1-2001

A Bayesian Decision Approach for Sample Size Determination in Phase II Trials

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

You-Gan WANG
Harvard University

DOI: https://doi.org/10.1111/j.0006-341X.2001.00309.x

Citation
LEUNG, Denis H. Y. and WANG, You-Gan. A Bayesian Decision Approach for Sample Size Determination in Phase II Trials. (2001). *Biometrics*. 57, (1), 309-312. Research Collection School Of Economics.
Available at: https://ink.library.smu.edu.sg/soe_research/32

Published in BiometricsVolume 57, Issue 1, 2001, Pages 309-312
https://doi.org/10.1111/j.0006-341X.2001.00309.x

A Bayesian Decision Approach for Sample Size Determination in Phase II Trials

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.

SUMMARY. Stallard (1998, *Biometrics* **54,** 279–294) recently used Bayesian decision theory for sample-size determination in phase II trials. His design maximizes the expected financial gains in the development of a new treatment. However, it results in a very high probability (0.65) of recommending an ineffective treatment for phase III testing. On the other hand, the expected gain using his design is more than 10 times that of a design that tightly controls the false positive error (Thall and Simon, 1994, *Biometrics* **50,** 337–349). Stallard's design maximizes the expected gain per phase II trial, but it does not maximize the rate of gain or total gain for a fixed length of time because the rate of gain depends on the proportion of treatments forwarding to the phase III study. We suggest maximizing the rate of gain, and the resulting optimal one-stage design becomes twice as efficient as Stallard's one-stage design. Furthermore, the new design has a probability of only 0.12 of passing an ineffective treatment to phase III study.

KEY WORDS: Bayesian; Decision theory; Gain function; Gittins Index; Sample size; Sequential design.

1. Introduction

Recently, Stallard (1998) considered sample size determination in phase II trials. The problem was developed in the context of developing a new treatment and so, if the treatment is accepted in the phase II trial, a phase III trial will be carried out to compare the new treatment with a standard. He assumed that there are costs of conducting the phase II and phase III trials and there are potential gains should the treatment be proven to be superior to the standard in the phase III trial. The costs and benefits were incorporated in a gain function. The treatment was to be tested on patients, and the response of each patient was Bernoulli depending on the underlying success probability, p, of the treatment. The success probability was unknown but assumed to follow a prior distribution π_0 at the beginning of the phase II trial. Based on this set-up, Stallard obtained optimal one-stage, two-stage, and fully sequential designs that maximize the overall expected gains. He found in his example (Stallard, 1998, Section 3) that the overall expected gains from the three optimal designs were very similar.

Stallard (1998) assumed that a treatment that passes phase II testing will be marketed only if its success probability, p, can be shown to be superior to the success probability p_0, of a standard in a phase III trial. However, he showed that the probability of passing a treatment with $p = p_0$ to phase III testing was 0.65 for his optimal fully sequential design and that this probability was 0.71 for his optimal one-stage design. Since any reasonable phase III design would reject a treatment with $p = p_0$ with high probability, why would an optimal phase II design choose to pass such a treatment to phase III testing with such a high probability?

The second unusual result from Stallard's (1998) article is in its evaluation of other existing designs. In particular, Stallard evaluated another Bayesian design suggested earlier by Thall and Simon (1994). Thall and Simon's design does not directly optimize gains but rather aims to control the posterior probabilities of accepting promising and rejecting nonpromising treatments. In his example, Thall and Simon's design had a much lower probability of passing a treatment with $p = p_0$ to phase III. But in terms of overall expected gain, Thall and

Simon's design was far inferior to any of Stallard's optimal designs. In fact, Stallard showed that Thall and Simon's design had even smaller expected gain than not carrying out a phase II trial altogether!

The goal of this article is to study and explain these results. We believe that the time to carry out the study to be an important component in a trial design, an issue that is not considered in Stallard's (1998) gain function. We suggest a design that maximizes the rate of gain, and we show that, when using this new gain function, the peculiarities described above disappear. The optimal design based on our new gain function will have only a small probability of passing an ineffective treatment for further study.

2. Maximizing the Rate of Gain

For ease of illustration, in this article, we only consider one-stage designs. We first briefly discuss Stallard's (1998) design. Suppose the phase II study consists of n patients. Let S_n be a random variable denoting the number of successes in the n patients and let s^* be the cut-off value to pass a treatment to phase III testing. Stallard's design is to find (n, s^*) by maximizing the following expected gain before the trial:

$$G_{\text{Stallard}} = -nk + \sum_{s=s^*}^{n} \{-m + l(n)\mathrm{E}(\kappa(p) \mid s, n)\}\Pr(S_n = s),$$

where k denotes the cost of testing each patient in the phase II study, m denotes the fixed set-up cost of the phase III study, $\kappa(p)$ denotes the probability that the phase III study will indicate the treatment to be effective when its success probability is p, and $l(n)$ denotes the potential gain from the phase III study,

$$l(n) = \begin{cases} l_0 - \lambda_1 n, & n \le n_0 \\ (l_0 - \lambda_1 n_0)\exp\{-\lambda_2(n - n_0)\}, & n > n_0. \end{cases} \qquad (1)$$

Stallard's (1998) design makes two implicit assumptions. First, the gain from a potentially successful phase III trial is immediately available upon completion of the phase II study. Second, each successful phase III trial is expected to have a constant profit. Because of these, the more phase III trials being conducted, as long as the expected gain in each is positive, the better. This explains why Stallard's design is much more profitable than Thall and Simon's (1994) design, which has a low overall probability of passing treatments.

We believe that Stallard's (1998) framework is reasonable but that his gain function is inappropriate. In this article, we suggest an alternative design. The design takes into consideration information that has been ignored, but that is readily available, in Stallard's set-up. This information is the time to carry out the phase III study.

We consider the general situation where a series of phase II trials, up to M, can be studied and one can stop at anytime before M and pass a promising treatment to a phase III study. The general set-up would make the optimal design appropriate to situations where drug companies may want to only pass promising treatments to phase III testing (e.g., Wang and Leung, 1998).

Assume that each phase II trial requires n patients and a phase III trial requires N patients. We will show later that N can be readily obtained from Stallard's (1998) set-up. We also assume that the time to carry out a trial is proportional to the number of patients required, so the time to carry out a phase II trial is n and that for a phase III trial is N. If a treatment passes the phase II study (with $S_n = s \ge s^*$), the potential gain from a phase II trial (with a possible successful phase III study) is

$$\begin{aligned} G_{\text{III}} &= \sum_{s=s^*}^{n} \{-m + l(n)\mathrm{E}(\kappa(p) \mid s, n)\}\Pr(S_n = s \mid S_n \ge s^*) \\ &\quad - m + \frac{l(n)\sum_{s=s^*}^{n} \mathrm{E}(\kappa(p) \mid s, n)\Pr(S_n = s)}{1 - \Pr(A)}, \end{aligned}$$

where $\Pr(A)$ denotes the probability of abandoning a phase II study. $\Pr(A)$ is a function of (n, s^*) and may be different for different designs. If a phase II study accepts a treatment for a phase III study after τ phase II studies (each with n patients), the total gain up to completion of the phase III study is

$$g_\tau = -nk\tau + G_{\text{III}}. \qquad (2)$$

Consider the truncated stopping time $\tau_M = \min(\tau, M)$, i.e., the study is stopped at τ or M if none of the M phase II trials is accepted for a phase III study. The probability of carrying out the jth phase II trial is $\Pr(\tau \ge j) = p^{j-1}(A)$, and the expected total gain becomes

$$g_{\tau_M} = -nk\mathrm{E}(\tau_M) + \left\{1 - \Pr^M(A)\right\} G_{\text{III}}.$$

Particularly, when $M = 1$, we have $\tau_1 = 1$ (one-step-look-ahead) and

$$\mathrm{E}(g_{\tau_1}) = -nk + \sum_{s=s^*}^{n} \{-m + l(n)\mathrm{E}(\kappa(p) \mid s, n)\}\Pr(S_n = s). \qquad (3)$$

This is in fact the objective function Stallard (1998) used, i.e., $\mathrm{E}(g_{\tau_1}) = G_{\text{Stallard}}$.

Although the expected gain from a phase III study is taken into account, the time and effort is not properly taken into account in (3). Based on (3), any treatment with a potential profit from marketing that can cover the set-up cost (m) will be passed to a phase III study. This is not desirable because phase II studies are to screen and recommend treatments with maximum gain (not treatments with a small positive gain).

If we consider the time, N, to carry out the phase III study, the average time for Stallard's (1998) design is

$$n + N\{1 - \Pr(A)\}. \qquad (4)$$

We now go back to the general case where M is any integer. The expected gain per unit time (gain rate) is

$$\begin{aligned} R_\tau &= \frac{\mathrm{E}(g_{\tau_M})}{n\mathrm{E}(\tau_M) + \left\{1 - \Pr^M(A)\right\} N} \\ &= \frac{-nk\frac{1-\Pr^M(A)}{1-\Pr(A)} + \left\{1 - \Pr^M(A)\right\} G_{\text{III}}}{n\frac{1-\Pr^M(A)}{1-\Pr(A)} + \left\{1 - \Pr^M(A)\right\} N} \\ &= \frac{-nk + \{1 - \Pr(A)\}G_{\text{III}}}{n + \{1 - \Pr(A)\}N}. \end{aligned}$$

This gain rate is very similar to the Gittins index for completing different projects (Gittins, 1989, p. 25). It is more reasonable to maximize this index R_τ. Note also that R_τ is independent of M.

In maximizing $R_{\mathcal{T}}$, we need to know the value of N, which is readily available from Stallard's (1998) set-up. In his set-up, he assumed that the phase III trial is to detect a difference from p_0 of a standard treatment to $p_0 + \delta$ with a two-tailed significance test at α-level with a power of $1-\beta$. Given $p_0, \delta, \alpha, \beta$, the sample size N can be obtained using a formula originally by Whitehead (1993), as suggested by Stallard.

We now compare our design to Stallard's (1998) design as well as to Thall and Simon's (1994) design using the example from Stallard's Section 3. In that example, the following parameters were used: $k = 0.5$ (million dollars), $m = 200$ (million dollars), $l_0 = 7400$ (million dollars), $\lambda_1 = 5$ (million dollars), $\lambda_2 = 0.00173$, $p_0 = 0.2$, $\delta = 0.15$, $\alpha = 0.05$, and $\beta = 0.1$. The prior $\pi_0(p)$ is beta(a, b) distributed with $a = 0.845, b = 10 - a$. Given this information, we find that $N = 350$. The optimal design by maximizing $R_{\mathcal{T}}$ gives $(n, s^*) = (10, 4)$. Stallard's design gives $(n, s^*) = (29, 5)$.

Our optimal design gives a value of $R_{\mathcal{T}} = \$4.37$ million, which is the maximum expected gain per unit time (patient) in the trial under the assumptions of the example. Using this design, the value of $\Pr(A)$ is 0.955, which means that over 95% of the treatments will not be passed to a phase III study. The passing criteria seem to be in line with the prior assumption that only 10% of the treatments have success probability $p \geq p_0$ (and fewer still with $p = p_0 + \delta$). More important, the probability of passing a treatment with $p = p_0$ is now 0.12, compared with 0.71 using Stallard's (1998) one-stage design. The expected gain per unit of time (patient) from Stallard's design can be calculated from (3), (4), and $\Pr(A) = 0.8$, based on $(n, s^*) = (29, 5)$, as

$$\begin{aligned} G_{\text{Stallard}}/(n + N(1 - \Pr(A))) &= 222.3/(29 + 350(1 - 0.8)) \\ &= \$2.2 \text{ million.} \end{aligned}$$

So our expected gain per unit time is twice as that of Stallard's. The reason for this improvement can be seen by considering the following. A treatment is marketable if $p \geq p_0 + \delta = 0.35$. The number of patients that needs to be tested before a marketable treatment can be found is $N_{\mathcal{T}} = (n/(1 - P(A)) + N)/p_{++}$, where p_{++} is the probability of passing a treatment that is eventually marketable. For Stallard's (1998) design, $N_{\mathcal{T}} = (29/0.2 + 350)/0.072 = 6875$ patients; for our design, $N_{\mathcal{T}} = (10/0.045 + 350)/0.21 = 2725$ patients. Therefore, our design requires only 40% of the number of patients used in Stallard's design, and consequently, our design has a higher rate of gain.

Using our design, we found that the probability of passing a borderline effective treatment (one with $p = p_0 + \delta$) is only 0.49—quite a bit smaller than one would conventionally wish and much smaller than Stallard's (1998) value of 0.99. But a moment's reflection would help to explain this behavior. In the current example, the prior suggests that there are many more ineffective treatments than effective ones. Therefore, if one wants to have a high probability of passing an effective treatment, one must also be prepared to pass many other ineffective ones, none of which will yield any gains. So a good design can only pass a treatment when there is sufficient evidence that it is effective. To illustrate this last point, we see that the probability of acceptance using our design rises to 0.83 for a treatment with $p = 0.5$.

We note that our optimal design recommends a much smaller n (10) than that recommended by Stallard (1998) ($n = 29$) in this example. This is because the prior suggests that only few treatments will have $p \geq p_0 + \delta$. Therefore, a large phase II trial would mean a large number of patients will be wasted on nonpromising treatments. If we had used $n = 29$ for our design in this example, then the optimal s^* would be eight with the rate of gain decreased to \$3.4 million, but still an improvement of 50% over Stallard's design (Table 1).

Table 1
Value of Stallard's (1998) *design, Thall and Simon's* (1994) *design, and the new design proposed in this article under different objectives*

		Objectives[a]			
Design	(n, s^*)	G_{Stallard}	$R_{\mathcal{T}}$	$N_{\mathcal{T}}$	α, β
Stallard	(29, 5)	222.2	2.2	6875	0.71, 0.01
New design	(10, 4)	112.4	4.4	2725	0.12, 0.51
New design[b]	(29, 8)	180.2	3.4	3889	0.21, 0.15
Thall and Simon	(100, 28)	138.3	1.2	8247	0.03, 0.06

[a] G_{Stallard}, overall gain based on Stallard; $R_{\mathcal{T}}$, rate of gain based on this article; $N_{\mathcal{T}}$, number of patients required before a marketable treatment is found; α, Pr(passing a treatment with efficacy p_0 to phase III); β, Pr(abandoning a treatment with efficacy $p = p_0 + \delta$).

[b] New design with same value of n as Stallard's design.

Thall and Simon's (1994) design gives $(n, s^*) = (100, 28)$ when the parameters p_U and p_L in their method are set to 0.95 and 0.05, respectively. In comparison, the expected gain per unit time for Thall and Simon's design is \$1.3 million and $N_{\mathcal{T}} = 8247$. These unfavorable results are due to the fact that Thall and Simon's method is not designed to optimize gain. Moreover, we must acknowledge that Table 1 is not suggesting that one should recommend a phase II trial with $n = 100$.

3. Conclusion

The purpose of this article is to gain a better understanding of Stallard's (1998) optimal design, which calls for a high probability of passing ineffective treatments for further study. In the case of a one-stage design, we showed that, by taking into consideration the time to carry out the phase III trial and by maximizing the gain per patient, the optimal design has a much smaller chance of passing a treatment with little possibility of being successful beyond a phase II study. Using our optimal one-stage design, it is shown that the expected gain per unit time (patient) is twice as much as that using Stallard's one-stage design.

Our work is based on the consideration that a number of candidate treatments are available for development, such as the case in a drug development program in a company or for a particular disease. As such, the conclusions drawn are limited to these situations. There are other issues related to this type of study. We refer the readers to the review in Stallard's (1998) paper.

Acknowledgements

We are grateful to the associate editor and the referees for their insightful comments and suggestions. This research was supported in part by NIH grants CA74207 and CA47179 for DHL and AI24643 for YGW.

Résumé

Dans une récente livraison de Biometrics (1998, *Biometrics* **54,** 279–294), Stallard a utilisé l'approche bayésienne de la théorie de la décision pour le calcul d'effectif dans le cadre d'essais cliniques de Phase II. Sa méthode maximise l'espérance du gain financier lors du développement d'un nouveau traitement. Cependant, elle entraîne aussi une très forte probabilité (0.65) de recommander le passage en Phase III d'un traitement non efficace. D'un autre côté, elle correspond à une espérance de gain de plus de 10 fois celle associée à une méthode qui contrôle étroitement le risque d'erreur de première espèce (Thall et Simon, 1994, *Biometrics* **50,** 337–349). La méthode de Stallard maximise l'espérance du gain par essai de Phase II mais elle ne maximise pas le gain totalisé sur une période de temps donnée car le gain total dépend de la proportion de traitements passant en Phase III. Nous suggérons plutôt de maximiser le gain total, dans le cas d'un essai en une seule étape (essai non séquentiel) la méthode optimale correspondante est deux fois plus efficace en terme de gain total que la méthode de Stallard. De plus, notre méthode a une probabilité de seulement 0.12 de faire passer un traitement non efficace en Phase III.

References

Gittins, J. C. (1989). *Multi-Armed Bandit Allocation Indices.* Chichester, U.K.: Wiley.

Stallard, N. (1998). Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics* **54,** 279–294.

Thall, P. and Simon, R. (1994). Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* **50,** 337–349.

Wang , Y. G. and Leung, D. H. Y. (1998). An optimal design for screening trials. *Biometrics* **54,** 243–250.

Whitehead, J. (1993). Sample size calculation for ordered categorical data. *Statistics in Medicine* **12,** 2257–2271.