Point of view

Some aspects of the design and monitoring of clinical trials

The pre-eminence of the randomised clinical trial to assess medical interventions has been firmly established, although there has been some recent disquiet.[1,2] It would seem appropriate then, to review some aspects of trial design and termination in the light of new developments. By way of introduction, we highlight some details from a trial in the critically ill which was prematurely stopped.

The ALVEOLI trial[3] compared two ventilatory strategies in patients with acute lung injury and acute respiratory distress syndrome (ALI/ARDS): (a) higher end-expiratory lung volume/lower F_IO_2 (HEELV/LF_IO_2) versus (b) lower end-expiratory lung volume/higher FIO_2 (LEELV/HF_IO_2), the latter ventilatory strategy being identical to that of the treatment group in the initial ARDS Network trial.[4] The trial was stopped for lack of efficacy in early 2002, based upon a protocol specified futility stopping boundary.[5] The trialists had planned to enroll a maximum of 750 patients (power 89%) over 2-3 years assuming a 10% mortality difference between the LEELV/HF_IO_2 group (mortality 28%) and the HEELV/LF_IO_2 group (mortality 18%) with 89% power to detect such a difference. Two interim analyses were planned at 250 and 500 subjects. The trial stopping criteria for (i) efficacy, were O'Brien-Fleming boundaries,[6] with one-sided p = 0.025, and (ii) futility were, at the first interim analysis, a mortality in the HEELV/LF_IO_2 group greater than that observed in the LEELV/HF_IO_2 group and, at the second interim analysis, a mortality in the HEELV/LF_IO_2 group not at least 2% better than the LEELV/HF_IO_2 group. If there were no true difference in mortality between the study groups, the chances of stopping at the first and second interim analyses were 50% and 24% respectively. An "informal" statement of the statistical requirements of ARDS Network trials has been provided by one of the co-authors of the initial ARDS Network trial,[7] who noted that the clinical consequences of such trial designs would be to reduce the cost of long-term drug development (by early stopping of futile trials) and guard against harm to patients. This would seem eminently suitable for trials in critically patients, as opposed to say, cardiovascular or cancer trials, where long term outcomes are of substantive interest[8] (although recent recommendations for critically ill patients enrolled in trials suggest that they should be followed for ≥ 90 days).[9]

What does this mean for the clinician? The use of interim analyses and efficacy stopping boundaries are not uncommon in clinical trials and we have commented upon this previously in the journal.[10] It would therefore seem appropriate to extend these observations and canvass some wider questions of monitoring and estimation in clinical trials.

History

The paradigm encompassing the randomised clinical trial has developed over a relatively short period.[11,12] Notable watersheds were:

(i) the impact of randomised allocation.[13] Although randomisation was introduced into agricultural science by RA Fisher in 1926, it was not formally adopted into medical trials until the 1930's by Amberson and then, in the 1940's with the work of Hill.[14,15] It has been suggested that the early history of experimental design may have been different if RA Fisher[16] had not been employed initially in agricultural research, where experiments are essentially non-sequential in nature.[17]

(ii) sequential analysis. The sequential nature of medical and industrial experimentation leads to the accumulation of data over time and the formulation of appropriate statistical approaches to sequential or interim analysis of such data, arising from the early work of Wald in the industrial sphere,[18,19] has been crucial.[17,20,21]

(iii) Data monitoring committees (DMCs).[22,23] The progressive integration of DMCs into the trial scenario has complemented the developing statistical approaches, to the extent that decisions regarding the termination or otherwise of trials are the result of a constellation of factors.[24,25] This is exemplified by subsequent published reports of the deliberations of the DMCs in two early pivotal randomised controlled trials in the 1970's (the University Group Diabetes Project[26] and the Coronary Drug Project[27]) and, of more interest to critical care practitioners, the 1987 Department of Veteran's Affairs Cooperative study of steroid therapy for systemic sepsis.[28]

An understanding of "futility", as it applies to clinical trials, requires consideration of both sequential analysis and design in clinical trials and the role of the DMC in the context of the question: when to stop a clinical trial?[29,30] Some attention has been directed to these questions in the intensive care literature, although the focus has been more general.[31]

Definitions

If there are two groups of patients (n in each arm) in a trial, allocated to treatments A and B, then the interest is in testing the null hypothesis of no treatment difference H_0: $u_A = u_B$, where u = the mean response (continuous or binary), against the (two-sided) alternative $u_A \neq u_B$, with a Type I and Type II error probability of α and β respectively (power being defined as [1-β]). These error probabilities are defined at a particular value of $(u_A - u_B) = \pm \delta$, where δ = treatment difference or effect. The standardised Z statistic can be used to assess this difference and it can be further shown that:

(i) $Z \sim N((u_A - u_B)\sqrt{\{n/(2\sigma^2)\}}, 1)$, where n is total number, σ^2 = variance (and 1= SD of a standardised variable) and

(ii) the necessary (fixed) sample size (per arm), $n_f(\alpha, \beta, \delta, \sigma^2) = \{\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\}^2 2\sigma^2 / \delta^2$, where Φ is the standard normal cumulative distribution function, α is the Type I error and β is the Type II error.[32]

We are also interested in specific rules or tests to stop the trial at an early stage for (a) efficacy, or in a worst case scenario, effect reversal and (b) futility, which we will define as the conditional probability that a clinical trial will fail to demonstrate the superiority of a new therapy given the accumulated data and projected sample size of the study.

Controversy

The question of one- or two-sided (tailed) tests in clinical trials has engendered, perhaps not surprisingly, much controversy. One-sided tests have greater power for rejecting the null hypothesis when the one-sided alternative applies, but for alternatives on the opposite side, there is less sensitivity. The larger sample size required of a two-sided test (“modest” increase only for some[33]) for the same power as a one-sided test in the same direction, is obviously offset by the power for the alternative in the opposite direction.

For a one-sided test assessing H_0: $u_A - u_B = \delta = 0$ vs H_s: $u_A - u_B = \delta > 0$, the power is:

$1-\Phi\{Z_{1-\alpha} - (\delta/\sigma_d)\}$, where σ_d is the pooled variance.

The two-sided power is:

$[1-\Phi\{Z_{1-(\alpha/2)} - (\delta/\sigma_d)\} + \Phi\{Z_{\alpha/2} - (\delta/\sigma_d)\}]$.

Arguments for one-sided tests (at both the 0.05 and 0.025 level) in mainly drug-placebo trials, where the alternative hypothesis being tested is one-sided, have been advanced,[34-36] but also vigorously contested,[33,37] especially in the light of the experiences of The Cardiac Arrhythmia Suppression Trial (CAST) trial. The CAST trial, in the belief that antiarrhythmics were (only) beneficial, was designed as a one-sided trial, albeit at the conservative 0.025 level, to assess “beneficial or .. no beneficial effect….(it)..was not designed to prove that an antiarrhythmic drug could cause harm”,[38] which in fact occurred, much to the surprise of the investigators. Currently, in line with conservative regulatory requirements (e.g. FDA), most trials are conducted with two-sided tests.

Sequential analysis

The effect of repeated “looks” at accumulating data, in terms of inflation of the Type I error has been known for some time,[39] but needs to be re-iterated:

> For a pre-defined number of data inspections (say, 5) a nominal p value level of ≤ 0.0159 must be obtained in any of the 5 tests for the results to be “truly” significant at ≤ 0.05.[40] The same may be said somewhat differently: a trial will continue if the test statistic[41] does not exceed some critical value (for example, the standardised Z statistic with an α level = 0.05, is 1.96*)*. If this critical value was used for each inspection of accumulating trial data (interim analyses), the probability of stopping would be 0.05 for the first test, 0.14 for the 5th and 0.19 for the 10th.[21]

That this is a real concern was convincingly demonstrated by the profile of the z statistic over the course of the Coronary Drug Project trial: the magnitude of the statistic fluctuated markedly between efficacy ($z < -2$) and null effect ($z \equiv 0$), but the final mortality curves (clofibrate- versus placebo-treated patients) were almost identical.[11,27]

The multiple and diverse patient monitoring requirements in clinical trials (detection of treatment trends and toxicities, minimisation of patient numbers, ethical concerns) mandate interim analyses of trial data, and a sophisticated statistical analytic apparatus, incorporating at its core the above insights, has been developed.

Sequential design

Classical sequential designs, introduced by Wald,[19] used the likelihood ratio statistic (as opposed to the more familiar standardised Z statistic), but could not prescribe exact sample sizes, merely that the experiment would stop (“open plan”).[11,21] The extension of “fully” sequential designs to the medical arena, initially by Armitage[42] (“closed plan”), required effective continual assessment of patients (or rather, patient pairs) and the test statistic (the size of which reflected the treatment effect magnitude) used to estimate this treatment effect was recalculated after each new outcome (hence increased as the trial progresses) and compared with certain criteria to control Type I error.[43] Continual patient assessment is obviously burdensome and for the most part impracticable, except in single institution studies,[44] and fully sequential designs have seen limited application. A notable exception was the relatively small (N = 196) MADIT trial (implanted defibrillator for high risk ventricular arrhythmias) which used weekly assessments after the first 10 deaths.[45] The test statistic (log-rank)[46] was used to terminate the trial when it "crossed" one of the preset termination boundaries (efficacy (+ve value of the log-rank statistic), inefficacy (-ve value of the log-rank statistic) or null effect (values close to zero)) as determined by a two-sided triangular sequential design, proposed by Whitehead.[47] Similarly, the trial of prolonged methyprednisolone in unresolving acute respiratory distress syndrome,[48] a 4 institution single city study, used a one-sided triangular test at the 0.05 level (power 0.95) to demonstrate the superiority of drug to placebo. The boundaries for the triangular test are the score test statistic $S_k = Z_k\sqrt{I_k}$, where Z is the standardised statistic and I is the accrued information (see below).

Group sequential design

The seeming impracticality of the fully sequential design led to the development, initially by Pocock[49] and O'Brien & Fleming,[6] of group sequential designs where data is analysed at intervals; the number of analyses being a function of factors such as anticipated sample size and recruitment rates (although in practice there appears little to be gained from more than 5 interim analyses).[18,50] A maximum number (K) of patient blocks is determined: for example, in a 2 armed trial that would enroll 400 patients (200 in each arm) according to a fixed sample scenario, it may be decided to have 4 groups of 100 patients and interim analyses (of the appropriate outcome(s)) would then occur at 100, 200 300 and 400 patients. The standardised statistic Z_k is computed over the k blocks as data accumulates. Rejection of H_0 and trial termination occurs if $| Z_k | >$ critical value (c), there being a sequence of critical values $\{c_1,\ldots, c_K\}$ for each particular design, whereas continuation of the trial occurs if $| Z_k | < c$. Type I and II error probabilities are preserved under repeated testing.

The approach adopted by Pocock[49] was essentially one of a repeated significance test with a constant nominal significance level α', such that, for a given number of analyses (of *equal* patient number), the overall trial significance level will be α. So, for K = 5 analyses and α = 0.05 (two-sided), Z'_K = 2.413. The total sample size (assuming no stopping) needed for Pocock boundaries is in excess of the fixed sample design and is a function of a constant, R_p, which is defined for particular values of the parameters K, α and β; formally, $R_p(K,\alpha,\beta)$. That is, total $N = (N_{\text{fixed sample}} \times R_p)$. At a power (1-β) = 0.8 and (two-sided) α = 0.05, R_p = 1, for K = 1; R_p = 1.137 for K = 3 and R_p = 1.187 for K = 5.[32]

The O'Brien-Fleming[6] boundaries are such that the nominal significance level (for rejection of H_0) increases with data accumulation. Thus rejection is difficult early in the trial and becomes progressively easier; the critical value of the last test (K) is approximately the same as if a single test were done. The standardised statistic is computed as: $c_k = C_B(K, \alpha)\sqrt{(K/k)}$, where the constant $C_B(K, \alpha)$ ensures overall Type I error probability. Total sample size requirements for O'Brien-Fleming boundaries are less than for Pocock, but still somewhat greater (again, by a function R_B) than for a fixed sample design, again assuming no early stopping: at a power (1-β) = 0.8 and (two-sided) α = 0.05, R_B = 1, for K = 1; R_B = 1.007 for K = 3 and R_B = 1.015 for K = 5.[32]

For both the Pocock and O'Brien-Fleming designs, given that there is a possibility to stop (increased at early analyses for the Pocock compared with the O'Brien-Fleming design), the expected number of patients (average sample number, ASN = ∑(probability of significance at jth test)×(jn)) is less than that of a fixed sample design, especially for large treatment effects (δ), although ASN for the Pocock design is uniformly greater than for O'Brien-Fleming.

We see the classic configuration of both these designs, using the normalised Z statistic scale, in Figure 1 for a two sided trial. The sequential design, 4 interim analyses and stopping only for the alternative hypothesis, has been applied to a trial postulating a reduction in mortality from 50% to 35% (-15%) with α = 0.05 and power (1-β) = 0.9. The critical boundary (Z) values of the two designs at each analysis are indicated by the small circles (Pocock) and small plus (O'Brien-Fleming) at the end of the vertical lines on the graph. Although the boundaries are discontinuous, they have been joined by horizontal solid lines (for the Pocock design) and sloping dotted lines for the O'Brien design to aid in visualisation. Efficacy (a reduction in mortality) would be accompanied by a negative value of the Z statistic and the trial would (potentially) stop at any of the interim analyses if the Z statistic were less than these values. Thus the (trial) continuation regions are the "white-space" areas outside the boundaries. As seen in the upper panel, stopping is possible earlier with the Pocock design but at the cost of inflation of the sample size. The sample size for a fixed design of same power is indicated by a vertical dash-dot line. In the lower panel, the ASN is seen to be less for O'Brien-Fleming design (left insert), and the 75th percentile of the sample size distribution (over many possible values of the true treatment effect) is also less (right insert).

Wang and Tsiatis[51] proposed a "power" family[32] of two-sided tests, indexed by a parameter Δ that determines the shape of the continuation region; Δ = 1 is equivalent to the fixed sample size scenario, high values of Δ entail higher probabilities of early stopping, and low (0.3, 0.4) ensure minimal ASN. The standardised statistic is computed as: $c_k = C_{WT}(K, \alpha, \Delta)\ (k/K)^{\Delta-1/2}$ and the Pocock and O'Brien-Fleming designs are special cases. If $\Delta = 0.5$, the Pocock design results; $\Delta = 0$ gives the O'Brien-Fleming design. Unequal group sizes may be also accommodated, with the O'Brien-Fleming and Wang and Tsiatis (with low values of Δ) designs being robust to early group size variation.

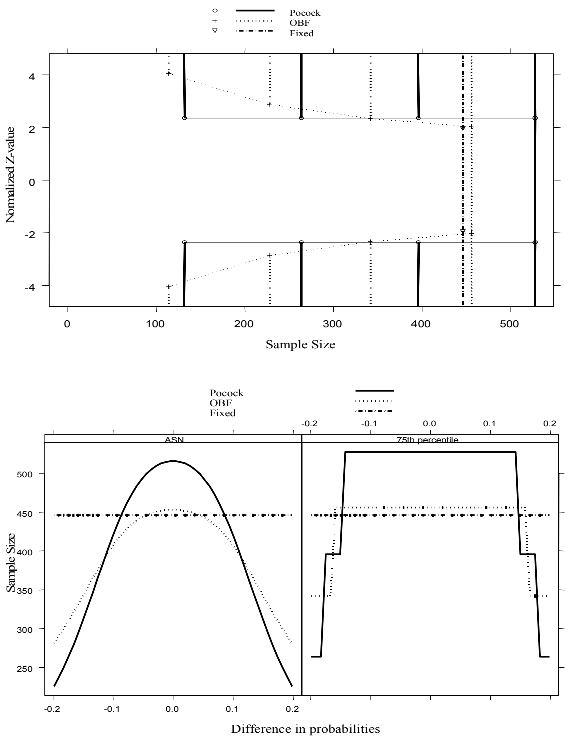

Figure 1. Upper panel. Pocock and O'Brien-Fleming group sequential boundaries for the designed trial. Horizontal axis: sample size. Vertical axis: normalized Z statistic. Interim analyses are indicated by vertical lines (solid, Pocock design; dot, O'Brien-Fleming) and critical boundary values of Z for each interim analysis are indicated by the small circle (Pocock) and small plus (O'Brien-Fleming). Boundaries are discontinuous, but have been joined to aid in visualization. Vertical dash-dot line indicates the sample size for a fixed design of same power. OBF = O'Brien-Fleming. Fixed = fixed sample size for same power.
Lower panel. Left insert: ASN is seen for both the Pocock and O'Brien-Fleming designs. Vertical axis: sample size; Horizontal axis: sample mean scale of difference in probabilities of the treatment effect (efficacy associated with –ve probabilities). Right insert: Sample efficiency for Pocock and O'Brien-Fleming designs at 75th percentile of the sample size distribution (over many possible values of the true treatment effect)

For trials comparing outcomes with survival analytic methods, the log-rank or a generalisation of the Wilcoxon test are appropriate statistics, although interim analyses are done after certain numbers of events have occurred, rather than patients enrolled; again unequal event number may be tolerated.[52]

An alternative less formal approach had been introduced by Haybittle[53] and Peto[8] who suggested conservative criteria at each interim analyses ($z = 3.09$: that is, for k analyses, setting $\alpha_1 = \alpha_2 .. = \alpha_{k-1} \equiv 0.001$). At the final ($k^{th}$) analysis, $\alpha_k = 0.05$ and the experimental error is nearly 0.05;[54] that is, providing the number of looks is not great, a fixed sample analysis can be undertaken at the final stage with no allowance for interim analyses.[32] Any inflation of the α error may be avoided by application of the Bonferroni inequality to the final stage,[55] such that, for example, after 5 interim analyses (at $p \leq 0.001$), the final p value, for a one-sided test $\alpha = 0.025$, would be 0.02 (z boundary of 2.05). A formal application of this strategy using the exact distribution of the test statistic has also been described.[54] A conceptual problem with this type of rule is that a z value of 2.9 at the k-1 analysis does not suggest early termination, but is striking at the kth analysis.[21] Haybittle-Peto boundaries were used in the EMAIT trial (European Myocardial Infarct Amiodarone Trial)[56] and will be used for the ongoing Australian and New Zealand Intensive Care Society (ANZICS) sponsored SAFE (saline versus albumin fluid evaluation) study.[57]

The nominal requirement of group sequential methods to specify equal numbers of patients or events (for survival analyses) in advance was formally overcome by the α spending function of Lan & DeMets,[58,59] which produced flexible discrete boundaries by the specification of an increasing function $\alpha(t^*)$ which characterises the rate at which the α error is "spent" over the interim analyses. The exact number and/or timing of interim analyses need not be specified in advance, although the total sample size does. For any calendar time t, a certain fraction of the total information t^* is observed ($0 < t^* < 1$, is given as the ratio of the inverse of the variance of the test statistic at any interim analysis and the final analysis). At trial beginning, $t^* = 0$ and $\alpha(t^*) = 0$; at trial end $t^* = 1$ and $\alpha(t^*) = \alpha$. Information fractions (of patients or events) may be variously defined (for example, 0.2, 0.4, 0.6, 0.8, 1.0) and critical test statistic values can be appropriately computed, using the techniques of Armitage *et al*.[39] The specification of "information" may be problematic and trials have been categorised as:

(i) maximal information: maximal information is prescribed and trial termination is at either boundary crossing or maximum information and
(ii) maximal duration: maximum trial duration is pre-specified and (maximum) information is estimated, based upon trial design; current information is estimated on empiric grounds (current interim analysis) or on calendar time fractions.[58,60]

The approximate O'Brien-Fleming spending function is defined as $\alpha_1(t^*) = 2\text{-}2\Phi(Z_{\alpha/2} / \sqrt{t^*})$ and the Pocock, $\alpha_2(t^*) = \alpha \times \ln(1 + (e\text{-}1)t^*)$; other spending functions have also been described.[11] The difference in the rate of "spending" of the α (Type I error) over information fractions (0 through 1) for the two different functions, $\alpha_1(t^*)$ and $\alpha_2(t^*)$, is seen clearly in Figure 2 for a total (two sided) $\alpha = 0.05$. Of interest, the PROWESS study utilised "the O'Brien-Fleming spending function according to the method of Lan and DeMets".[61]

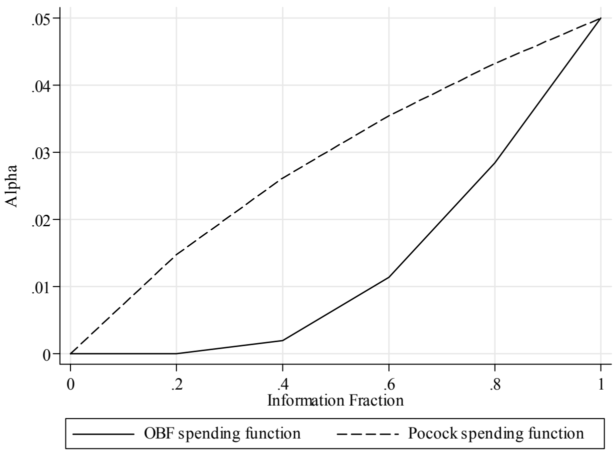

Figure 2. Vertical axis: Type I error (α). Horizontal axis: information fraction

In classical group sequential designs, originally described as two-sided tests, if the null hypothesis was in fact not rejected, then patient enrollment would occur until the final (K^{th}) analysis with probability of at least 1-α to satisfy Type I error constraints. Further refinement of group sequential design saw the introduction of one-sided tests, both one and two sided tests for early stopping under the null[32,60,62-65] and asymmetric boundaries.[11,22,29] Figure 3 displays, in the upper panel, a two-sided design with stopping for both the null and the alternative for the same trial conditions as above in Figure 1, using both the O'Brien-Fleming and Pocock boundaries. The stopping boundaries for the null are seen as the so-called "inner-wedge";[32] stopping for the null occurs earlier (potentially) for the Pocock design. This being so, it is perhaps not surprising to see in the lower panel, that the ASN for both the Pocock and O'Brien-Fleming design is less than the fixed sample size and that the ASN for the Pocock is the less than the O'Brien-Fleming (left insert). The 75th percentile of the sample distribution is also greater for the O'Brien-Fleming design (right insert).

Stochastic curtailment

The progressive review of trial data may be analysed for a trend (positive, negative or null) by estimating what is termed the Conditional Power or the probability that, given current information, the trial will yield at endpoint a "significant" result.[66] Two well known instances of this were the Beta-Blocker Heart Attack Trial (BHAT), for positive effect,[67] and a trial of prophylactic barbiturate coma in head injury for the null.[68,69] If the conditional power under the *alternate* hypothesis is say, ≤ 0.15, given the information at the interim analysis (for a two sided test, minimum recommended t = 0.64), then consideration should be given to terminating the trial for futility.[44,70] This stopping may result in a loss of power, but the loss is not substantial. If a trial is designed to have a (unconditional) power of 90%, the Type II error (β) = 0.1; if with curtailment, the (conditional) power is computed as 0.15 and the trial is stopped, the overall ("true") Type II error probability has found to be β / γ = 0.8(1-0.15) = 0.12, (where γ =1- conditional power).[71] Similarly, the overall Type I error is given by α / γ. Curtailment boundaries may be generated and it is of interest that the O'Brien-Fleming and the 50% Conditional Power boundaries are coincident.

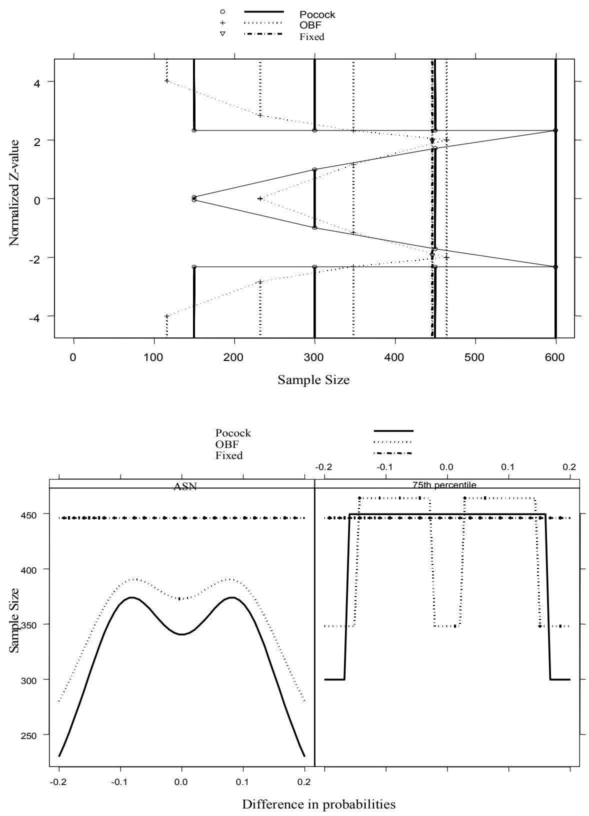

Figure 3. Upper panel. Pocock and O'Brien-Fleming (OBF) group sequential boundaries for the designed trial. Horizontal axis: sample size. Vertical axis: normalized Z statistic. Interim analyses are indicated by vertical lines (solid, Pocock design; dot, O'Brien-Fleming) and critical boundary values of Z for each interim analysis are indicated by the small circle (Pocock) and small plus sign (O'Brien-Fleming). Boundaries are discontinuous, but have been joined to aid in visualization. Fixed sample size for equal power ("Fixed") is seen by the vertical dash-dot line. Null boundaries are seen as the "inner wedge".
Lower panel. Left insert: ASN for Pocock (solid line) and O'Brien-Fleming (OBF = dotted line) and fixed sample size seen as horizontal dash-dot line. Right insert: sample size for the 75th percentile of the sample size distribution (over many possible values of the true treatment effect).

In the presence of low unconditional (that is, at initial design stage) power and, for time-to-event outcomes with non-proportional hazards, aggressive futility monitoring and consequent early stopping may be problematic.[72] A second approach to stochastic curtailment, Predictive futility, has a Bayesian flavor; a prior distribution for the treatment effect is specified and, given the data, the posterior distribution is then computed.[73]

Inference

It is well known that early stopping of a trial introduces bias into the usual point estimates and confidence intervals of the various maximum likelihood estimators of treatment effect (for example, risk ratio, odds ratio). A number of revised estimators have been described and the bias adjusted mean would appear to have advantageous properties.[74] Suffice it to say that the reason for this bias is that the sampling distribution of the estimator, under the conditions of interim analysis, has an asymmetric jagged profile, quite unlike the smooth profile of standard distributions.[32,75,76] We have previously commented upon both this and the lack of use of these adjusted estimators in trial reports.[10]

A unified approach

The above seeming plethora of group sequential designs was simplified with the introduction of the “Unified Family” by Kittelson and Emerson,[77] resulting in a hybrid design incorporating aspects of both equivalence and superiority test designs. The particular insight was that the various methods involving seemingly different boundary functions (standardised *Z* statistic, partial sum statistic (used in the triangular test), alpha spending function, maximum likelihood estimate of treatment effect (sample mean scale), stochastic curtailment (conditional and predictive power)) are transformations of each other.[25,75] The unified family, described initially on the sample mean scale which is invariant to hypothesis shifts, includes the above *formal* designs and allows a continuum between the four basic boundary designs, as seen in Figure 4, which we generate by adapting code given in the S+SeqTrial2 User’s guide.

The top panel uses the normalised Z scale and the bottom, the sample mean scale; again negative vales of the Z statistic and negative mean differences are associated with efficacy; O’Brien-Fleming boundaries are used. In clock-wise direction (repeated in each panel): a single boundary design for a one-sided test (early stopping against the null, or non rejection of the null); two boundary design for a one-sided test (early stopping for or against the null); four boundary design for two-sided hypothesis (early stopping for or against the null); and two boundary design for a two-sided test (early stopping against the null, but decision for the null only at final analysis).[78]

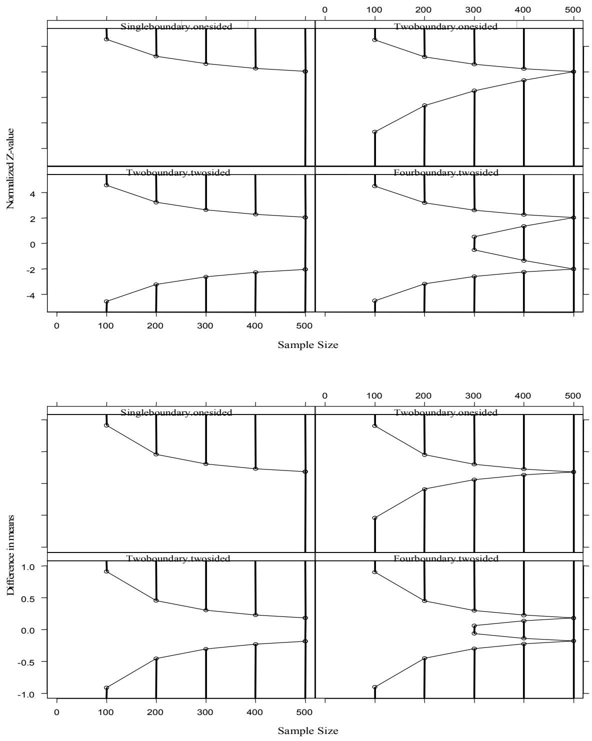

Figure 4. Top panel: O’Brien-Fleming boundaries for the 4 basic boundary designs (see Text). Horizontal axis: sample size. Vertical axis: normalised *Z* statistic. Bottom panel: Same design using the sample mean scale. Horizontal axis: sample size. Vertical axis: Difference in means. Efficacy is associated with negative values of the *Z* statistic and difference in means

This innovation allows assessment of a spectrum of designs and “connection” between distinct families Figure 5 shows the same trial as considered as in Figures 1 and 3, but with a one-sided test and early stopping for both the null and alternate hypothesis. The null boundaries (upper panel) have been progressively changed to allow earlier stopping (for futility) by manipulation of the Δ parameter in the Wang-Tsiatis power family, from a standard O’Brien-Fleming boundary through boundaries approaching the Pocock. That this is potentially advantageous is seen in the lower panel where ASN for the various designs is plotted; the design “Alt: OBF; null P = 0.6” which has a Δ value in the Wang & Tsiatis “power series” approaching the Pocock design has obvious reduced sample size as the probability difference approaches the null.

It was mentioned above (*Stochastic curtailment*) that the O’Brien-Fleming design and the 50% Conditional Power boundaries were coincident. Using the facility of the Unified Family, we are able to change scale and see this for the initial O’Brien-Fleming design for the two-sided trial illustrated in Figure 1. The lower and upper boundaries (there is no stopping for the null) are "a" and "d" respectively (Table 1) and the conditional probability (*computed at the particular boundary*) is 0.5 that at the last analysis the estimated treatment effect would correspond to an *opposite* decision.

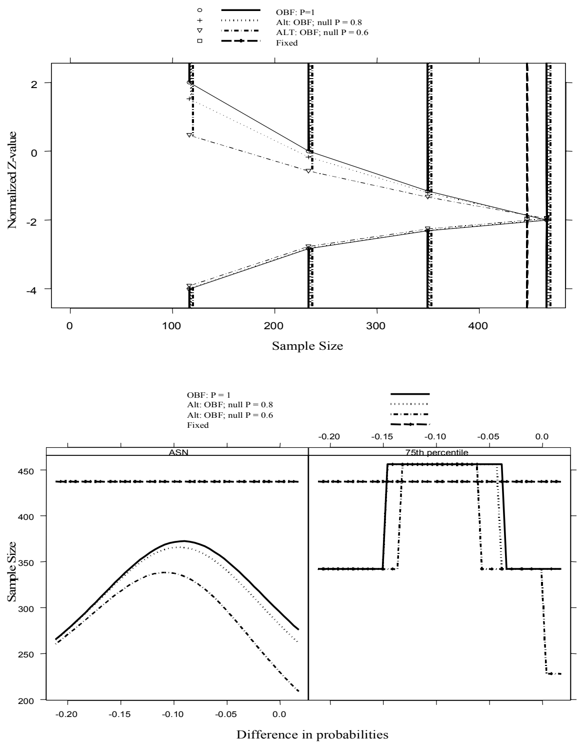

Figure 5. Upper panel: Horizontal axis: sample size. Vertical axis: normalized *Z* statistic. Efficacy boundaries are the lower (O'Brien-Fleming) boundaries, which are co-incident. Null boundaries are the upper boundaries which are three in total: O'Brien-Fleming P = 1 (solid line), "null P = 0.8" (dotted line) and "null P = 0.6" (dash-dot line). The "P" values are specific to the implementation in S+SeqTrial2, but indicate progressive movement away from O'Brien-Fleming boundaries (P = 1) to Pocock boundaries (P = 0.5). OBF = O'Brien-Fleming. Alt = Alternative hypothesis (efficacy in a one-sided test). Vertical long-dashed line indicates the fixed-sample ("Fixed") size for equal power.
Bottom panel. Left insert: ASN for three one-sided designs in the Upper panel. O,Brien-Fleming P = 1 (solid line), "null P = 0.8" (dotted line) and "null P = 0.6" (dash-dot line). Right insert: sample size for the 75th percentile of the sample size distribution (over many possible values of the true treatment effect). Vertical axis: sample size. Horizontal axis: difference in probabilities (null effect indicated by low or zero difference).

Table 1. Stopping boundaries: Conditional probability scale

	a	b	c	d
Time 1 (N= 114)	0.5	NA	NA	0.5
Time 2 (N= 228)	0.5	NA	NA	0.5
Time 3 (N= 342)	0.5	NA	NA	0.5
Time 4 (N= 456)	0.5	0.5	0.5	0.5

Therefore stopping occurs if the conditional power is < 50%, a rather elevated level. However, if we assume that the treatment effect is the *current best estimate* (and here we follow the discussion in the S+SeqTrial2 Manual, Chapter 6),[78] we can revisit the boundaries on this scale, as seen in table 2.

Table 2. Stopping boundaries: Conditional probability scale

	a	b	c	d
Time 1 (N= 114)	0.0000	NA	NA	0.0000
Time 2 (N= 228)	0.0021	NA	NA	0.0021
Time 3 (N= 342)	0.0886	NA	NA	0.0886
Time 4 (N= 456)	0.5000	0.5	0.5	0.5000

We now see, looking at the boundaries "a" and "b" for the early analyses, that the probability of a reverse decision at the final analysis is small, which is consistent with the known early conservatism of the O'Brien-Fleming design.

Software implementation

The above "Unified family" approach has seen implementation in S+SEQTRIAL;[78] alternate software packages are those of East[79] and, from the sequential perspective, PEST.[80] The latter two software programs, in earlier versions, have been the subject of formal review.[81] Public domain routines to implement error-spending tests are also available.[82]

Conclusions

The specific statistical approaches described above are an aid to the "real-world" concerns of trial conduct.[83] The termination of trials, for benefit, harm or the null, is a complex procedure. The ramifications of these decisions may have a life of their own, as witnessed by the 10 year debate of the decisions of the DMC in the University Group Diabetes Project trial.[26] The factors recommending the adoption of a particular method of interim analysis perhaps reduce to the specifications of the individual trial, but some general determinants can be offered: the type of patient (critically ill or otherwise) and disease (acute or chronic); the nature of the intervention (life saving or equivalence testing); the extent of follow-up (short term 28 day mortality or long term observation); the need to define the "history" of the treated disease and the secondary end-points / toxicities; the estimate of the treatment effect;[84] and any pressing requirement for early recognition of toxicity or null effect.[85] For the ARDS Network at least, stopping for the null effect would appear to have assumed a priority. Current statistical developments in the design of randomised clinical trials allow a large degree of flexibility which will aid in the tailoring of appropriate designs for the individual clinical trial.

Technical note: All group sequential design graphics were produced using S+SeqTrial2 as a module running in S-Plus 6.1 Professional Edition, Release 1.

Acknowledgments

CSIRO Mathematical & Information Sciences Division (http;//www.cmis.csiro.au/S-PLUS/) and Ms Sue Clancy for the acquisition of S+SeqTrial2 module.

J. L. MORAN
Intensive Care Unit, Queen Elizabeth Hospital, Woodville, SOUTH AUSTRALIA

P. J. SOLOMON
School of Applied Mathematics, University of Adelaide, Adelaide, SOUTH AUSTRALIA

REFERENCES

1. Britton A, McKee, M, Black, M, McPherson, K, Sanderson, C, Bain, C. Choosing between randomised and non-randomised studies: a systematic review. Health Technol Assess 1998;2:1-123.
2. Ioannidis JP, Haidich, A B, Lau, J. Any casualties in the clash of randomised and observational evidence? BMJ 2001;322:879-880.
3. ARDS Network. Report: Prospective, Randomized, Multi-Center Trial of Higher End-expiratory Lung Volume/Lower FiO2 versus Lower End-expiratory Lung Volume/Higher FiO2 Ventilation in Acute Lung Injury and Acute Respiratory Distress Syndrome. http://hedwig.mgh.harvard.edu/ardsnet/ards04.html
4. The ARDS Network Authors for the ARDS Network. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. The Acute Respiratory Distress Syndrome Network. N Engl J Med 2000;342:1301-1308.
5. ARDS Network. Protocol: Prospective, Randomized, Multi-Center Trial of Higher End-expiratory Lung Volume/Lower FiO2 versus Lower End-expiratory Lung Volume/Higher FiO2 Ventilation in Acute Lung Injury and Acute Respiratory Distress Syndrome. http://hedwig.mgh.harvard.edu/ardsnet/alveoli.pdf
6. O'Brien PC, Fleming, T R. A multiple testing procedure for clinical trials. Biometrics 1979;35:549-556.
7. Schoenfeld DA. A simple algorithm for designing group sequential clinical trials. Biometrics 2001;57:972-974.
8. Peto R, Pike, M C, Armitage, P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 1976;34:585-612.
9. Cohen J, Guyatt, G, Bernard, G R, et al. New strategies for clinical trials in patients with sepsis and septic shock. Crit Care Med 2001;29:880-886.
10. Moran JL, Peake, S L, Solomon, P J. Reporting of clinical trials using group sequential methods. Critical Care and Resuscitation 2001;3:146-147.
11. Friedman, L. M., Furberg, C. D., and DeMets, D. L. Fundamentals of Clinical Trials. 3rd ed. New York: Springer-Verlag; 1998.
12. Meldrum ML. A brief history of the randomized controlled trial. From oranges and lemons to the gold standard. Hematol Oncol Clin North Am 2000;14:745-760.
13. Pocock SJ. Allocation of patients to treatment in clinical trials. Biometrics 1979;35:183-197.
14. Doll R. Controlled trials: the 1948 watershed. BMJ 1998;317:1217-1220.
15. Fisher LD. Advances in clinical trials in the twentieth century. Ann Rev Pub Health 1999;20:109-124.
16. Box GEP. Science and statistics. Journal of the American Statistical Association 1976;71:791-799.
17. Armitage P. Interim analysis in clinical trials. Stat Med 1991;10:925-935.
18. McPherson K. Sequential stopping rules in clinical trials. Stat Med 1990;9:595-600.
19. Wald, A. Sequential analysis. New York: John Wiley & Sons; 1947.
20. DeMets DL. Stopping guidelines vs stopping rules: A practitioner's point of view. Communications in Statistics - Theory and Methods 1984;13:2395-2417.
21. DeMets DL, Lan, K K. An overview of sequential methods and their application in clinical trials. Communications in Statistics - Theory and Methods 1984;13:2315-2338.
22. Califf RM, Ellenberg, S S. Statistical approaches and policies for the operations of Data and Safety Monitoring Committees. Am Heart J 2001;141:301-305.
23. Ellenberg SS. Independent data monitoring committees: rationale, operations and controversies. Stat Med 2001;20:2573-2583.
24. Organization, review, and administration of cooperative studies (Greenberg Report): a report from the Heart Special Project Committee to the National Advisory Heart Council, May 1967. Control Clin Trials 1988;9:137-148.
25. Ellenberg, S. S., Fleming, T. R., and DeMets, D. L. Data Monitoring Committees in Clinical Trials : A Practical Perspective. Hoboken, NJ: John Wiley & Sons; 2002.
26. Kolata EB. Controversy over study of diabetes drugs continues for nearly a decade. Science 1979;203:986-990.
27. Practical aspects of decision making in clinical trials: the coronary drug project as a case study. The Coronary Drug Project Research Group. Control Clin Trials 1981;1:363-376.
28. Peduzzi P. Termination of the Department of Veterans Affairs Cooperative Study of steroid therapy for systemic sepsis. Control Clin Trials 1991;12:395-407.
29. DeMets DL, Pocock, S J, Julian, D G. The agonising negative trend in monitoring of clinical trials. Lancet 1999;354:1983-1988.
30. Pocock SJ. When to stop a clinical trial. BMJ 1992;305:235-240.
31. Hebert PC, Cook, D J, Wells, G, Marshall, J. The design of randomized clinical trials in critically ill patients. Chest 2002;121:1290-1300.

32. Jennison, C. and Turnbull, B. W. Group Sequential Methods with applications to clinical trials. Boca Raton: Chapman & Hall/CRC; 2000.
33. Moye LA, Tita, A T. Defending the rationale for the two-tailed test in clinical research. Circulation 2002;105:3062-3065.
34. Fisher LD. The use of one-sided tests in drug trials: an FDA advisory committee member's perspective. J Biopharm Stat 1991;1:151-156.
35. Knottnerus JA, Bouter, L M. The ethics of sample size: two-sided testing and one-sided thinking. J Clin Epidemiol 2001;54:109-110.
36. Overall JE. Tests of one-sided versus two-sided hypotheses in placebo-controlled clinical trials. Neuropsychopharmacology 1990;3:233-235.
37. Fleiss JL. One-tailed versus two-tailed tests: Rebuttal. Control Clin Trials 1988;10:227-228.
38. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. N Engl J Med 1989;321:406-412.
39. Armitage P, McPherson, C K, Rowe, B C. Repeated significance tests on accumulating data. Journal of the Royal Statistical Society Series A 1969;132:235-244.
40. McPherson K. Statistics: the problem of examining accumulating data more than once. N Engl J Med 1974;290:501-502.
41. Fleming TR, Green, S J, Harrington, D P. Considerations for monitoring and evaluating treatment effects in clinical trials. Control Clin Trials 1984;5:55-66.
42. Armitage, P. Sequential Medical Trials. 2nd ed. New York: John Wiley & Sons; 1975.
43. DeMets DL. Sequential designs in clinical trials. Cardiac Electrophysiology Review 1998;2:57-60.
44. Ware JH, Muller, J E, Braunwald, E. The futility index. An approach to the cost-effective termination of randomized clinical trials. Am J Med 1985;78:635-643.
45. Moss AJ, Hall, W J, Cannom, D S, et al. Improved survival with an implanted defibrillator in patients with coronary disease at high risk for ventricular arrhythmia. Multicenter Automatic Defibrillator Implantation Trial Investigators. N Engl J Med 1996;335:1933-1940.
46. Mantel N. Evaluation of survival data and two new rank-order statistics arising in its consideration. Cancer Chemother Rep 1966;50:163-170.
47. Whitehead, J. The design and analysis of sequential clinical trials. 2nd ed. Chichester: Ellis Horwood; 1992.
48. Meduri GU, Headley, A S, Golden, E, et al. Effect of prolonged methylprednisolone therapy in unresolving acute respiratory distress syndrome: a randomized controlled trial. JAMA 1998;280:159-165.
49. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64:191-199.
50. McPherson K. On choosing the number of interim analyses in clinical trials. Stat Med 1982;1:25-36.
51. Wang SK, Tsiatis, A A. Approximately optimal one-parameter boundaries for group sequential trials. Biometrics 1987;43:193-199.
52. DeMets DL, Gail, M H. Use of logrank tests and group sequential methods at fixed calendar times. Biometrics 1985;41:1039-1044.
53. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. Br J Radiol 1971;44:793-797.
54. Fleming TR, Harrington, D P, O'Brien, P C. Designs for group sequential tests. Control Clin Trials 1984;5:348-361.
55. Proschan MA. Statistical methods for monitoring clinical trials. J Biopharm Stat 1999;9:599-615.
56. Julian DG, Camm, A J, Frangin, G, et al. Randomised trial of effect of amiodarone on mortality in patients with left-ventricular dysfunction after recent myocardial infarction: EMIAT. European Myocardial Infarct Amiodarone Trial Investigators. Lancet 1997;349:667-674.
57. Finfer S, Bellomo, R, Myburgh, J, Norton, R. Efficacy of albumin in critically ill patients. BMJ 2003;326:559-560.
58. DeMets DL, Lan G. The alpha spending function approach to interim data analyses. In: Thall, P. F. Recent advances in clinical trial design and analysis. Boston: Kluwer Academic Publishers; 1995: 1-27.
59. Lan KKG, DeMets, D L. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659-663.
60. Spiessens B, Lesaffre, E, Verbeke, G, Kim, K, DeMets, D L. An overview of group sequential methods in longitudinal clinical trials. Stat Methods Med Res 2000;9:497-515.
61. Bernard GR, Vincent, J L, Laterre, P F, et al. Efficacy and safety of recombinant human activated protein C for severe sepsis. N Engl J Med 2001;344:699-709.
62. Pampallona S, Tsiatis, A A, Kim, K M. Interim monitoring of group sequential trials using spending functions for the Type I and Type II error probabilities. Drug Inf J 2001;35:1113-1121.
63. Overall JE, Atlas, R S. Selecting an interim analysis procedure. Psychopharmacology Bulletin 1993;29:141-147.
64. Emerson SS, Fleming, T R. Symmetric group sequential test designs. Biometrics 1989;45:905-923.
65. Pampallona S, Tsiatis, A A. 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 1994;42:19-35.
66. Betensky R.A. Alternative derivations of a rule for early stopping in favor of H0. The American Statistician 2000;54:35-39.
67. DeMets DL, Hardy, R, Friedman, L M, Lan, K K. Statistical aspects of early termination in the beta-blocker heart attack trial. Control Clin Trials 1984;5:362-372.
68. Choi SC, Smith, P J, Becker, D P. Early decision in clinical trials when the treatment differences are small. Experience of a controlled trial in head trauma. Control Clin Trials 1985;6:280-288.
69. Ward JD, Becker, D P, Miller, J D, et al. Failure of prophylactic barbiturate coma in the treatment of severe head injury. J Neurosurg 1985;62:383-388.

70. Davis BR, Hardy, R J. Data monitoring in clinical trials: the case for stochastic curtailment. J Clin Epidemiol 1994;47:1033-1042.
71. Lan KK, Simon, R, Halperin, M. Stochastically curtailed tests in long-term clinical trials. Communications in Statistics - Sequential Anlaysis 1982;1:207-219.
72. Freidlin B, Korn, E L. A comment on futility monitoring. Contol Clin Trials 2002;23:355-366.
73. Spiegelhalter DJ, Freedman, L S, Blackburn, P R. Monitoring clinical trials: conditional or predictive power? Control Clin Trials 1986;7:8-17.
74. Emerson SS, Fleming, T R. Parameter estimation following group sequential hypothesis testing. Biometrika 1990;77:875-892.
75. Emerson, S. S+SEQTRIAL: Technical Overview. Research Report No 98. http://www.insightful.com/DocumentsLive/seqtech.pdf
76. Pinheiro JC, DeMets, D L. Estimating and reducing bias in group sequential designs with Gaussian independent incremental structure. Biometrika 1997;84:831-845.
77. Kittelson JM, Emerson, S S. A unifying family of group sequential test designs. Biometrics 1999;55:874-882.
78. S+SEQTRIAL 2 User's Manual. Seattle, WA: Insightful Corporation; 2002.
79. East V3.0 software. http://www.cytel.com/new.pages/EAST.2.html
80. PEST 4. http://www.rdg.ac.uk/mps/mps_home/software/software.htm
81. Emerson SS. Statistical Packages for group sequential methods. The American Statistician 1996;50:183-192.
82. Reboussin DM, DeMets, D L, Kim, K M, Lan, K K. Computations for group sequential boundaries using the Lan-DeMets spending function method. Control Clin Trials 2000;21:190-207.
83. Delgado-Herrera L, Anbar, D. A model for the interim analysis process: a case study. Control Clin Trials 2003;24:51-65.
84. Wheatley K, Clayton, D. Be skeptical about large apparent treatment effects: the case of an MRC AML12 randomization. Control Clin Trials 2003;24:66-70.
85. Emerson SS. Stopping a clinical trial very early based on unplanned interim analyses: a group sequential approach. Biometrics 1995;51:1152-1162.