Published in final edited form as:

Contemp Clin Trials. 2013 May ; 35(1): 170–178. doi:10.1016/j.cct.2013.03.006.

Optimal Continuous-Monitoring Design of Single-arm Phase II Trial Based on the Simulated Annealing Method

Nan Chen and J. Jack Lee*

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, U.S.A.

Abstract

Simon's two-stage design is commonly used in phase II single-arm clinical trials because of its simplicity and smaller sample size under the null hypothesis compared to the one-stage design. Some studies extend this design to accommodate more interim analyses (i.e., three-stage or four-stage designs). However, most of these studies, together with the original Simon's two-stage design, are based on the exhaustive search method, which is difficult to extend to high-dimensional, general multi-stage designs. In this study, we propose a simulated annealing (SA)-based design to optimize the early stopping boundaries and minimize the expected sample size for multi-stage or continuous monitoring single-arm trials. We compare the results of the SA method, the decision-theoretic method, the predictive probability method, and the posterior probability method. The SA method can reach the smallest expected sample sizes in all scenarios under the constraints of the same type I and type II errors. The expected sample sizes from the SA method are generally 10–20% smaller than those from the posterior probability method or the predictive probability method, and are slightly smaller than those from the decision-theoretic method in almost all scenarios. The SA method offers an excellent alternative in designing phase II trials with continuous monitoring.

Keywords

Simulated annealing; Simon's design; Early stopping; Adaptive design; Bayesian inference; Phase II trial; Type I error; Type II error; Optimal design

1 Introduction

Simon's two-stage design has been extensively applied in phase II cancer trials because of its simplicity and good operating characteristics, such as the smaller sample size required compared to the one-stage design under the null hypothesis [1]. In this design, the primary endpoint is binary, e.g., response or no response. A single-arm trial is performed to evaluate whether a new treatment is more efficacious than the standard treatment. Early stopping due to futility is possible but no early stopping for efficacy is allowed in the design.

Crown Copyright © 2013 Published by Elsevier Inc. All rights reserved.

*Corresponding author at: Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Unit 1411, Houston, Texas 77030, U.S.A., Tel: 1-713-794-4158, jlee@mdanderson.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Although it is logistically easier to conduct a two-stage design than a multi-stage ($n > 2$) design in practice, the interim monitoring schemes are more flexible in multi-stage or continuous monitoring designs, which can lead to smaller sample sizes. Ensign et al. proposed a three-stage design in which the trial is stopped if none of the patients respond to the treatment in the first stage [2]. Chen extended Simon's two-stage design to a three-stage design that does not have this first stage constraint [3]. Chen and Shan proposed a three-stage design that includes both futility and efficacy early stopping rules [4].

Most of these methods, including Simon's original design, are based on the exhaustive search algorithm that searches the entire space of the possible stopping boundary domain to find the optimal design. In the search process, the sample size and early stopping boundary in each stage is considered as a variable. Each variable is assigned a calculation loop and all calculation loops are embedded in a nested fashion to search for the solution with the smallest expected sample size in the entire parameter space. The implementation of the exhaustive search method is straightforward and works well if the number of variables is small. But it is difficult to extend this algorithm to high-dimensional problems. The calculation time increases exponentially with an increase in the number of variables, making it difficult to apply the exhaustive search method to high-dimensional multi-stage or continuous monitoring designs.

Alternatively, the Bayesian framework can be applied to construct multi-stage or continuous monitoring designs. Two commonly used methods are the posterior probability method and the predictive probability method. For the posterior probability method, at each interim monitoring point, the posterior distribution of the treatment's efficacy rate, calculated based on the current patients' outcomes, is compared with a pre-specified threshold value to determine whether to stop or continue the trial. For the predictive probability method (e.g., Lee and Liu [5]), the predictive probability, obtained by predicting the remaining patients' outcomes based on the current observed information, is compared with the pre-specified threshold value to inform early stopping decisions. These methods are not optimized for minimizing the expected sample sizes.

The simulated annealing method has been extensively used in many research areas as a general optimization method [6]. The method is essentially a stochastic optimization process based on the Metropolis algorithm. To utilize this method, an initial configuration of a system is required and the Metropolis algorithm randomly perturbs this configuration to minimize the energy of the system. The temperature constant in the Metropolis algorithm is gradually reduced and the final equilibrium state will have minimum energy, corresponding to the optimal solution of the problem at hand. The name of this method comes from the analogy of the physical annealing process in which the temperature of a substance is gradually reduced until an equilibrium state is reached. Geman and Geman applied this method to image processing [7]. Sheldrick carried out crystallographic phase structure calculation using the SA method [8]. Based on a method related to the simulated annealing algorithm, Müller et al. developed a framework for clinical trial design optimization [9].

In this study, we propose to apply the simulated annealing method to searching for the optimal design of multi-stage or continuously monitored single-arm trials. The basic idea behind this method is simple: the expected sample size is a function of the sample size and the early stopping boundary of each stage in a multi-stage design. We consider the stopping boundary and the corresponding expected sample size as the configuration and the corresponding energy of a system, respectively. We then apply the simulated annealing method to this system to find the optimal configuration (sample size and early stopping boundary in each stage) and the corresponding minimum energy (expected sample size).

2 Method for Phase II Design

Although a new hypothesis setting is suggested in a recent publication [10], in this study we use the one from Simon's two stage design due to its popularity. Let p denote the true response rate of a new treatment. A phase II clinical trial design with a binary outcome often requires testing a null hypothesis $H_0: p = p_0$ against an alternative hypothesis $H_1: p = p_1$, where p_0 and p_1 are the null response rate of the standard treatment and the target response rate of the new treatment, respectively. It is also required that the probability $P(\text{accept the new treatment} | H_0)$ is no larger than α (type I error) and the probability $P(\text{reject the new treatment} | H_1)$ is no larger than β (type II error). Please note that the above commonly used hypothesis test notation in fact refers to test $H_0: p = p_0$ vs $H_1: p > p_0$, but the statistical power is evaluated at $p = p_1$.

Let $N_{\max}$ denote the maximum number of patients for an s stage ($s \geq 2$) single-arm trial, and n_i ($i = 1, \dots, s-1$) denote the number of patients at each interim monitoring point. If the new treatment is not effective, we want to stop the trial as early as possible. Otherwise, the trial is continued until the number of treated patients reaches $N_{\max}$. Similar to Simon's two-stage design, only futility stopping is allowed, and no efficacy stopping is implemented in the design. At each interim monitoring point, if the number of patients with positive response x_i satisfies $x_i \leq r_i$ where r_i is the rejection boundary, we reject the new treatment and accept the null hypothesis (subject to the type II error β). Otherwise we continue the trial to the next stage until the maximum sample size $N_{\max}$ is reached. At the final decision point $n_s = N_{\max}$, we compare the number of patients with positive response x_s with the rejection boundary r_s to make the decision. If $x_s \leq r_s$, we reject the new treatment.

For a given response rate p of a new treatment, the probability of having m positive responses at stage g , represented by $C_g(m, p)$, is calculated by applying a recursive formula given by Shultz et al.[11]

$$C_g(m, p) = \sum_{j=E}^m C_{g-1}(j, p) \text{bin}(m-j; p, n_g - n_{g-1}), \quad (1)$$

where $E = \max(r_{g-1} + 1, m - n_g + n_{g-1})$ and bin denotes the binomial probability function. Thus, the probability of rejecting the new treatment and accepting the H_0 at stage g is given by

$$R_g(p) = \sum_{m=r_{g-1}+1}^{r_g} C_g(m, p) \quad (2)$$

and the cumulative probability of rejecting the new treatment is then

$$R(p) = \sum_{g=1}^s R_g(p). \quad (3)$$

The type I and type II errors are $1 - R(p_0)$ and $R(p_1)$, respectively. Other operating characteristics such as expected sample size EN and probability of early termination PET can be easily derived from these formulas.

Based on the previous analysis, the operating characteristics of the single-arm trial designs are determined by the interim monitoring of the patient number vector $\mathbf{n} = (n_1, n_2, \dots, n_s)$, rejection boundary vector $\mathbf{r} = (r_1, r_2, \dots, r_s)$, p_0 and p_1 . An optimal design to minimize the expected sample size under the null hypothesis is described as the solution of

$$\underset{\mathbf{n}, \mathbf{r}}{\operatorname{argmin}} EN_{p_0}(p_0, p_1, \mathbf{n}, \mathbf{r}) \quad (4)$$

under the pre-specified type I and type II error constraints. Using the exhaustive search method, Simon obtained the solution with $s = 2$, which is frequently used in designing phase II cancer clinical trials. However, it is very difficult or even impractical to use the exhaustive search method for problems with a large value of s because the calculation time increases exponentially with an increase in the s value. In this study, to simplify this problem, we fix the value of the vector $\mathbf{n}$ at a pre-specified value and use the simulated annealing method to optimize only the rejection boundary vector $\mathbf{r}$. Users need to specify the interim monitoring points in this design, which makes the calculation easier. The problem thus changes to one of optimizing the rejection boundaries $\mathbf{r}$:

$$\underset{\mathbf{r}}{\operatorname{argmin}} EN_{p_0}(p_0, p_1, \mathbf{n}, \mathbf{r}) \quad (5)$$

under the pre-specified type I and type II error constraints.

2.1 Simulated Annealing Method

The simulated annealing algorithm is based on a stochastic process that consists of many iteration steps. In each iteration step, a new rejection vector is proposed in the form of a small perturbation on the previous one. For each component r_i of the vector $\mathbf{r}$, the new attempt is generated by $r_i^* = r_i + \Delta r_i$, where Δr_i is an integer uniform random variable in the range $(-d, \dots, d)$ and d is the maximum jump distance. If the operating characteristics, based on the new rejection boundary vector, do not satisfy the type I and type II error constraints, the algorithm rejects the new attempt and proceeds to the next iteration step. Let $\mathbf{r}^*$ denote the new attempt of the rejection boundary vector and $\Delta EN = EN_{p_0}(\mathbf{r}^*) - EN_{p_0}(\mathbf{r})$ denote the change in the expected sample size, caused by the new attempt of the rejection boundary vector. The new attempt $\mathbf{r}^*$ is accepted or rejected according to the following rules: if $\Delta EN < 0$, then the algorithm accepts the new attempt, otherwise it accepts the new attempt with the probability

$$p(\Delta EN) = \exp(-\Delta EN/T), \quad (6)$$

where T is the temperature corresponding to the fluctuation amplitude of the expected sample size in the stochastic search. The process gradually reduces the temperature until it reaches the lowest temperature that will obtain the optimal solution. The parameters of choice are discussed in the next section. At a high temperature, the fluctuation amplitude of the expected sample size is large and it is easy for the sampling solution to escape the local minima.

When the pre-specified temperature and the number of iteration steps reach the pre-specified values, the rejection boundary vector with the smallest EN_{p_0} is the solution. Because the simulated annealing method is based on a stochastic optimization process, the solutions obtained from different runs can be different. In order to solve the convergence problem, we carry out multiple runs with different annealing paths (different random number generation sequences). If the variance of the solutions is small compared to the precision required for the practical application (e.g., a standard deviation value of 0.1 for a sample size estimation), we consider the solution to be converged and select the one with the smallest expected sample size as the optimal solution. Otherwise, we need to increase the number of iteration steps or change other parameters such as the temperature settings and maximum jump distance d .

The simulated annealing method requires an initial system configuration for the further stochastic annealing process. Each component $r_i (i = 1, \dots, s - 1)$ of the initial rejection boundary vector $\mathbf{r}$ is assigned -1 such that no early termination is allowed with the initial configuration. The initial value of the final decision boundary r_s is given by $\text{Bin}^{-1}(1 - \alpha; n_g, p_0)$, where Bin^{-1} is the inverse cumulative binomial distribution function. With this initial setting, the type I error constraint is satisfied. In an annealing iteration step, if the operating characteristics of an attempted rejection boundary vector do not satisfy the type I and type II error constraints, the new attempt will be rejected. When the temperature is low (e.g., 0.1), the sampling rejection vectors may be stuck at the local minima. The objective function (expected sample size) is not a smooth function of the early stopping boundaries and the change on the stopping boundary is no less than 1 (at least one patient). Kirkpatrick suggested that if the objection function of a system is not smooth, it may be more effective to use a moderate temperature and record the optimal solution in the search [12]. In this study, we fix the temperature value at 0.5, corresponding to 0.5 on the expected sample size change, to perform the search. The total number of iteration steps depends on how many stages there are in the design. A design with a larger number of stages has a larger number of iteration steps. In two-stage designs, the total number of iteration steps is set at 25, 000. In multi-stage ($s > 15$) designs, large numbers (4×10^8) of iteration steps are used. The maximum jump distance d is 20 and 12 for two-stage and multi-stage designs, respectively. In multi-stage ($s > 15$) designs, after the first 2×10^6 steps of iterations, the value of d reduces to 5 for fast convergence. The parameters are determined by the convergence plots from multiple runs.

2.2 Bayesian Decision-Theoretic Method

In this study, we also implement a Bayesian decision-theoretic method for comparison purposes. In order to minimize the expected sample size under type I and type II error constraints, the loss function used in the decision-theoretic method consists of the cost of enrolled patients and the cost of wrong decisions. Let

$$L_{total} = n + L_d, \quad (7)$$

where n is the number of patients enrolled and L_d is the cost associated with wrong decisions. The L_d term is a constant loss function that is similar in form to that applied by Lewis and Berry [13]. However, in this study, rather than one, we used two parameters (k_1 , k_2) in the loss function to control the costs associated with the type I and type II error rates separately. The trial may be terminated at interim monitoring points and we use (A) and (R) to indicate the decisions of accepting and rejecting the null hypothesis (subject to the type II and type I errors, respectively), respectively. The cost for accepting the null hypothesis action is defined as

$$L_d(p, A) = 0 \text{ if } p \leq p_1 \quad (8)$$

$$= k_1 \text{ if } p > p_1 \quad (9)$$

and the cost for rejecting the null hypothesis action is

$$L_d(p, R) = 0 \text{ if } p > p_0 \quad (10)$$

$$= k_2 \text{ if } p \leq p_0. \quad (11)$$

At an interim monitoring point $g (g < s)$, let n_g and m denote the number of enrolled patients and the number of patients with positive responses, respectively. The prior probability

density function (pdf) of p follows a $\text{beta}(a, b)$ distribution and we select $a = 1$, $b = 1$ in our studies. Due to the conjugate property of the beta distribution, the posterior distribution of the response rate follows a $\text{beta}(a + m, b + n_g - m)$ distribution. Because there is no efficacy stopping in the design, the stopping loss at stage g is

$$L_g^{\text{stop}}(m) = n_g + EL_d(p, A) \quad (12)$$

$$= n_g + k_1 [1 - \text{Beta}(p_1; a+m, b+n_g - m)], \quad (13)$$

where EL_d is the expected value of L_d and Beta is the cumulative probability of the beta distribution. At the final decision point, with m positive responses,

$$EL_d(p, A) = k_1 [1 - \text{Beta}(p_1; a+m, b+N - m)] \quad (14)$$

and

$$EL_d(p, R) = k_2 \text{Beta}(p_0; a+m, b+N - m). \quad (15)$$

And the loss function at the final decision stage s is

$$L_s(m) = \min[N_{\max} + EL_d(p, R), N_{\max} + EL_d(p, A)]. \quad (16)$$

After we calculate the loss function value for each m at the final stage, we use the backward induction method to calculate the cost associated with the continuing action at each stage g

$$L_g^{\text{cont}}(m) = \sum_{k=m}^{\min(n_{g+1}, n_{g+1} - n_g + m)} L_{g+1}(k) \text{betaBinom}(k - m, n_{g+1} - n_g, a+m, b+n_g - m), \quad (17)$$

where $\text{betaBinom}(a, b, c, d)$ is the pdf of a beta-binomial distribution (the probability of obtaining a responses from b patients based on the response rate following a $\text{beta}(c, d)$ distribution). Therefore the loss at stage g is given by

$$L_g(m) = \min[L_g^{\text{cont}}(m), L_g^{\text{stop}}(m)]. \quad (18)$$

Based on this formula, we can calculate the value of L_g from L_{g+1} , and then the value L_{g-1} from L_g until we reach the first stage. At stage g , if $L_g^{\text{cont}}(m)$ is less than $L_g^{\text{stop}}(m)$, we continue the trial. Otherwise, we terminate the trial early. After we reach the decisions at each point, we determine the rejection boundaries based on the decisions we make at each point.

In this decision-theoretic framework, there are two constants k_1 and k_2 in the loss function. We search the entire two-dimensional parameter space to find the pair corresponding to the minimum expected sample size $E(N_{p_0})$ under both type I and type II error constraints. In order to speed up the search process, we perform a multiple-stage search with gradually reduced searching step sizes and searching areas to find the solution.

2.3 Other Comparison Methods

In Bayesian clinical trial designs, the posterior probability method is frequently applied for interim monitoring and early stopping. We use this method for comparison purposes in this study. The posterior probability method requires two posterior threshold values: one for early termination (θ_L), a threshold for declaring a lack of efficacy, and one for the final

decision inference (θ_T), a threshold for declaring efficacy. Similar to the decision-theoretic method, we search the entire parameter space to find the solution corresponding to the minimum expected sample size $E(N|p_0)$ under both type I and type II error constraints.

The predictive probability method we used in this study is based on Lee and Liu's study [5] and the software for the calculation is obtained from the MD Anderson software download site (https://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=84). Simon's two-stage design calculation is performed using software from the National Cancer Institute (<http://linus.nci.nih.gov/~brb/Opt.htm>).

3 Results

3.1 Two-stage Designs

We applied the simulated annealing method to the two-stage designs and compared the results with those from Simon's optimal two-stage design to check for accuracy. We tested all scenarios (optimal designs with response difference $\delta = 0.2$) listed in Table 1 of Simon's paper. The results are presented in the Supplemental Table 1. The number of patients at interim monitoring n_1 and the total number of patients n are from Simon's reference. The stopping boundaries and the final decision boundaries are calculated using the simulated annealing method. The results demonstrate that the simulated annealing method successfully reproduced all stopping boundaries, final decision boundaries, and operating characteristics from Simon's two-stage designs. The medians of the numbers of iteration steps to the minima are within 2000–3000 steps.

3.2 Continuous Monitoring Designs with Fixed N_{max}

In this section we compare the expected sample sizes calculated from different methods: the simulated annealing (SA), decision-theoretic (DT), predictive probability (PP), and posterior probability (PT) methods, and Simon's two-stage design. Continuous monitoring starts after the 10th patient receives treatment and proceeds after each new patient receives treatment until early stopping occurs or the maximum number of patients is reached. Two different type I/II error constraint combinations are evaluated in the study: the constraint with $\alpha = 0.1$, $\beta = 0.1$ (Table 1) and $\alpha = 0.05$, $\beta = 0.2$ (Supplemental Table 2), respectively. The maximum sample sizes N_{max} in Table 1 and Supplemental Table 2 are the same as those from Simon's two-stage designs.

From Table 1 and Supplemental Table 2 we observe that the SA method yields the smallest expected sample sizes for almost all (34 of 35) scenarios. In the only entry in Table 1 ($p_0 = 0.5$, $p_1 = 0.7$, Section A), the expected sample size obtained by the SA method is very close to the smallest one obtained by the DT method (25.78 vs. 25.75). We also performed an SA calculation with a large number (1.35×10^{10}) of iteration steps for this scenario (it took about 65 hours on a workstation with Intel Xeon 3Ghz CPUs), the SA method did find the same solution as the one obtained from the DT method. The DT method yields slightly larger expected sample sizes than the SA method for most scenarios, with one exception ($p_0 = 0.6$ of Table 1 Section A), in which the expected sample size (21.36) obtained from the DT method is 7.7% larger than the one (19.83) obtained from the SA method. The expected sample sizes obtained by the SA method are generally 10–20% smaller than those obtained by the PP method or the PT method. The expected sample sizes obtained by Simon's two-stage design are generally larger than the results obtained by the SA method, the DT method, and the PP method, but some of them are smaller than the results obtained by the PT method.

The type I/II error rates corresponding to the scenarios in Table 1 are summarized in Table 2. Most of the type I/II error rates obtained by different methods are close to the pre-

specified constraints. The error rates calculated by the SA method and the DT method are slightly larger than those obtained by other methods while the expected sample sizes calculated by the SA and the DT methods are smaller than those obtained by other methods, which implies that higher error rates are the cost of reducing the expected sample sizes. Some error rates obtained by the PP method and the PT method are low compared to the prespecified error rate constraints (10%), for instance, the type I error rate of the PT method with $p_1 = 0.1$ and $p_2 = 0.3$ (Table 2 Section B) is only 4.93% due to the discrete nature of the binomial distribution.

We compare the rejection boundaries from different methods and present the results in Figure 1a and Figure 1b, corresponding to the scenarios of minimax designs with $p_0 = 0.2$ and 0.3, respectively. We observe similar tendencies in both figures. The rejection boundaries obtained by the SA method and the DT method are close to each other, which explains why the expected sample sizes of the two methods are close to each other. The rejection boundaries obtained by the PP method and the PT method are lower than that obtained by the SA method, resulting in larger expected sample sizes than those obtained by the SA method.

In the rejection boundaries obtained by the PT method, there is a big gap between the late-stage rejection boundary values and the decision boundary, which is caused by the two different posterior probability threshold values applied for early stopping and a final decision, respectively. This gap may cause unnecessary wasting of patient resources in a trial. For instance, in a trial based on the PT design in Figure 1b, the final decision boundary is 16/42 and at the interim analysis point of 36 patients, the rejection boundary is 8/36. If we suppose the outcomes of 9 of the 36 patients are positive after the 36th patient, then we should continue the trial according to the rejection boundary obtained by the PT method. However, no matter what results we obtain from the remaining 6 patients, we always accept the null hypothesis because the final decision boundary is 16/42, which means that the outcomes of the remaining 6 patients have no impact on the final conclusion. Thus, the information obtained from treating the remaining 6 patients is wasted.

The PP method, stopping the trial by predicting the possible responses of the remaining patients, avoids this problem with a continuous stopping boundary until the final inference. However, during the trial's early stage, the rejection boundaries obtained by the PP method are low due to the high uncertainty of the remaining patients, leading to relatively large expected sample sizes compared to those obtained by the SA method.

3.3 Searching for the Optimal Simulated Annealing Design with Continuous Monitoring by Varying N_{max}

Under given type I and type II error constraints, we define the optimal design as the one that corresponds to the smallest expected sample size. In order to obtain an optimal design using the SA method, we gradually reduce the maximum number of patients N_{max} from a large value to a small one, and run the simulated annealing calculation with each N_{max} to find the one with the smallest expected sample size. For instance, we gradually reduce the value of N_{max} from 50 to 25 with type I and type II error constraints of (0.10, 0.10) and run 5×10^7 iteration steps using the SA method at each value of N_{max} . We set up the initial rejection boundary with no early stopping for the first $N_{max} = 50$ patients and use the previously calculated rejection boundary as the initial early stopping boundary for the subsequent N_{max} . The results for different scenarios using the SA method are presented in Table 3. The expected sample sizes of the optimal designs obtained by the SA method are smaller than those from Simon's two-stage designs and all the other designs shown in Table 1 for all scenarios. We also find that the optimal maximum sample sizes N_{max} of the continuous monitoring designs are close to but different from those of Simon's two-stage design.

In order to study the effects of the value of maximum sample sizes N_{max} on the resulting expected sample sizes of the SA designs, we applied the continuous monitoring SA designs with different N_{max} values for three testing scenarios: (1) $p_0 = 0.1$, $p_1 = 0.3$, (2) $p_0 = 0.2$, $p_1 = 0.4$, and (3) $p_0 = 0.3$, $p_1 = 0.5$, under the type I/II error constraints of $\alpha = 0.1$, $\beta = 0.1$. Interim monitoring starts after the 10th patient and the largest maximum sample size N_{max} is 50. The expected sample sizes of the SA method, along with the results of Simon's two stage designs, are presented in Figure 2. In each curve, the expected sample size at the smallest N_{max} is relatively large. However, the values of the expected sample sizes become smaller with increase of the N_{max} values. This tendency is very similar to Simon's two stage design that has a relatively large expected sample size for the minimax design. When the value of the N_{max} is larger than the critical value from the optimal design, the expected sample size gradually increases with the increase of N_{max} . The expected sample sizes are not a monotone function of the N_{max} values and fluctuates with the increase of the N_{max} values, which makes the N_{max} selection more complicated. The selected value of the N_{max} should be smaller than the one from the optimal design, otherwise both the expected sample size and the maximum sample size N_{max} become larger than those from the optimal design.

3.4 Effects of Cohort Sizes, and Interim Monitoring Distributions on Expected Sample Sizes

The effects of cohort sizes on the expected sample sizes of the SA design are studied with three scenarios and the results are presented in Table 4. The cohort size varies from 1 to 10 and the value of maximum sample size N_{max} is fixed for each scenario. Interim monitoring starts after the 10th patient for all designs. When the cohort size increases from 1 (continuous monitoring) to 3, the corresponding expected sample sizes increase in the range from 0.1 to 0.5 with different scenarios. When the cohort size is 5, the corresponding expected sample sizes increase in the range of 1.0 to 1.3 compared to the results with continuous monitoring. When the cohort size is 10, the expected sample sizes increase in the range of 2.4 to 2.9. If we consider that one patient increase in the expected sample size is tolerable in a trial design, the cohort sizes of 3 to 5 are good choices for the design.

The interim monitoring vector is fixed in the SA design to avoid the intensive computational complexity. However, it is of interest to know the effects of the distributions of interim monitoring on the expected sample sizes with a fixed number of interim monitoring stages. Five different SA designs with different distributions of interim monitoring are studied for this purpose and the results are presented in Table 5. Distributions of interim monitoring include the following scenarios: (1) Uniformly distributed, (2) Early stage monitoring preferred, (3) Extremely early stage monitoring preferred, (4) Late stage monitoring preferred, and (5) Extremely late stage monitoring preferred. The maximum sample size N_{max} is 45 and the number of interim monitoring stages is 8 for all designs. The design with the uniformly distributed interim monitoring (1) has the smallest expected sample size among all interim monitoring distributions. The more extreme the interim monitoring distribution is, the larger the expected sample size is. The expected sample sizes of the late stage monitoring preferred are larger than those from the early stage monitoring preferred, as expected.

3.5 Convergence of the Simulated Annealing Method

One concern frequently raised regarding the SA method is the convergence of the results. Because the SA method is essentially based on a stochastic process, it is possible to achieve varying results from different runs. It is difficult to completely solve this problem in general. However, for a practical application, we can improve the results by enlarging the number of iteration steps and running multiple simulations until the error is smaller than the precision requirement. We illustrate a trace plot of the expected sample size in a typical SA

calculation in Figure 3. Although after the first 5×10^6 iteration steps, the change in the smallest expected sample sizes is small (less than 0.5), we still run the entire simulation with a very large number (4×10^8) of iteration steps to reduce the probability of being stuck at the local minima. The fluctuation amplitude is controlled at around 2.0. We perform 40 runs with different annealing paths for each scenario and check the variance of the solutions to ensure the convergence. The standard deviations of the smallest expected sample sizes for scenarios in Table 1, based on the results of 40 simulations for each scenario, are summarized in Table 6. The largest standard deviation of all scenarios from 40 runs is 0.027 ($p_0 = 0.4$, Table 1 Section B), which is a small value from a practical point of view. We illustrate the distribution of the smallest expected sample sizes of 40 runs in Figure 4, where the difference between the maximum and minimum values is around 0.1, which is quite small for any given trial.

4 Discussion

The simulated annealing method is a useful tool for optimizing the expected sample size of single-arm phase II trials. The SA method successfully reproduces the early stopping boundaries of the commonly used Simon's two-stage designs. For multi-stage designs, we compare the results from the SA method, the posterior probability method, the predictive probability method, and the decision-theoretic method. The SA method has the smallest expected sample sizes in almost all scenarios. The expected sample sizes obtained by the SA method are generally 10–20% smaller than those obtained by the posterior probability method and the predictive probability method and provide a slight reduction from those obtained by the DT method.

The SA method can be applied in conjunction with other methods to further optimize the study design. For instance, the user can apply a stopping boundary obtained by another method as the initial early stopping boundary input to the SA method to obtain better results by further optimization.

As an optimization method, the SA method can be used for high-dimensional optimization problems that arise in clinical trial designs. It is difficult to solve this kind of problem using the traditional exhaustive search method, but the SA method is suitable for this purpose. Furthermore, SA methods can be extended to multi-stage two-arm trials or other more complex settings and provide a fast searching algorithm to find near optimal solutions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank LeeAnn Chastain for editorial assistance. The study was supported, in part, by NCI grants CA16672 and CA97007.

References

1. Simon R. Optimal 2-stage designs for phase-II clinical trials. *Controlled Clinical Trials*. 1989; 10:1–10. [PubMed: 2702835]
2. Ensign LG, Gehan EA, Kamen DS, Thall PF. An optimal three-stage design for phase II clinical trials. *Statistics in Medicine*. 1994; 13:1727–1736. [PubMed: 7997706]
3. Chen TT. Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine*. 1997; 16:2701–2711. [PubMed: 9421870]

4. Chen K, Shan M. Optimal and minimax three-stage designs for phase II oncology clinical trials. *Contemporary Clinical Trial*. 2008; 29:32–41.
5. Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. *Clinical Trials*. 2008; 5:93–106. [PubMed: 18375647]
6. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by Simulated Annealing. *Science*. 1983; 220:671–680. [PubMed: 17813860]
7. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence*. 1984; 6:721–741.
8. Sheldrick GM. Phase annealing in SHELX-90: direct methods for larger structures. *Acta Cryst*. 1990; A46:467–473.
9. Müller P, Sans B, De Iorio M. Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*. 2004; 99(467):788–798.
10. Zhong B. Single-arm phase IIA clinical trials with go/no-go decisions. *CCT*. 2012; 33:1272–1279.
11. Schultz JR, Nichol FR, Elfving GL, Weed SD. Multiple-stage procedures for drug screening. *Biometrics*. 1973; 29:293–300. [PubMed: 4709516]
12. Kirkpatrick S. Optimization by Simulated Annealing: Quantitative studies. *J. Stat. Phys*. 1984; 34:975–986.
13. Lewis DRJ, Berry DA. Group sequential clinical trials: a classical evaluation of Bayesian decision-theoretic designs. *Journal of the American Statistical Association*. 1994; 89:1528–1534.

Figure 1.

Early stopping boundaries using different methods. The stopping boundaries correspond to the scenarios of (a) $p_0 = 0.2, p_1 = 0.4$ and (b) $p_0 = 0.3, p_1 = 0.5$ of minimax designs in Table 1. SA, simulated annealing method; Decision, decision-theoretic method; PP, predictive probability method; Posterior, posterior probability method

Figure 2.

Expected sample sizes with different N_{max} values. The figure includes the expected sample sizes of the simulated annealing designs (SA) under the optimal criteria with $\alpha = 0.10$, $\beta = 0.10$ and Simon's two stage designs (Minimax and Optimal) for three scenarios: (1) $p_0 = 0.1$, $p_1 = 0.3$ (2) $p_0 = 0.2$, $p_1 = 0.4$ (3) $p_0 = 0.3$, $p_1 = 0.5$.

Figure 3.

The early stage of the trace plot of the expected sample size in one simulated annealing run with the scenarios of (a) $p_0 = 0.2$, $p_1 = 0.4$ and (b) $p_0 = 0.3$, $p_1 = 0.5$ optimal designs in Table 1.

Figure 4. Distribution of the average sample sizes obtained by 40 simulated annealing runs with the scenario $p_0 = 0.4$ of the optimal designs in Table 1.

Expected sample sizes with error constraints ($\alpha = 0.1, \beta = 0.1$) using different methods. The values of N_{max} correspond to the maximum number of patients in Simon's **minimax** and **optimal** two-stage designs. In the designs, after the first 10 patients receive treatment, continuous monitoring is performed for all methods except Simon's two-stage design. (SA, simulated annealing method; DT, decision-theoretic method; PP, predictive probability method; PT, posterior probability method; Two-Stage, Simon's two-stage design.)

Table 1

Methods						
p_0	p_1	n/N_{max}	SA	DT	PP	PT
Section A: N_{max} from Simon's minimax design						
0.1	0.3	10/25	18.70	18.70	19.99	20.61
						20.37
0.2	0.4	10/36	24.48	24.52	27.67	28.83
						28.26
0.3	0.5	10/39	31.63	**	**	38.34
						34.99
		10/42 *	25.73	25.85	32.47	30.50
0.4	0.6	10/41	28.75	28.91	31.07	34.42
						33.84
0.5	0.7	10/39	25.78 ***	25.75	29.15	31.04
						31.00
0.6	0.8	10/35	19.71	19.82	**	**
						28.47
		10/36 *	19.83	21.36	25.51	24.59
0.7	0.9	10/25	16.42	16.42	16.42	20.02
						20.05
Section B: N_{max} from Simon's optimal design						
0.1	0.3	10/35	17.35	18.71	26.01	21.16
						19.84
0.2	0.4	10/37	22.21	22.24	25.13	26.28
						26.02
0.3	0.5	10/46	23.94	24.04	25.11	36.34
						29.89
0.4	0.6	10/46	24.68	24.76	27.33	28.87
						30.22
0.5	0.7	10/45	22.88	23.00	24.31	39.36
						28.96
0.6	0.8	10/38	19.65	19.93	21.65	22.22
						25.38
0.7	0.9	10/28	14.78	15.00	15.73	**
						17.79
		10/29 *	14.88	15.29	19.55	17.56

*** This value is 25.75 with 1.35×10^{10} MC iteration steps.
** Does not meet the criteria

* Solutions with the expanded value of N_{max}

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Type I and type II error rates using different methods. The values of N_{max} correspond to the maximum number of patients in Simon's **minimax** and **optimal** two-stage designs. In the designs, after the first 10 patients receive treatment, continuous monitoring is performed for all methods except Simon's two-stage design. (SA, simulated annealing method; DT, decision-theoretic method; PP, predictive probability method; PT, posterior probability method; Two-Stage, Simon's two-stage design.)

(Type I and type II rates in %)						
p_0	p_1	n/N_{max}	SA	DT	PP	PT
Section A: N_{max} from Simon's minimax design						
0.1	0.3	10/25	(9.44 9.99)	(9.44 9.99)	(9.61 9.54)	(9.51 9.87)
0.2	0.4	10/36	(8.49 10.0)	(8.54 9.97)	(8.78 9.38)	(8.61 9.76)
0.3	0.5	10/39	(9.43 10.0)	**	**	(9.44 9.99)
		10/42*	(8.90 9.98)	(8.95 9.98)	(9.61 8.34)	(9.22 9.61)
0.4	0.6	10/41	(9.49 10.0)	(9.50 9.99)	(9.61 9.77)	(9.54 9.96)
0.5	0.7	10/39	(9.72 10.0)	(9.72 10.0)	(9.95 9.49)	(9.79 9.98)
0.6	0.8	10/35	(9.99 9.96)	(9.94 9.90)	**	**
		10/36*	(9.99 9.96)	(8.58 9.97)	(9.02 8.93)	(8.72 9.81)
0.7	0.9	10/25	(9.05 9.80)	(9.05 9.80)	(9.05 9.80)	(9.05 9.80)
Section B: N_{max} from Simon's optimal design						
0.1	0.3	10/35	(8.92 9.98)	(8.95 8.16)	(5.40 7.02)	(4.93 9.56)
0.2	0.4	10/37	(9.29 9.99)	(9.21 9.98)	(9.95 8.36)	(9.77 9.26)
0.3	0.5	10/46	(9.41 9.99)	(9.30 9.96)	(9.89 9.19)	(6.64 9.95)
0.4	0.6	10/46	(9.97 9.97)	(9.17 9.99)	(9.98 8.42)	(9.85 9.41)
0.5	0.7	10/45	(9.45 9.99)	(9.28 9.92)	(9.94 8.72)	(6.73 9.96)
0.6	0.8	10/38	(10.0 9.99)	(8.99 9.87)	(9.93 8.10)	(9.70 9.25)
0.7	0.9	10/28	(10.0 9.98)	(8.77 9.65)	(9.97 7.67)	**
		10/29*	(10.0 9.99)	(7.66 9.75)	(9.32 6.37)	(8.76 7.59)

** Does not meet the criteria

* Solutions with the expanded value of N_{max}

Table 3

Optimal simulated annealing (SA) design for different scenarios with α 0.10, β 0.10.

p_0	p_1	SA				Two-Stage			
		N_{max}	$EN(p_0)$	N_{max}	$EN(p_0)$	N_{max}	$EN(p_0)$	N_{max}	$EN(p_0)$
0.1	0.3	26	17.21	35	19.84				
0.2	0.4	38	21.42	37	26.02				
0.3	0.5	47	23.71	46	29.89				
0.4	0.6	47	24.21	46	30.22				
0.5	0.7	46	22.73	45	28.96				
0.6	0.8	40	19.39	38	25.38				
0.7	0.9	27	14.68	28	17.79				

Table 4

Expected sample sizes of different cohort sizes with $\alpha = 0.10$, $\beta = 0.10$.

Cohort Sizes	Scenarios		
	$p_0 = 0.1, p_1 = 0.3$	$p_0 = 0.2, p_1 = 0.4$	$p_0 = 0.3, p_1 = 0.5$
	$N_{max} = 35$	$N_{max} = 37$	$N_{max} = 46$
1	22.21	17.35	23.94
3	22.58	17.46	24.42
5	23.16	18.37	25.23
10	24.63	20.22	26.36

Table 5

Expected sample sizes of different interim monitoring distributions with $p_0 = 0.3$, $p_1 = 0.5$, $\alpha = 0.1$, $\beta = 0.1$

Scenario	Interim Analysis Points	Expected Sample Sizes
1	(10 15 20 25 30 35 40 45)	26.03
2	(10 13 16 19 22 25 35 45)	26.16
3	(10 12 14 16 18 20 25 45)	27.23
4	(10 20 30 33 36 39 42 45)	27.18
5	(10 30 35 37 39 41 43 45)	29.73

Table 6

Standard deviation of the smallest sample sizes with the different scenarios shown in Table 1. We performed 40 runs for each scenario.

p_0	p_1	n/N_{max}	S.D.
Section A: N_{max} from Simon's minimax design			
0.1	0.3	10/25	0.000
0.2	0.4	10/36	0.000
0.3	0.5	10/39	0.002
		10/42	0.014
0.4	0.6	10/41	0.013
0.5	0.7	10/39	0.017
0.6	0.8	10/35	0.013
		10/36	0.017
0.7	0.9	10/25	0.009
Section B: N_{max} from Simon's optimal design			
0.1	0.3	10/35	0.000
0.2	0.4	10/37	0.002
0.3	0.5	10/46	0.019
0.4	0.6	10/46	0.027
0.5	0.7	10/45	0.024
0.6	0.8	10/38	0.023
0.7	0.9	10/28	0.008