3-2001

Optimal Designs for Evaluating a Series of Treatments

Denis H. Y. Leung
Singapore Management University, denisleung@smu.edu.sg

You Gan WANG
Harvard University

Published in Biometrics, 2001 March, 57 (1), 168-171.
DOI: 10.1111/j.0006-341X.2001.00168.x

Optimal Designs for Evaluating a Series of Treatments

Denis Heng-Yan Leung

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center,
1275 York Avenue, New York, New York 10021, U.S.A.
email: leung@biost.mskcc.org

and

You-Gan Wang

Department of Biostatistics, Harvard School of Public Health,
655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
email: ygwang@hsph.harvard.edu

SUMMARY. Several articles in this journal have studied optimal designs for testing a series of treatments to identify promising ones for further study. These designs formulate testing as an ongoing process until a promising treatment is identified. This formulation is considered to be more realistic but substantially increases the computational complexity. In this article, we show that these new designs, which control the error rates for a series of treatments, can be reformulated as conventional designs that control the error rates for each individual treatment. This reformulation leads to a more meaningful interpretation of the error rates and hence easier specification of the error rates in practice. The reformulation also allows us to use conventional designs from published tables or standard computer programs to design trials for a series of treatments. We illustrate these using a study in soft tissue sarcoma.

KEY WORDS: Bayesian; Optimality; Phase I studies; Screening; Sequential trials; Soft tissue sarcoma.

1. Introduction

A number of research programs at Memorial Sloan-Kettering Cancer Center (MSKCC) have prompted the need to test a series of treatments to identify promising ones for more intensive studies. One example is in the Immunology Program, where a large number of vaccines are to be screened for potentially efficacious ones for further study. Another example is in the Soft Tissue Sarcoma Chemotherapy Program, where a large number of Phase II trials are to be implemented for identifying promising chemotherapy agents. In response to these needs, a number of design strategies have recently been proposed (Yao, Begg, and Livingston, 1996; Wang and Leung, 1998; Yao and Venkatraman, 1998). A novel aspect in these designs is the recognition that it is not possible to predict when a promising treatment will appear and therefore no upper limit is placed on the number of treatments to test in these designs. The goal is to find the first promising treatment with the smallest expected number of patients while the errors of falsely accepting a nonpromising treatment and of rejecting one or more promising treatments are controlled for.

Yao et al. (1996) (referred to as YBL hereafter) used a single-stage design to solve this problem. That is, n patients are to be tested on each treatment and a treatment will be accepted or rejected depending on whether $>k$ or $\leq k$ successes are seen. More recently, Yao and Venkatraman (1998) (hereafter called YV) considered a two-stage design after YBL. The design consists of a series of two-stage trials. The maximum sample size in the two stages are n_1 and n_2, respectively. If k_1 or fewer successes at the end of the first stage or a total of k_2 or fewer successes at the end of the second stage are observed, the treatment will be rejected. They further suggested a truncated two-stage design. For example, if (k_1, n_1, k_2, n_2) are the parameters, then we need not test the treatment any further as soon as $n_1 - k_1$ failures are observed because this treatment will definitely be rejected even if we finish the first-stage testing. Similarly, the design can be truncated at $n_2 + n_1 - k_2$ failures in the second stage. Therefore, they suggested stopping testing as soon as these events have been observed. Hence, they have changed their two-stage monitoring to fully sequential monitoring except the stopping boundaries remained defined by an optimal two-stage design. In fact, the optimal fully sequential design is given by Wang and Leung (1998) (hereafter called WL), in which the two types of error rates are achieved by using Lagrangian constraints.

Although it is more realistic to consider testing a series of treatments as an ongoing process, this approach adds substantial conceptual complexity to the design problem. In this article, we show that the design for testing a series of treatments is merely a series of identical designs for testing the

individual treatments. The design for individual treatments can be one of the conventional designs (one stage, two stage, multistage) in the literature. We will establish a link between the error rates in the conventional design and the error rates in the design for the ongoing process. This link will allow us to convert the error rates for different designs. We identify two uses of our work. First, we show in Sections 2 and 3 that the error rates for the individual treatments are easier to interpret and hence easier to specify in practice. When designing trials involving a series of treatments, we suggest specifying the error rates for the individual treatments. As long as the converted error rates are used in the conventional design for the individual treatments, the overall errors for testing a series of treatments will be maintained at the desired level. Second, the design problem for a series of treatments can be solved using conventional designs, which are easily available from published tables or from standard computer programs. We illustrate these uses in an example in soft tissue sarcoma.

2. Ongoing Process Versus Single Treatment

The problem of testing or screening a series of treatments can be written formally as follows. Suppose the success probability of the ith selected treatment is θ_i and a treatment is considered to be promising if $\theta_i \geq \theta^*$. The trial consists of a series of testing of

$$\mathrm{H}_0\colon \theta_i < \theta^* \text{ versus } \mathrm{H}_1\colon \theta_i \geq \theta^*.$$

In general, θ_i is unknown but we may impose a prior distribution on θ_i to reflect our knowledge about the treatments. In practice, it is impossible to predict how many treatments to test before a promising one can be identified. Therefore, we assume that there is an indefinite series of treatments available and that a rejected treatment will never be tested again, so after a rejection, the screening problem returns to its original state (see YBL, YV, and WL). This assumption is important because the order in which a treatment arrives becomes irrelevant and so the same testing strategy can be used for all treatments in the screening trial.

For any particular design used in the screening process, two types of errors can be identified, (1) that of accepting a nonpromising treatment (α_1) and (2) that of rejecting one or more promising ones (α_2). If we denote the testing outcome for a treatment as Y (0 = rejected and 1 = accepted), then the two error rates can be expressed as (see YBL)

$$\alpha_1 = \Pr(\theta_i < \theta^* \mid Y = 1\} = \frac{p_{+-}}{P_{+-} + p_{++}},$$
$$\alpha_2 = \frac{p_{-+}}{1 - p_{--}}, \tag{1}$$

where

$$p_{--} = \Pr(Y = 0, \theta_i < \theta^*),$$
$$p_{-+} = \Pr(Y = 0, \theta_i \geq \theta^*),$$
$$p_{+-} = \Pr(Y = 1, \theta_i < \theta^*),$$
$$p_{++} = \Pr(Y = 1, \theta_i \geq \theta^*).$$

The design problem is to minimize the expected overall number of patients to find the first promising treatment when α_1 and α_2 are controlled. Note that α_1 and α_2 are posterior probabilities and are different from traditional Type I and II errors.

By considering the screening trial as an ongoing hypothesis testing problem, each treatment is considered to have the same response distribution in the design for the ongoing process. Therefore, the same design can be used for all treatments in the trial. However, the ongoing process still adds considerable complexities to the determination and interpretation of the design. Here we show that the ongoing process is intimately linked to the tests of the individual treatments.

2.1 *Error Rates*

From (1), it is clear that the four quantities (p_{--}, p_{+-}, p_{-+}, p_{++}) are determined entirely by the errors (α_1, α_2) together with the two constraints: $p_{++} + p_{-+} = \Pr(\theta \geq \theta^*)$ and $p_{++} + p_{+-} + p_{-+} + p_{--} = 1$. The value $\Pr(\theta_i \geq \theta^*)$ will be denoted as p for convenience.

Now consider the error rates for any particular treatment rather than the overall error rate of the screening process. The false-positive error, α_1, is unaffected but the false-negative error is the probability that a rejected treatment is in fact a promising one, i.e.,

$$\alpha_2^* = \Pr(\theta_i \geq \theta^* \mid Y = 0\} = \frac{p_{-+}}{p_{-+} + p_{--}}.$$

From (1), we can obtain the relationship between α_2 and α_2^* as

$$\alpha_2^* = \frac{\alpha_2 p}{1 - (1 - \alpha_2)(\alpha_1 + p)}. \tag{2}$$

Consequently, if one is interested in an optimal design for the screening trial with error rates (α_1, α_2), one simply needs to find the optimal design for a single treatment trial with error rates (α_1, α_2^*). This conventional design for a single treatment trial with error rates (α_1, α_2^*) is also the optimal design for the ongoing process with error rates (α_1, α_2). Some values of α_2 and α_2^* for various values of p and $\alpha_1 = 0.1$ are given in Table 1. We note that, from Table 1, α_2 is in general larger than α_2^*. But this is not true when p is large, as it results in a large probability of acceptance each time (i.e., when the expected number of treatments needed to test before a promising one can be found is small).

Because the screening of new treatments is modeled as an ongoing process, the use of α_2 is appealing. However,

Table 1
Values of α_2, α_2^* for different values of $p = \Pr(\theta \geq \theta^*)$, $\alpha_1 = 0.1$

p	α_2	α_2^*
0.100	0.050	0.006
0.300	0.050	0.024
0.500	0.050	0.058
0.100	0.150	0.018
0.300	0.150	0.068
0.500	0.150	0.153
0.100	0.250	0.029
0.300	0.250	0.107
0.500	0.250	0.227
0.100	0.350	0.040
0.300	0.350	0.142
0.500	0.350	0.287

α_2 is defined in terms of the expected number of rejected treatments before a promising one is declared. This limits its practical use in two ways. First, it cannot be used to make statements about the false-error rate among rejected treatments in any particular trial because the number of rejected treatments is a random number. Second, its interpretation is only meaningful if the expected number of rejected treatments is also reported. Therefore, we suggest specifying α_2^* instead. The interpretation of α_2^* is clear; it is simply the probability of a false negative when a treatment is rejected. Because of (2), the use of α_2^* will also control the overall error of the trial to be bounded by α_2. If we choose to control (α_1, α_2^*), the design would have (p_{++}, p_{+-}) given by

$$p_{++} = \frac{(1-\alpha_1)(p-\alpha_2^*)}{1-\alpha_1-\alpha_2^*}$$

$$p_{+-} = \frac{\alpha_1(p-\alpha_2^*)}{1-\alpha_1-\alpha_2^*}.$$

From a Bayesian viewpoint, $(1-\alpha_1, \alpha_2^*)$ are the two posterior probabilities of being promising when the treatment is accepted or rejected. Note that the prior probability of being promising is p. Therefore, $(1-\alpha_1)/p$ measures the efficiency of screening and α_2^*/p can be interpreted as the remaining efficiency among those rejected. Other variations of the error probabilities may be used depending on the context of the trials. Or alternatively, one can just specify the desired proportions p_{++} and p_{+-}, and a variety of error probabilities can then be evaluated. As we mentioned, the error rates are all essentially determined by p_{++} and p_{+-}.

2.2 *Sample Size*

If $\mathrm{E}(\tau)$ is the expected sample size for each treatment to be tested, the expected sample size until a promising treatment is identified is $N = \mathrm{E}(\tau)N_v$, where N_v is the expected number of treatments to be tested before a promising one is accepted. For any given error probabilities (α_1, α_2), the expected number of trials required to identify a promising one is

$$N_v = \frac{1-\alpha_1(1-\alpha_2)}{(1-\alpha_2)p}.$$

The design aims to minimize the expected sample size, N, until a promising treatment is identified. In fact, given two error probabilities, (α_1, α_2) or (α_1, α_2^*), or any other variations (functions of p_{++} and p_{+-}), N_v is a constant for any design, i.e., $N_v = [1-\alpha_1(1-\alpha_2)]/[(1-\alpha_2)p]$. So an optimal design can be obtained by minimizing the expected sample size for testing each treatment, $\mathrm{E}(\tau)$, with controlled error rates (α_1, α_2^*). Therefore, treating screening as an ongoing process does not change any essential aspects of the design, i.e., a design that is optimal for each individual treatment is also optimal for the overall ongoing process (until a promising one is identified) and vice versa.

2.3 *Connection to Conventional Designs Under a Two-Point Prior*

So far, we have formulated the hypothesis testing problem as a test of H_0: $\theta_i > \theta^*$ versus H_1: $\theta_i \geq \theta^*$. But often, we simplify the problem to that of testing H_0': $\theta_i = \theta_0 < \theta^*$ against H_1': $\theta_i = \theta_a > \theta^*$. In that case, we can assume a two-point prior for the success probability θ_i so that $\Pr(\theta_i = \theta_0) = 1-p$ and $\Pr(\theta_i = \theta_a) = p$. Let

$$\alpha = \Pr(Y = 1 \mid \theta_i = \theta_0); \qquad \beta = \Pr(Y = 0 \mid \theta_i = \theta_a)$$

be the frequentist Type I and II error rates. Then one can show (cf., Lee and Zelen, 2000) that α and β can be obtained from α_1 and α_2^* using the following relationships:

$$\alpha = \frac{\alpha_1(p-\alpha_2^*)}{(1-p)(1-\alpha_1-\alpha_2^*)}$$

$$\beta = \frac{\alpha_2^*(1-\alpha_1-p)}{p(1-\alpha_1-\alpha_2^*)}. \tag{3}$$

Therefore, to design a trial for testing a series of treatments, with errors (α_1, α_2), we first use (2) to convert (α_1, α_2) to (α_1, α_2^*). Then we use (3) to convert (α_1, α_2^*) to the frequentist errors (α, β). Based on (α, β), the optimal design for a series of treatments can be easily obtained from standard tables or programs.

3. Soft Tissue Sarcoma Trials

Every year, MSKCC conducts a large variety of clinical trials on treatments for different types of malignancies. Among these are the trials in the chemotherapy program for soft tissue sarcomas (NIH grant PO1-47179). The primary goal of this program is to identify promising chemotherapy treatments, through a series of Phase I and II trials, for further study. A total of 12 different Phase II trials are planned for the next 5 years. We illustrate the implications of the issues raised in Section 2 in this context.

We used prior knowledge from 52 historical Phase II trial results (Yang et al., 1993) to help us in designing the new trials in the program. We acknowledged that there is a possibility of publication bias, which will be ignored in this illustration. The 52 historical trials arise from 30 different combination chemotherapy treatments. The number of patients used on these combination treatments ranges from 8 to 732. In forming the prior, we included only studies with doxorubicin-, epirubicin-, and ifosfamide-based treatments because the new treatments to be tested are either combinations based on one of these three or with similar activities to these three. The efficacy of each treatment is measured in terms of the percent of complete response, which is defined as complete disappearance of the tumor. We assumed that each treatment has a probability of complete response that follows a beta distribution. We found the best fitting beta-binomial model to these data to be beta$(1.3, 8.6)$.

Our clinical collaborators have indicated that only treatments with a complete response rate of $\geq$20% are worthy of further investigation. Under the prior, the proportion of treatments worthy of further study is 0.217. We limit the false-positive error (α_1) to be under 10% and false-negative error (α_2) to be under 30%. YV's optimal design, when restricting $N = n_1 + n_2 \leq 100$, is $(k_1, n_1, k_2, n_2) = (2, 17, 16, 56)$. The average number of patients required to find a promising treatment using this design is 225.5. Using WL's fully sequential design and constraining the maximum sample size of each Phase II trial to be the same as YV's design, the optimal design requires, on average, 145.2 patients to reach a promising treatment. Therefore, the savings by adopting the fully sequential design over a two-stage design is about 35%. The fully sequential design reduces patient numbers by having

the opportunity to stop earlier. Note that the sample size per Phase II trial is quite large in order to control for moderate values of α_1 and α_2. However, from (2), we can obtain α_2^* to be 0.084 only. In other words, when a treatment is rejected, the probability that it has a response rate of $\geq$20% is 8.4%. Therefore, the design controls (α_1, α_2^*) to be under 10% each even though α_2 is 30%.

Using $\alpha_1 = 0.1$ and $\alpha_2 = 0.3$, the average number of treatments to test before a positive one is declared is $(N_v - 1) = 5.2$. This number is a constant for the design, with $(\alpha_1, \alpha_2) = (0.1, 0.3)$ or, equivalently, $(\alpha_1, \alpha_2^*) = (0.1, 0.084)$. The interpretation of α_1 is clear; it is the probability of a false-positive error. However, the interpretation of α_2 is not as straightforward. Suppose, in a particular trial, we rejected four treatments before accepting a promising one. Then, even though we know $\alpha_2 = 0.3$, we cannot use it to make statements about the chance of false-negative errors in any of the four rejected treatments. On the other hand, using $\alpha_2^* = 0.084$, we know the chance of a false negative for each rejected treatment is 8.4% in each of the four rejected treatments.

A second problem in interpreting α_2 is as follows. The quantity $\alpha_2 = 0.3$ refers to the probability that at least one of the rejected treatments is promising and is equal to 0.3. This probability is based on an average of 5.2 rejected treatments (i.e., $N_v - 1 = 5.2$). If we keep $\alpha_2 = 0.3$ but specify α_1 to be 0.2, then $(N_v - 1)$ becomes 4.7. In this case, even though α_2 is still 0.3, it is now based on an average of only 4.7 rejected treatments, which is more serious than when it is based on an average of 5.2 rejected treatments. Therefore, two identical specifications of α_2 represent two different levels of errors committed. The interpretation of α_2 depends on N_v, which in turn depends on (α_1, α_2), and the observed N_v value follows a geometric distribution. This makes it difficult to specify the error rates in terms of (α_1, α_2) in practice. On the other hand, the meaning of α_2^* is unchanged, as its interpretation is for individual treatments rather than for a series of treatments. When α_1 changes to 0.2 from 0.1 while fixing α_2 at 0.3, α_2^* changes to 0.091 from 0.084.

Finally, we illustrate how the design problem for a series of trials can be simplified when one assumes a two-point prior for θ_i. Based on the data, we assume that we are interested in testing $\theta_i = 0.1$ versus $\theta_i = 0.3$ and $\Pr(\theta_i = 0.1) = 0.7$ and $\Pr(\theta_i = 0.3) = 0.3$. Furthermore, suppose we allow errors of $\alpha_1 = 0.1$ and $\alpha_2 = 0.1$. For ease of illustration, we use the one-stage design of YBL. Then using (2), we obtain $\alpha_2^* = 0.042$. From (3), we then have $(\alpha, \beta) = (0.043, 0.098)$. Using standard programs from standard single-treatment trials, with $(\alpha, \beta) = (0.043, 0.098)$, we obtain the optimal one-stage design as in YBL (Table 1). This design is given by $(n, k) = (33, 6)$ and it means to use 33 patients for each treatment and then declare the treatment as promising if more than six successful responses are observed.

4. Discussion

In this article, we established a link between the overall error probability for the ongoing process framework and the errors for each individual treatment. Error probabilities in terms of the individual treatments may be more meaningful in practice. This concept was also suggested in a recent discussion paper by Lee and Zelen (2000) in a different context. The interpretation of these error rates is also easier. For example, if (α_1, α_2^*) are controlled at (0.1, 0.05), we would expect 90% of those accepted for further study are truly promising and only 5% among those rejected are promising.

On the other hand, α_2 is the probability of rejecting one or more promising treatments before we recommend a promising one. But since the number of treatments tested before recommending a promising one is a random variable (geometrically distributed with a mean of N_v), the interpretation of α_2 is awkward. We therefore believe specifying α_2^* is more meaningful in practice. Furthermore, when the prior is a two-point prior, standard programs or published tables can be used to obtain optimal designs with specified (α_1, α_2^*) values.

Acknowledgements

This research has been partially supported by NIH grants CA74207 and CA47179 for DHL and AI24643 for YGW. We thank Drs T. J. Yao and E. S. Venkatraman for providing their program to calculate their two-stage designs. We are also grateful to the associate editor and to two anonymous reviewers for their constructive comments, which led to a much improved article.

Résumé

Plusieurs articles, dans cette revue, ont étudié des plans optimaux permettant de tester différents traitements—les uns à la suite des autres—dans le but d'identifier ceux qui s'avéreraient suffisamment prometteurs pour faire l'objet d'investigations ultérieures. L'analyse de ces plans peut être considérée comme l'analyse d'une série de tests qui se poursuivent jusqu'à ce que soit identifié un traitement prometteur. Plus réaliste, cette formulation accroît cependant, et de façon non négligeable, la complexité calculatoire du problème. Dans cet article, nous montrons que ces nouveaux plans expérimentaux, où l'on contrôle les risques d'erreur pour l'ensemble des traitements, peuvent aussi être formulés comme des plans classiques où l'on contrôle les risques d'erreur pour chaque traitement pris séparément. Cette reformulation permet, lorsque l'on a besoin de construire un plan expérimental concernant une série de traitements, d'utiliser les tables publiées pour les plans classiques, ainsi que les logiciels correspondants. Nous illustrons tout ceci à partir d'une étude dans le sarcome des tissus mous.

References

Lee, S. J. and Zelen, M. (2000). Clinical trials and sample size considerations: Another perspective (with discussion). *Statistical Science* **15,** 95–110.

Wang, Y.-G. and Leung, D. H. Y. (1998). An optimal design for screening trials. *Biometrics* **54,** 245–252.

Yang, J., Glatstein, E., Rosenberg, A., and Antman, K. (1993). Sarcomas of soft tissue. In *Cancer: Principles and Practice of Oncology,* V. DeVita, S. Hellman, and S. Rosenberg (eds), 1478–1479. Philadelphia: Lippincott.

Yao, T. J. and Venkatraman, E. S. (1998). Optimal two-stage design for a series of pilot trials of new treatments. *Biometrics* **54,** 1183–1189.

Yao, T. J., Begg, C. B., and Livingston, P. O. (1996). Optimal sample size for a series of pilot trials of new treatments. *Biometrics* **52**, 992–1001.