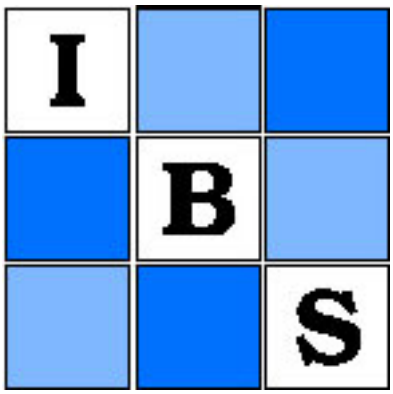

A Simple Algorithm for Designing Group Sequential Clinical Trials
Author(s): David A. Schoenfeld
Source: *Biometrics*, Vol. 57, No. 3 (Sep., 2001), pp. 972-974
Published by: International Biometric Society
Stable URL: http://www.jstor.org/stable/3068440
Accessed: 04/06/2009 09:39

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=ibs.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.

International Biometric Society is collaborating with JSTOR to digitize, preserve and extend access to *Biometrics.*

http://www.jstor.org

BIOMETRICS 57, 972–974
September 2001

A Simple Algorithm for Designing Group Sequential Clinical Trials

David A. Schoenfeld

Massachusetts General Hospital–Biostatistics,
50 Staniford Street, Boston, Massachusetts 02114, U.S.A.
email: dschoenfeld@partners.org

SUMMARY. This article describes a simple algorithm for calculating probabilities associated with group sequential trials. This allows the choice of boundaries that may not be among those implemented in available software.

KEY WORDS: Acute respiratory distress syndrome; Clinical trials; Futility stopping rules; Group sequential designs; Hypothesis test; Numerical integration.

1. Introduction

In 1995, the National Heart, Lung, and Blood Institute established the Acute Respiratory Distress Syndrome (ARDS) Network to conduct clinical trials on ARDS, a syndrome that is initiated by pneumonia, trauma, or other serious condition. Patients' lungs fill up with fluid, and the patients require mechanical ventilation. The exact cause is unknown and there are no approved drug treatments. The mortality rate is 30–50%, and death usually occurs within a month of onset. There are an estimated 50,000 cases a year in the United States.

Given the failure of drug treatments in the past and the paucity of data on the drugs that they were planning to test, Network investigators wanted a trial design where studies would stop if there was not a trend toward efficacy at the first interim analysis. They wanted to review the data at a few specified interim analyses. At each analysis, they would stop the trial in favor of the null hypothesis if the treatment was not 3% better, in terms of mortality, than the control. They would also stop in favor of the alternative hypothesis using an O'Brien and Fleming (1979) upper boundary.

The lower boundary was designed to stop the study if continuing it was futile. It nested a pilot study inside a phase III study since the first group of patients could be thought of as the pilot study. These boundaries lower the power of the study by the probability under the alternative hypothesis that the lower boundary would be crossed.

There are several ways of designing group sequential trials. One approach is to use published designs (O'Brien and Fleming, 1979; DeMets and Ware, 1982). The futility stopping boundaries developed by DeMets and Ware (1982) are described in their paper. To use them, no computer program is necessary. They require the treatment to be worse than the control at the first look, so they would not stop a negative study as early as the ARDS Network investigators wanted.

An alternative is to use commercial software. The PEST and EAST software are reviewed in Emerson (1996). EAST 2000 provides support for designing, monitoring, and analyzing trials with boundaries described in Wang and Tsiatis (1987) and Pampallona and Tsiatis (1994). These include the boundary introduced by Pocock (1977) and those introduced by O'Brien and Fleming (1979). The futility boundaries implemented in this software are the family of lower boundaries developed by Pampallona and Tsiatis (1994). In EAST, they are constrained to have a crossing probability, given the alternative hypothesis, that is equal to the type II error and to converge at the last look to the upper boundary. This bound allows one to demonstrate that the alternative hypothesis is false. Although this would have been a reasonable choice of bound, it was not what was specified by the investigators. EAST allows the user to customize boundaries, using a simulation to calculate type I and II errors.

PEST 4 implements methods described by Whitehead (1999). Its triangular test and truncated sequential probability ratio test design both have lower futility boundaries, but there is little flexibility in specifying the boundary shape. It has a module for calculation of p-values and confidence intervals that can be used to calculate type I and II errors for customized boundaries (Whitehead, personal communication). To calculate a p-value using this module, one specifies the boundaries and the value of the test statistic at each look at the data. The value of the test statistic is irrelevant before the last look. It calculates a two-sided p-value equal to $2\min(p, 1-p)$, where p is the probability of crossing the upper boundary before the last look or achieving a test statistic value greater than that observed at the last look. The program then uses an algorithm presented in Jennison and Turnbull (2000), which is similar to the one presented here. The program can be used to calculate power by subtracting the mean of the statistic under the alternative hypothesis from the boundaries.

Use of these features in EAST and PEST to design studies would be very difficult because there is no straightforward way to repeatedly change the sample size and rerun the programs in order to find the sample size to obtain a specified power.

Thus, to design trials with customized boundaries, it is necessary to calculate type I and II errors either by simulation or by numerical integration. Simulation is very easy to program but is too slow when one has to repeatedly modify the sample size and rerun the simulation in order to find the sample size that will achieve a specified power. Armitage, McPherson, and Rowe (1969) developed an algorithm for recursive numerical integration that reduces the calculation of type I and II errors for group sequential trails to the calculation of several one-dimensional integrals. The purpose of the present article is to present a version of this algorithm for calculating boundaries and stopping probabilities that is easy to program to facilitate the development of stopping rules that are not implemented in available software. The algorithm could be used to calculate p-values as described in Fairbanks and Madsen (1982).

2. Description of Algorithm

Assume that a clinical trial has two treatments, that we look at the data m times, and that, at each time, the data are summarized by a statistic Y_k, $k = 1, \ldots, m$, that has a standard normal distribution under H_0. In a group sequential trial, there are defined boundaries, $a_1, \ldots, a_m$, $b_1, \ldots, b_m$, such that, if $a_k \leq Y_k \leq b_k$, the trial continues to the next look unless $k = m$, in which case, the null hypothesis is accepted. If at any time $Y_k < a_k$ or $Y_k > b_k$, the decision is made to stop the trial for futility or efficacy, respectively.

Designing a trial consists of choosing the boundaries and sample size so that the probabilities of crossing the upper boundary or crossing the lower boundary equal specified quantities under the null and alternative hypotheses. Under H_0, the probabilities of interest are $P_U(a,b) = \Sigma_{k=1}^{m} P(a_1 \leq Y_1 \leq b_1, \ldots, a_{k-1} \leq Y_{k-1} \leq b_{k-1}, b_k < Y_k)$ and $P_L(a,b) = \Sigma_{k=1}^{m} P(a_1 \leq Y_1 \leq b_1, \ldots, a_{k-1} \leq Y_{k-1} \leq b_{k-1}, Y_k < a_k)$.

Under the alternative, the probabilities have the same expression except that one uses as boundaries $(a_k - \mu_k)$ and $(b_k - \mu_k)$, where μ_k is the mean of Y_k under the alternative hypothesis. Denote the summands above as $P_{L,k}$ and $P_{U,k}$.

3. Calculation of the Probabilities

Under the additional assumption that Y_k, $k = 1, \ldots, m$, has the same distribution as $t_k^{-1/2} \Sigma_{j=1}^{k} Z_i$, where $Z_1, Z_2 \cdots$ are independent normal random variables with variances $t_1, t_2 - t_1, \ldots, t_k - t_{k-1}$, the algorithm of Armitage et al. (1969) provides a way of calculating these quantities without multivariate integration. The algorithm presented here recasts their algorithm as the multiplication of a sequence of matrices. Suppose that we divide the interval from a_k to b_k into N equally spaced intervals. Let $d_k = (b_k - a_k)/N$ be the width of each interval and let $x_{k,j} = a_k + (j - 0.5)d_k$, $j = 1, \ldots, N$, be the midpoint.

Let the ith element of the matrix M_1 approximate the probability that Y_1 is in the interval centered at $x_{1,i}$. Thus, $(M_1)_i = d_1 \exp(-x_{1,i}^2/2)(2\pi)^{-1/2}$.

Let the i,jth element of the matrix $M_k, k = 2, \ldots, m-1$, approximate the probability that Y_k is in the interval centered at $x_{k,i}$ given that Y_{k-1} is in the interval centered at $x_{k-1,j}$. Thus, $(M_k)_{i,j} = d_k[t_k/\{2\pi(t_k - t_{k-1})\}]^{1/2} \exp[-(t_k^{1/2} x_{k,i} - t_{k-1}^{1/2} x_{k-1,j})^2/\{2(t_k - t_{k-1})\}]$.

Let $(V_{U,1})_i = \Phi(-b_1)$ be the probability that $Y_1 > b_1$ and let the ith element of $V_{U,k}$ approximate the probability that $Y_k > b_k$ given that Y_{k-1} is in the interval centered at $x_{k-1,i}$,

$$(V_{U,k})_i = \Phi\left\{-\left(t_k^{1/2} b_k - t_{k-1}^{1/2} x_{k-1,i}\right)/(t_k - t_{k-1})^{1/2}\right\}.$$

Similarly, $(V_{L,1})_i = \Phi(a_1)$ and

$$(V_{L,k})_i = \Phi\left\{\left(t_k^{1/2} a_k - t_{k-1}^{1/2} x_{k-1,i}\right)/(t_k - t_{k-1})^{1/2}\right\}.$$

Then $P_{U,k} \approx V_{U,k} M_{k-1} \cdots M_1$ and similarly for $P_{L,k}$. Note that the calculated probabilities do not change if all of the t_k are multiplied by a constant.

The algorithm is proposed for its ease of programming rather than for its accuracy or speed. The accuracy of the algorithm letting $N = 60$ compares favorably with using Simpson's rule and using a fast Fourier transform (Elson, 1995), with small differences in the last decimal place. The value of N should be increased until the value of the calculated probabilities do not change in the third decimal place. Execution time is less of an issue now than it was in 1995 when Elson proposed using a fast Fourier transform. It was found to be adequate for interactive experimentation with different boundaries and sample sizes. A Matlab m-file and a computer program that runs under Windows 95 are available from the author.

4. Example

The first trial using this method tested ketoconazole (see ARDS Network (2000) for details). This trial was to have interim looks at 100, 200, and 300 patients on each treatment and a final analysis when there were 400 patients on each treatment. The mortality rate under the null hypothesis was 0.35 in both treatments, and under the alternative, the mortality rates were 0.3 and 0.4. To use the method above, let n_i be the number of patients on each treatment at the ith look and let p_1 and p_2 be the estimated mortality rates on each treatment. Then, letting $v_i = (p_1(1-p_2)/n_i + p_2(1-p_2)/n_i)^{1/2}$, $Y_i = (\hat{p}_1 - \hat{p}_2)/v_i$ and $t_i = v_i^2$.

The lower futility bound is therefore $a_i = 0.03/v_i$. In this case, with $p_1 = p_2 = 0.35$, it would be $a = (0.44, 0.63, 0.77, 0.89)$. To perform the calculation under H_a, set $a_i = (0.03 - 0.1)/v_i$, which gives $a = (-1.04, -1.48, -1.81, -2.09)$.

The upper O'Brien Fleming boundary is of the form $b_i = C(t_m/t_i)^{1/2}$. As noted by DeMets and Ware (1982), the value of C must be adjusted downward to counteract the effect of the futility boundary. In this case, with a one-sided 0.05 significance level, $C = 1.57$, which was found by trial and error, starting with $C = 1.654$. Rejecting the null hypothesis of treatment equality in favor of the hypothesis that the control is better than the treatment is nearly impossible with this design, which justifies a one-sided p-value. However, the choice of a significance level of 0.025 may be preferable because it is consistent with two-sided p-values, which are more commonly used. Under the null hypothesis, $b = (3.14, 2.22, 1.81, 1.57)$, and under the alternative, it is $b_i = C(t_m/t_i)^{1/2} - 0.1/v_i$, i.e., $b = (1.65, 0.11, -0.77, -1.41)$.

With these boundaries, there is a 67% chance that we would stop at the first look and an 81% chance we would stop by the second look under the null hypothesis. Under the alternative hypothesis, the probability of rejecting H_0 would be

80%. If we removed the futility stopping rule entirely, the power would increase to 90%. The biggest loss of power occurs on the first look since there is a 15% chance of stopping under the alternative hypothesis. The power would be 88% if the futility boundary at the first look is removed. The power would be 85% if the boundary is changed to require that the treatment not be worse at the first look.

If the time of the interim analysis changed, we used the DeMets and Lan (1994) α-spending function strategy. We used linear interpolation to find a spending function $\alpha(n)$ defined on the interval $(0, \Sigma_{i=1}^{m} n_i)$ with $\alpha(\Sigma_{i=1}^{k} n_i) = \Sigma_{i=1}^{k} P_{U,i}$. If a look was before or after the allotted number of patients, say at n_k', we adjusted a_k so that $\Sigma_{i=1}^{k} P_{U,i} = \alpha(n_k')$. This method correctly preserved the size of the trial. We also observed that the futility stopping rule might have a greater negative impact on power than planned if the mortality rate was closer to 0 or 100%. This was handled by allowing an adjustment of the lower boundary to preserve a power of at least 80%.

One problem with futility boundaries is that if the trial is not stopped when the futility boundary is crossed, then the type I error of the trial will be over 5% whichever stopping rule is subsequently used (Lan, 2001 personal communication).

5. Conclusion

Futility stopping rules for clinical trials can radically reduce the cost of a long-term drug development program by stopping futile trials early. They simultaneously guard against the possibility that a new treatment will harm clinical trial participants. The lack of software is often the largest impediment to the use of new statistical methods such as futility stopping rules. Hopefully, this algorithm will allow more flexibility in the use of sequential stopping rules.

Acknowledgements

The author thanks Cyrus Mehta, John Whitehead, Dianne Finkelstein, and the referees for several improvements. Funding was provided by the National Institutes of Health, HR-46064 and CA-78784.

Résumé

Ce papier décrit un algorithme simple pour le calcul des probabilités associées aux essais séquentiels groupés. Elle autorise le choix de frontières pouvant ne pas être parmi celles proposées dans les logiciels diffusés.

References

Anonymous. (2000). Ketoconazole for early treatment of acute lung injury and acute respiratory distress syndrome: A randomized controlled trial. *Journal of the American Medical Association* **283,** 1995–2002.

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

DeMets, D. L. and Lan, K. G. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine* **13,** 1341–1352.

DeMets, D. L. and Ware, J. H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* **69,** 661–663.

Elson, P. J. (1995). Use of the fast Fourier transform algorithm in the calculation of the operating characteristics of group sequential clinical trials. *Computational Statistics and Data Analysis* **20,** 491–498.

Emerson, S. S. (1996). Statistical packages for group sequential methods. *The American Statistician* **50,** 183–192.

Fairbanks, K. and Madsen, R. (1982). *P* values for tests using repeated significance test design. *Biometrika* **69,** 69–74.

Jennison, C. and Turnbull, B. (2000). *Group Sequential Methods with Applications to Clinical Trials.* Boca Raton: Chapman and Hall.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35,** 549–556.

Pampallona, S. and Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor 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,** 191–200.

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

Whitehead, J. (1999). A unified theory for sequential clinical trials. *Statistics in Medicine* **18,** 2271–2286.

Received June 2000. *Revised May* 2001.
Accepted May 2001.