(I_{k_1}/I_{k_2})}, \quad \text{for } 1 \le k_1 \le k_2 \le K. \end{aligned} \tag{1}$$

We refer to the set of properties (1) as the *canonical joint distribution* for a sequence of statistics $Z_1, \ldots, Z_K$ with information levels $\{I_1, \ldots, I_K\}$ for the parameter θ. In fact, Jennison & Turnbull (1997) and Scharfstein *et al*. (1997) show this joint distribution arises in a great many situations. Examples include: unbalanced two-sample comparisons; normal responses adjusted for baseline covariates; longitudinal data; parallel and crossover designs. The same canonical joint distribution also holds approximately for binary and survival data. For further details of how to construct $\{Z_k\}$ and $\{I_k\}$ sequences in specific applications, see Chapter 3 of Jennison &

Turnbull (2000). Our key conclusion is that we can build a unified theory of group sequential tests since properties of particular decision boundaries computed using (1) will be applicable to a wide variety of situations.

A PROBLEM OF OPTIMAL STOPPING

Consider a clinical trial where θ denotes the treatment effect and it is desired to test the null hypothesis H_0: $\theta \leq 0$ against the one-sided alternative $\theta > 0$ using a group sequential design with up to K analyses. The type I error rate is set at α under $\theta = 0$ and power $1-\beta$ is required when $\theta = \delta$. A fixed sample size test would need information for θ equal to

$$I_{fix} = \{\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)\}^2/\delta^2,$$

where Φ denotes the standard normal cumulative distribution function. In a group sequential design, the maximum information level has to be higher and we set this to be

$$I_{max} = R\, I_{fix}$$

for a chosen value $R > 1$. Assuming equal increments in information between analyses, we have

$$I_k = (k/K)\, I_{max}, \quad k = 1, \ldots, K.$$

Figure 1 illustrates a typical stopping boundary on the Z scale for a group sequential test with five analyses.

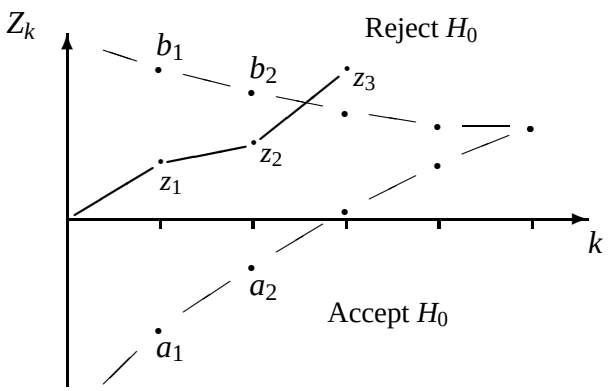

Fig. 1. Stopping boundary for a group sequential one-sided test with 5 analyses

The lower boundary points a_k and upper boundary points b_k are plotted for $k = 1, \ldots, 5$. Note that $a_5 = b_5$ to ensure a decision is reached at the final analysis. The example of a sample path stays within the continuation region at analyses 1 and 2, then crosses the upper boundary at analysis 3, resulting in termination of the trial to reject H_0 at this point.

We shall consider the problem of deriving a boundary satisfying the error rate requirements, with given values of R and K, which minimises

$$\{E_0(I) + E_\delta(I)\}/2, \tag{2}$$

where I denotes the level of information observed at termination. In our initial example of a two-treatment comparison with normal responses, information is proportional to sample size so minimising the expression (2) is equivalent to minimising the average of the expected sample sizes under $\theta = 0$ and $\theta = \delta$. In optimising the group sequential design we can choose the $2K-1$ boundary points freely subject to the constraints imposed by the error rate requirements under $\theta = 0$ and δ. This leaves a high dimensional space of possible boundaries in which to search. Before considering the optimisation problem, we discuss the calculation of properties for a particular boundary.

COMPUTATIONS FOR GROUP SEQUENTIAL TESTS

We need to be able to calculate the probabilities of basic events such as the outcome

$$a_1 < Z_1 < b_1, \;\; a_2 < Z_2 < b_2, \;\; Z_3 > b_3$$

illustrated in Figure 1. Combining such probabilities gives key properties, such as $P_\theta\{\text{Reject } H_0\}$. For a one-sided test with K analyses, define the events

$$\mathcal{A}_1 = \{Z_1 < a_1\}, \quad \mathcal{R}_1 = \{Z_1 > b_1\},$$

$$\mathcal{A}_k = \{a_1 < Z_1 < b_1, \; \ldots \; , a_{k-1} < Z_{k-1} < b_{k-1}, Z_k < a_k\}, \quad k = 2, \ldots, K,$$

and

$$\mathcal{R}_k = \{a_1 < Z_1 < b_1, \; \ldots \; , a_{k-1} < Z_{k-1} < b_{k-1}, Z_k > b_k\}, \quad k = 2, \ldots, K.$$

Then

$$P_\theta\{\text{Accept } H_0\} \;=\; P_\theta\{\mathcal{A}_1\} + \ldots + P_\theta\{\mathcal{A}_K\}, \;\; P_\theta\{\text{Reject } H_0\} \;=\; P_\theta\{\mathcal{R}_1\} + \ldots + P_\theta\{\mathcal{R}_K\}$$

and the observed information on termination is

$$E_\theta(I) \;=\; (P_\theta\{\mathcal{A}_1\} + P_\theta\{\mathcal{R}_1\}) \; I_1 \; + \; \ldots \; + \; (P_\theta\{\mathcal{A}_K\} + P_\theta\{\mathcal{R}_K\}) \; I_K.$$

Armitage *et al*. (1969) present recursive formulae for the densities of statistics at interim analyses. Working on the Z-statistic scale, the density $f_1(z_1)$ of Z_1 is that of a $N(\theta\sqrt{I_1}, 1)$ variate and the joint distribution of the Z_ks implies that

$$Z_2|Z_1 \;\sim\; N(\theta(I_2 - I_1)/\sqrt{I_2} + Z_1\sqrt{(I_1/I_2)}, (I_2 - I_1)/I_2).$$

We denote this conditional density by $f_2(z_2|z_1)$. Since analysis 2 is only reached if $a_1 < Z_1 < b_1$, the sub-density for Z_2 is

$$f_2(z_2) \;=\; \int_{a_1}^{b_1} f_1(z_1)\, f_2(z_2|z_1)\, dz_1.$$

In the general recursive step, the sub-density for Z_k at analysis k can be written as

$$f_k(z_k) \;=\; \int_{a_{k-1}}^{b_{k-1}} f_{k-1}(z_{k-1})\, f_k(z_k|z_{k-1})\, dz_{k-1},$$

where $f_k(z_k|z_{k-1})$ is the density of the distribution

$$N(\theta(I_k - I_{k-1})/\sqrt{I_k} + Z_{k-1}\sqrt{(I_{k-1}/I_k)},\, (I_k - I_{k-1})/I_k).$$

Numerical quadrature can be used to evaluate each of the functions f_1, f_2, etc., in succession on a grid of points. Hence, we can compute the probabilities of specific events, such as

$$P_\theta\{a_1 < Z_1 < b_1, a_2 < Z_2 < b_2, Z_3 > b_3\} \;=\; \int_{a_2}^{b_2} f_2(z_2)\,\Phi\left(\frac{\theta(I_3 - I_2) + z_2\sqrt{I_2} - b_3\sqrt{I_3}}{\sqrt{(I_3 - I_2)}}\right) dz_2.$$

As an alternative approach to the same calculations, we can write probabilities as nested integrals, for example,

$$P_\theta\{a_1 < Z_1 < b_1, a_2 < Z_2 < b_2, Z_3 > b_3\} \;=\; \int_{a_1}^{b_1}\int_{a_2}^{b_2}\int_{b_3}^{\infty} f_1(z_1)\; f_2(z_2|z_1)\; f_3(z_3|z_2)\, dz_3\, dz_2\, dz_1.$$

Applying numerical integration, we replace each integral by a sum of the form

$$\int_a^b f(z)\, dz \;=\; \sum_{i=1}^{n} w(i)\, f(z(i)),$$

where $z(1),\ldots,z(n)$ is a grid of points from a to b. Thus, we have

$$P_\theta\{a_1 < Z_1 < b_1, a_2 < Z_2 < b_2, Z_3 > b_3\} \;\approx$$
$$\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\sum_{i_3=1}^{n_3} w_1(i_1)\, f_1(z_1(i_1))\, w_2(i_2)\, f_2(z_2(i_2)|z_1(i_1))\, w_3(i_3)\, f_3(z_3(i_3)|z_2(i_2)).$$

Multiple integrations and summations arise in these calculations and for an outcome at analysis k we need to evaluate a k-fold sum of the form

$$\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\cdots\sum_{i_k=1}^{n_k} w_1(i_1)\, f_1(z_1(i_1))\, w_2(i_2)\, f_2(z_2(i_2)|z_1(i_1))\ldots\, w_k(i_k)\; f_k(z_k(i_k)|z_{k-1}(i_{k-1})).$$

However, the structure of the k nested summations is such that the computation required is of the order of $k-1$ double summations, much less than a general k-fold summation. We have found that using Simpson's rule with 100 to 200 grid points per integral gives probabilities to an accuracy of 5 or 6 decimal places. For details of sets of grid points that will provide accurate results efficiently, see Chapter 19 of Jennison & Turnbull (2000).

COMPUTING OPTIMAL GROUP SEQUENTIAL TESTS

We can now apply the methods of efficient computation for group sequential boundaries described in the previous section to derive optimal group sequential tests. Recall that we seek a test of H_0: $\theta \le 0$ against $\theta > 0$ with type I error rate α under $\theta = 0$ and power $1-\beta$ at $\theta = \delta$. Among all group sequential designs which achieve this using K analyses at information levels $I_k = (k/K)\, I_{max}$, $k = 1,\ldots,K$, where $I_{max} = R\, I_{fix}$, we seek the design minimising $\{E_0(I) + E_\delta(I)\}/2$.

Following Eales & Jennison (1992) and Barber & Jennison (2002), we deal with the constraints on error rates by introducing Lagrangian multipliers to create the *unconstrained problem* of minimising

$$\{E_0(I)+E_\delta(I)\}/2+\lambda_1 P_{\theta=0}\{\text{Reject } H_0\}+\lambda_2 P_{\theta=\delta}\{\text{Accept } H_0\}.$$

Once we have developed a method for solving this problem, we search for a pair of multipliers (λ_1,λ_2) such that the solution has type I and II error rates α and β, then this design solves the *constrained problem* too. The Lagrangian approach has a Bayesian interpretation. Suppose we put a prior distribution on θ with

$$P\{\theta=0\}=P\{\theta=\delta\}=0.5$$

and specify costs of: 1 per unit of information observed; $2\lambda_1$ for rejecting H_0 when $\theta=0$; and $2\lambda_2$ for accepting H_0 when $\theta=\delta$. Then, the total Bayes risk is

$$\{E_0(I)+E_\delta(I)\}/2+\lambda_1 P_{\theta=0}\{\text{Reject } H_0\}+\lambda_2 P_{\theta=\delta}\{\text{Accept } H_0\},$$

just as in the Lagrangian problem. The Bayes interpretation of the problem is useful in understanding how to solve it using the technique of "Dynamic Programming" or "Backwards Induction". For each $k=1,\ldots,K$, we denote the posterior distribution of θ given $Z_k=z_k$ at analysis k by $p^{(k)}(\theta|z_k)$ for $\theta=0$ and $\theta=\delta$. In applying dynamic programming to find the optimal Bayes rule, we work backwards from the final analysis as follows.

At analysis K

There is no further sampling cost once analysis K has been reached, so we simply compare the two possible decisions

$$\begin{aligned}\text{Reject } H_0:&\quad E(\text{Cost}) = \lambda_1\, p^{(K)}(0|z_K),\\ \text{Accept } H_0:&\quad E(\text{Cost}) = \lambda_2\, p^{(K)}(\delta|z_K).\end{aligned}$$

The boundary point a_K is the value of z_K where these expected costs are equal and the optimum decision rule at analysis K is to reject H_0 for $Z_K\geq a_K$ and to accept H_0 if $Z_K<a_K$.

At analysis $K-1$

We now know the optimal procedure to follow if we continue on to analysis K and we use this information in assessing that option. Consider an outcome in which the trial has continued to analysis $K-1$ where we observe $Z_{K-1}=z_{K-1}$, as shown in Figure 2.

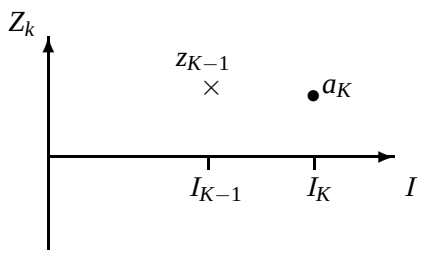

Fig. 2. Dynamic programming: State of the process at analysis $K-1$

If the trial is terminated at analysis $K-1$, there is no further cost of sampling and the expected additional costs for the two possible decisions are

$$\text{Reject } H_0\text{:} \quad E(\text{Cost}) = \lambda_1\, p^{(K-1)}(0|z_{K-1}),$$
$$\text{Accept } H_0\text{:} \quad E(\text{Cost}) = \lambda_2\, p^{(K-1)}(\delta|z_{K-1}).$$

If we let the trial continue on to analysis K, the expected additional cost is

$$1 \times (I_K - I_{K-1}) + \lambda_1\, p^{(K-1)}(0|z_{K-1})\, P_{\theta=0}\{Z_K \geq a_K | Z_{K-1} = z_{K-1}\}$$
$$+ \lambda_2\, p^{(K-1)}(\delta|z_{K-1})\, P_{\theta=\delta}\{Z_K < a_K | Z_{K-1} = z_{K-1}\}.$$

Equating the costs of pairs of decisions gives the optimal boundaries. The upper boundary point b_{K-1} is the value of z_{K-1} for which

$$E(\text{Cost of continuing}) = E(\text{Cost of stopping to reject } H_0)$$

and the lower boundary point a_{K-1} is the value of z_{K-1} where

$$E(\text{Cost of continuing}) = E(\text{Cost of stopping to accept } H_0).$$

After determining the optimal values of a_{K-1} and b_{K-1}, we set up a grid of points for use in numerical integration over the range a_{K-1} to b_{K-1}, as illustrated in Figure 3.

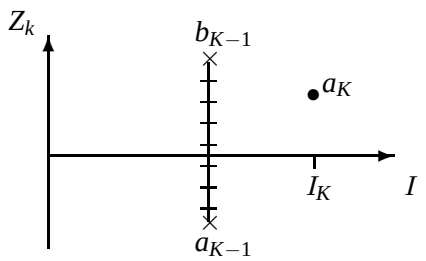

Fig. 3. Dynamic programming: Completed calculations for stage $K-1$

For each grid point z_{K-1}, we sum over the posterior distribution of θ to calculate

$$\beta^{(K-1)}(z_{K-1}) = E(\text{Additional cost when continuing to analysis } K \,|\, Z_{K-1} = z_{K-1})$$

and store this information. We are now ready to move back to analysis $K-2$.

At analysis $K-2$

Analysis $K-2$ has all the features of a generic analysis k. Calculating the expected additional cost when continuing on to the next analysis involves an integral over values z_{K-1} between a_{K-1} and b_{K-1}, but we have already set up a grid of points covering this interval, as seen in Figure 4, and stored values of the expected future cost $\beta^{(K-1)}(z_{K-1})$ on reaching z_{K-1} and proceeding optimally thereafter.

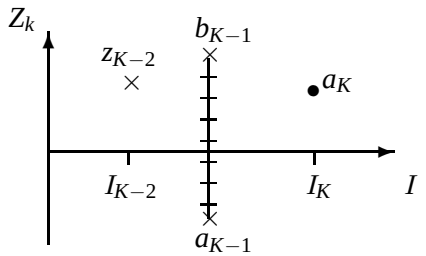

Fig. 4. Dynamic programming: State of the process at analysis $K-2$

If the trial is stopped at analysis $K-2$ with $Z_{K-2} = z_{K-2}$, the expected additional costs for the possible decisions are

$$\begin{aligned} \text{Reject } H_0\text{:} \quad E(\text{Cost}) &= \lambda_1\, p^{(K-2)}(0|z_{K-2}), \\ \text{Accept } H_0\text{:} \quad E(\text{Cost}) &= \lambda_2\, p^{(K-2)}(\delta|z_{K-2}). \end{aligned}$$

If the trial continues on to analysis $K-1$, the expected additional cost is

$$\begin{aligned} &1 \times (I_{K-1} - I_{K-2}) + \\ &\lambda_1\, p^{(K-2)}(0|z_{K-2})\, P_{\theta=0}\{Z_{K-1} > b_{K-1}|Z_{K-2} = z_{K-2}\} + \\ &\lambda_2\, p^{(K-2)}(\delta|z_{K-2})\, P_{\theta=\delta}\{Z_{K-1} < a_{K-1}|Z_{K-2} = z_{K-2}\} + \\ &\int_{a_{K-1}}^{b_{K-1}} \{p^{(K-2)}(0|z_{K-2})\, f_0^{(K-1)}(z_{K-1}|z_{K-2}) + \\ &\qquad p^{(K-2)}(\delta|z_{K-2}))\, f_\delta^{(K-1)}(z_{K-1}|z_{K-2})\}\, \beta^{(K-1)}(z_{K-1}) dz_{K-1}, \end{aligned}$$

where $f_\theta^{(K-1)}(z_{K-1}|z_{K-2})$ is the conditional density under θ of Z_{K-1} given $Z_{K-2} = z_{K-2}$. As before, equating the costs of the decisions to reject H_0 and to continue sampling gives the optimal upper boundary point b_{K-2} and equating the costs of accepting H_0 and continuing sampling gives the optimal lower boundary point a_{K-2}. It remains to set up a grid of points for use in numerical integration over the range a_{K-2} to b_{K-2} and calculate

$$\beta^{(K-2)}(z_{K-2}) \;=\; E(\text{Additional cost when continuing}\,|\,Z_{K-2} = z_{K-2})$$

at each of these points. The dynamic programming process then moves back to analysis $K-3$, and so on all the way back to analysis 1, at which point we have the full solution to our problem.

We can now return to the original problem of finding an optimal group sequential test with the specified type I and II error probabilities. Having set up a method of finding the Bayes optimal design for a particular pair of costs (λ_1, λ_2), we add another layer and search for a pair (λ_1, λ_2) such that the type I and type II error rates of the Bayes optimal design are α and β, respectively. The resulting design will be the optimal group sequential test, with the required frequentist error rates, for our original problem. It is important to remember that the output

of the dynamic programming routine will be fed into a numerical search algorithm, so results should not only be of high accuracy but also possess the continuity properties, etc., that the higher level search algorithm expects. This continuity requirement has implications for the definition of the grids of points used in numerical integration if discontinuities in the calculated values are to be avoided as the range of integration varies.

That the solution of a frequentist problem is found by solving a Bayes problem is in keeping with the general principle that good frequentist procedures should be similar to Bayes procedures. See Jennison & Turnbull (2006a) for further discussion of the relation between admissible group sequential procedures, in the frequentist sense, and solutions of Bayes problems.

The methods we have described are of broad applicability. In financial mathematics, dynamic programming arguments are commonly used to establish theory underlying the pricing of financial derivatives, but their use in direct computation of optimal strategies for executing an option has been more limited. In considering the translation of methods, it is important to note that the name "optimal stopping problem" is used with a specific meaning in probability theory: the quantity being optimised is a function of the sample path observed prior to the stopping time. This definition includes the unconstrained problem we have just solved but does not extend to the original problem of finding an optimal group sequential test with given type I and II error probabilities.

PROPERTIES OF OPTIMAL DESIGNS

Inspection of the properties of optimised designs shows the potential benefits of group sequential testing. As an example, consider one-sided tests with $\alpha = 0.025$, $1-\beta = 0.9$, a maximum of K equally spaced analyses and $I_{max} = R\, I_{fix}$. Table 1 presents the values of $\{E_0(I) + E_\delta(I)\}/2$ achieved by designs minimising this criterion for a variety of values of K and R.

Table 1. Minimum values of $\{E_0(I) + E_\delta(I)\}/2$ expressed as a percentage of I_{fix}

	R					*Minimum*
K	1.01	1.05	1.1	1.2	1.3	*over R*
2	80.8	74.7	73.2	73.7	75.8	73.0 at R =1.13
3	76.2	69.3	66.6	65.1	65.2	65.0 at R =1.23
5	72.2	65.2	62.2	59.8	59.0	58.8 at R =1.38
10	69.2	62.2	59.0	56.3	55.1	54.2 at R =1.6
20	67.8	60.6	57.5	54.6	53.3	51.7 at R =1.8

The results show that the minimised expected information (or, equivalently, expected sample size) decreases as the number of analyses K increases, but with diminishing returns. Similarly, expected information decreases with increasing values of R, up to a point. Given the costs

associated with conducting interim analyses and the desire to avoid too high a maximum sample size, designs with between 3 and 5 analyses and R around 1.05 or 1.1 appear attractive options.

The methods we have described can be applied with a variety of optimality criteria. We have used them to minimise general criteria of the form $\sum_i w_i E_{\theta_i}(I)$ or

$$\int f(\theta)\, E_\theta(I)\, d\theta$$

for a normal density $f(\theta)$. As well as providing specific designs directly, optimal procedures serve as benchmarks for other methods which may have additional useful features, for example, "error spending tests" which are designed to handle unpredictable information sequences.

RELATED PROBLEMS

The methods we have described for optimising a group sequential design can be applied to more general forms of group sequential procedure. We shall summarise three examples.

Adaptive choice of group sizes in a group sequential test

It is intuitive to think that a group sequential design might benefit from taking a smaller group size when the current test statistic lies close to the stopping boundary and a larger group size when the test statistic is mid-way between the boundaries. Schmitz (1993) proposed such procedures in which the group sizes are chosen adaptively. These designs are most easily defined in terms of the score statistics $S_k = Z_k\sqrt{I_k}$, $k = 1, \ldots, K$. For the first group of subjects, I_1 is fixed and we observe

$$S_1 \sim N(\theta\, I_1,\, I_1).$$

The next group size, and hence I_2, is then chosen as a function of S_1 and the statistic S_2 is observed. The increment $S_2 - S_1$ is conditionally independent of S_1 given I_2 and

$$S_2 - S_1 \,|\, I_2 \sim N(\,\theta(I_2 - I_1),\, (I_2 - I_1)\,).$$

The procedure continues with data-dependent choice of each I_k until stopping occurs with a decision to accept or reject H_0, or the final analysis K is reached. The sampling rule and stopping rule are pre-specified and defined so as to achieve the desired overall type I error rate and power.

Various methods have recently been proposed for modifying sample size during the course of a clinical trial in response to interim estimates of the treatment effect. The paper of Cui *et al*. (1999) addresses the problem of "rescuing" an under-powered study but other authors have recommended this approach as a prospective strategy for dealing with uncertainty about the likely treatment effect when planning a study. The resulting "adaptive" methods fall into the general class of procedures proposed by Schmitz (1993).

Jennison & Turnbull (2006a) derive optimal versions of these adaptive group sequential tests in order to assess the efficiency gains they can offer. Their findings are disappointing. Measuring expected sample size as a percentage of that required in a fixed sample size design,

optimal adaptive designs improve on the efficiency of optimal non-adaptive group sequential tests with equal group sizes by about 2 percentage points. If the group sizes of the non-adaptive test are allowed to be unequal, but still fixed in advance, this difference reduces to about 1 percentage point.

The positive message is that standard group sequential designs offer a simple and efficient methodology for interim monitoring of clinical trials and their properties cannot be significantly improved on by more complex adaptive designs. Trials can be designed to achieve power over the range of effect sizes of possible interest: if the treatment effect is particularly high, this is likely to lead to early stopping for a positive conclusion and a smaller sample size (Jennison & Turnbull, 2006b).

Testing for either superiority or non-inferiority

When an accepted treatment for a medical condition is already available, it is not appropriate to test a new treatment against placebo. In comparing a new treatment against an active control, there are two types of positive outcome: the new treatment may be shown to be *superior* to the current standard; or the new treatment may be shown to be *non-inferior* to the standard. Demonstrating non-inferiority is achieved by rejecting a null hypothesis of the form $H_{0,NI}$: $\theta \leq -d$ in favour of $\theta > -d$, where θ is a measure of the difference in effect between the new treatment and the standard and d is the accepted "non-inferiority margin", which should be set (and agreed with regulators) before the trial begins.

Adaptive trial designs have been proposed for such a situation. If a trial is instigated with the intention of demonstrating superiority of a new treatment over the standard, this goal may be adapted to proving non-inferiority if results are not as good as anticipated. The fact that there are two null hypotheses is not an issue since these are nested: the null hypothesis for non-inferiority, $H_{0,NI}$: $\theta \leq -d$, is a subset of that for superiority, $H_{0,S}$: $\theta \leq 0$. A more important issue is that the two hypothesis tests may require different sample sizes. Wang *et al*. (2001) note the non-inferiority margin d is often smaller than the effect size $\theta = \delta$ at which power for declaring superiority is specified and, hence, a larger sample size is needed to give adequate power at $\theta = 0$ in the test for non-inferiority. Thus, if early data indicate that the key issue is to test for non-inferiority, there may be reason to increase the trial's sample size at an interim stage.

However, a non-adaptive group sequential approach is also possible. Öhrn & Jennison (2010) embed tests for both superiority and non-inferiority in a group sequential design with fixed group sizes. The example of a stopping boundary displayed in Figure 5 shows three outcomes are possible: to reject $H_{0,S}$: $\theta \leq 0$ (establishing superiority); to reject $H_{0,NI}$ but not $H_{0,S}$ (showing non-inferiority only); or to accept $H_{0,NI}$: $\theta \leq -d$ (failing even to show non-inferiority).

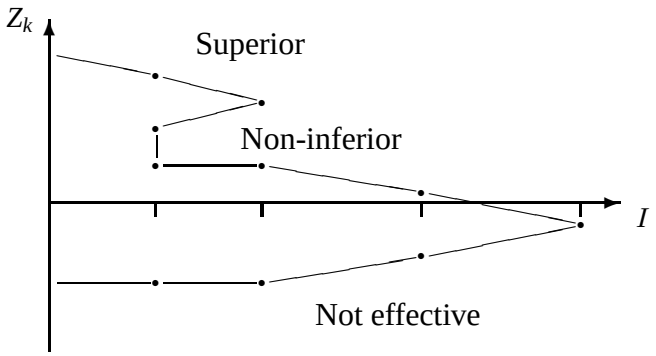

Fig. 5. A four-stage group sequential design to test for both superiority and non-inferiority

Early stopping is allowed for each outcome. It is significant that the lower arm of the continuation region, which involves discrimination between no effect and non-inferiority, is longer than the upper arm which tests between non-inferiority and superiority. Öhrn & Jennison (2010) derive tests which minimise expected sample size while satisfying two type I error rate constraints and two power requirements. This is achieved by defining related Bayes decision problems, solving these by dynamic programming, and searching for a set of costs such that the optimal procedure has the specified error rates and power.

Group sequential tests for a delayed response

In many trials, the clinical response is measured some time after each patient is randomised and the allocated treatment administered. Delays can also occur while validating and analysing responses. Thus, after a group sequential test stops, additional data will accrue from "pipeline" subjects who have entered the study but not yet responded.

In her PhD thesis, Hampson (2009) presents a framework for group sequential testing which recognises a delay in observing responses and models this appropriately. Formally, termination of the trial proceeds in two stages: first, recruitment of new patients ceases; then, after waiting to observe responses from all subjects enrolled at this time, a final decision is made. Again, it is possible to derive an optimal design which minimises a stated efficiency criterion by creating related Bayes problems and solving these by dynamic programming. A search for the costs in the Bayes decision problem that produce a procedure with the required type I error rate and power gives the optimal frequentist design for a delayed-response.

Examination of optimised designs reveals the extent to which the benefits of lower expected sample sizes usually provided by group sequential tests are reduced when response is subject to delay. However, there may be opportunities to recover these benefits. For example, if a second, more rapidly observed endpoint has a high correlation with the primary endpoint, a stopping rule based on the joint analysis of this pair of endpoints can mitigate the effects of the delay in observing the primary endpoint. A full account of this work is presented by Hampson & Jennison (2013).

CONCLUSIONS

We have seen that the monitoring of clinical trials poses a range of problems of statistical inference and optimal design. A general distribution theory gives a basis for generic methodology with wide applicability. Moreover, efficient computational methods, based on iterated numerical integration, are available to calculate properties of group sequential clinical trial designs.

The optimisation of a group sequential test for a specific sample size criterion is an important issue. Such problems can be tackled by using Dynamic Programming to solve related Bayes decision problems and searching for a set of costs so that the optimal Bayes procedure also solves the original problem with frequentist error rate constraints. The examples we have presented illustrate the versatility of this methodology for tackling a variety of problems of practical significance.

REFERENCES

Armitage, P., McPherson, C.K., & Rowe, B.C. 1969. Repeated significance tests on accumulating data. Journal of the Royal Statistical Society, Series A 132: 235–244.

Barber, S. & Jennison, C. 2002. Optimal asymmetric one-sided group sequential tests. Biometrika 89: 49–60.

Cui, L., Hung, H.M.J., & Wang, S-J. 1999. Modification of sample size in group sequential clinical trials. Biometrics 55: 853–857.

DeMets, D.L., Hardy, R., Friedman, L.M., & Lan, K.K.G. 1984. Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. Controlled Clinical Trials 5: 362–372.

Eales, J.D. & Jennison, C. 1992. An improved method for deriving optimal one-sided group sequential tests. Biometrika 79: 13–24.

Hampson, L.V. 2009. Group Sequential Tests for Delayed Responses. Ph.D. thesis, University of Bath, Bath, UK.

Hampson, L.V. & Jennison, C. 2013. Group sequential tests for delayed responses (with discussion). Journal of the Royal Statistical Society, Series B 75: 3–54.

Jennison, C. & Turnbull, B.W. 1997. Group sequential analysis incorporating covariate information. Journal of the American Statistical Association 92: 1330–1341.

Jennison, C. & Turnbull, B.W. 2000. Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall/CRC: Boca Raton.

Jennison, C. & Turnbull, B.W. 2006a. Adaptive and nonadaptive group sequential tests. Biometrika 93: 1–21.

Jennison, C. & Turnbull, B.W. 2006b. Efficient group sequential designs when there are several effect sizes under consideration. Statistics in Medicine 25: 917–932.

O'Brien, P.C. & Fleming, T.R. 1979. A multiple testing procedure for clinical trials. Biometrics 35: 549–556.

Öhrn, F. & Jennison, C. 2010. Optimal group sequential designs for simultaneous testing of superiority and non-inferiority. Statistics in Medicine 29: 743–759.

Pocock, S.J. 1977. Group sequential methods in the design and analysis of clinical trials. Biometrika 64: 191–199.

Rosner, G.L. & Tsiatis, A.A. 1989. The impact that group sequential tests would have made on ECOG clinical trials. Statistics in Medicine 8: 505–516.

Scharfstein, D.O., Tsiatis, A.A. & Robins, J.M. 1977. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. Journal of the American Statistical Association 92: 1342–1350.

Schmitz, N. 1993. Optimal Sequentially Planned Decision Procedures. Lecture Notes in Statistics, 79, Springer-Verlag: New York.

Wang, S.J., Hung, H.M.J., Tsong Y., & Cui, L. 2001. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. Statistics in Medicine 20: 1903–1912.