Optimal Two-stage Designs for Single-arm Phase II Oncology Trials with Two Binary Endpoints

C. U. Kunz; M. Kieser
Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany

1. Introduction

In phase II clinical trials in oncology, the potential efficacy of a new treatment regimen is assessed in terms of antitumor activity. Although there is a debate in recent literature on whether these studies should be performed as randomized trials (see, e.g., [1, 2]), the standard approach still consists in a single-arm two-stage design where a binary endpoint of interest (typically tumor response) is compared to a specified target value. The most popular of such designs are still those proposed by Simon in 1989 [3] minimizing the expected sample size (“optimal design”) or the maximum sample size (“minimax design”), respectively, under the null hypothesis of insufficient activity and subject to the constraints imposed by type I error rate and desired power. Since then, this design has been extended in various ways, for example by applying alternative optimality criteria (see, e.g., [4–6]), more than two stages [7], or for evaluating whether a treatment is active in a population as a whole or in a targeted subgroup defined by biomarker status [8]. Motivated by an example of a planned phase II trial within an oncology drug development program, Lin et al. [9] recently proposed a two-stage design where the hypothesis test considers two endpoints, namely “objective response” and “clinical response” (including both objective response and disease stabilization). This approach faces the fact that a new drug would still be considered as promising if it showed a lower tumor response rate than the target level but would lead to a prolonged stabilization of the disease. A major drawback of the work of Lin et al. [9] is the use of Monte Carlo simulations to find designs with pre-specified significance level and power. However, in the example they used—which is quite typical for phase II cancer trials—there exist about 30 million candidate designs that may potentially fulfill the constraints subject to the type I and type II error rate when restricting the search to a maximum of 70 enrolled patients. A systematic search and comparison of designs fulfilling defined optimality criteria using simulations is therefore impractical.

After introducing the notation and the test problem, we give an analytical solution for the calculation of the type I and type II error rates for this design in the next section. By enumeration using exact binomial and trinomial probabilities optimal designs can be determined. Furthermore, the problem of inference about the two single endpoints following rejection of the global null hypothesis is addressed by deriving a multiple test procedure that controls the experimentwise type I error rate in the strong sense.
2. Methods

2.1 Test Problem and Design

We adopt the notation used in Lin et al. [9] and denote by $p_i$ the true probability of success for the binary endpoint $i$, $i = 1, 2$, where success in endpoint 2 includes success in endpoint 1 (i.e., $p_2 > p_1$). Examples for such a scenario are “objective response” determined by RECIST guidelines [10] as endpoint 1 and “clinical response” consisting of complete response or partial response (i.e., more than 50% reduction in size of designated indicator lesions) as endpoint 2. The individual test problem for endpoint $i$ is denoted as

$H_0^i: p_i \leq p_{i0}$ versus $H_1^i: p_i \geq p_{i+}, i = 1, 2,$ (1)

where $p_{i0} \leq p_{20}$ and $p_{i+} \leq p_{2+}$. Here $p_{0+}, i = 1, 2,$ is an uninteresting level indicating too low activity and $p_{+} > p_{0+}, i = 1, 2,$ is a desirably high target level. Within drug development any indication for anti-tumor activity for at least one endpoint justifies further research. Consequently, the following test problem is assessed:

$H_0: H_0^1 \cap H_0^2$ versus $H_1: H_1^1 \cup H_1^2. $ (2)

The design proposed by Lin et al. [9] uses the following decision rules: Stage 1: Enroll $n_1$ patients. If the observed number of successes with respect to endpoint 1 is greater than $r_1$, go to stage 1; otherwise stop and accept $H_0$. Stage 2: Enroll further $n - n_1$ patients. Reject $H_0$ if for the total of $n = n_1 + n_2$ patients more than $r$ successes with respect to endpoint 1 or more than $s$ successes with respect to endpoint 2 are observed; otherwise accept $H_0$.

The type I error rate and the type II error rate at the point $p_1 = p_{1+}$ and $p_2 = p_{2+}$ as well as the expected sample size (ESS) and the probability of early termination (PET) can be derived as shown in the Appendix.

For each value of total sample size $n$ the algorithm searches over $r$ in the range of $[0, n-1]$ and over $s$ in the range of $[r, n-1]$. Since the cumulative binomial distribution at the point $(r, s, n)$ is a lower boundary of the type II error, we check if

$$\sum_{x=0}^{s} \sum_{x_1=0}^{x} m(x_1, x_2 - x_1; n, p_{i1}, p_{2a} - p_{i0}) < \beta$$

is satisfied for the parameters $(r, s, n)$, where $m(x_1, x_2 - x_1; n, p_{i1}, p_{2a} - p_{i0})$ denotes the binomial distribution (Appendix). If the condition is not fulfilled, the rest of the range of $s$ is skipped and the search continues with the next $r$. If the condition is fulfilled, the algorithm searches over $r_1$ in the range of $[0, r]$ and backward over $n_1$ in the range of $[r_1 + 1, n - (r - r_1)]$. As can be seen from the formula given in the Appendix, the type II error is calculated using a sum consisting of a cumulative binomial distribution and the product of two trinomial distributions. The parameters $(r_1, n_1, r, s, n)$ can only then be a solution, if the sum is smaller than $\beta$. Since each term of the sum is positive, both terms themselves must be smaller than $\beta$.

Therefore, we check if

$$\sum_{x=0}^{s} m(x, n_1, n_2)$$

is smaller than $\beta$, with $b(x, n_1, n_2)$ denoting the binomial distribution (Appendix). If this condition is not fulfilled, the rest of the range of $n_1$ is skipped and the algorithm continues with the next $r_1$. Otherwise, we further check if $P_{H_1^i}(R_i > r_1 \text{ and } R > r)$ is less than $\alpha$. As shown in the following section, $P_{H_1^i}(R_i > r_1 \text{ and } R > r)$ is less or equal to $P_{H_0^i}(R_i > r_1 \text{ and } (R > r \text{ or } S > s))$ and, therefore, any given parameter set can only then be a solution if it meets the above mentioned condition.

There are usually a number of solutions $(n_1, n_2, r_1, r, s)$ that fulfill the requirements with respect to $\alpha$ and $\beta$. According to the criteria proposed by Simon [3], we define optimal designs as those with minimum expected sample size under $H_0$; if this solution is not unique, the one with lowest $\beta$ (i.e., with highest power $1 - \beta$) is chosen. Minimax designs are defined as those with minimum total sample size $n$ under $H_0$; if this solution is not unique, the one with minimum expected sample size is chosen.

By a systematic search, designs satisfying a desired optimality criterion can be identified.

Optimal and minimax designs may show highly divergent characteristics with respect to expected and maximum sample size. For this reason, compromise designs have been proposed minimizing a weighted sum of expected and maximum sample size under the null hypothesis [6, 11]. As these designs are admissible according to Bayesian decision theory with a loss function justified by ethical reasons [6], they are denoted as admissible designs in the literature. Such designs can also be identified with the methods described above, and we illustrate their application with an example in Section 3.1.

2.2 Multiple Test Procedure

An important issue in practical applications that was not addressed by Lin et al. [9] is whether after rejection of the global null hypothesis $H_0$ inference about the single endpoints can be made under control of the type I error rate. The closure principle [12] assures control of the experimentwise type I error rate in the strong sense [13] by $\alpha$, if level-$\alpha$ tests are applied for the assessment of $H_0$ and $H_0^i, i = 1, 2$, and if $H_0^i$ is rejected if and only if both $H_0$ and $H_0^i$ are rejected by the corresponding tests at level $\alpha$. Within the two-stage design considered here, every level-$\alpha$ for $H_0$ defines a test for $H_0^i$ with type I error rate of at most $\alpha$ when the same decision boundaries $r_1$ and $r$ for the number of successes for endpoint 1 are used:

$$\alpha \geq P_{H_0^i}(R_i > r_1 \text{ and } (R > r \text{ or } S > s)) \rightarrow P_{H_0^i}(R_i > r_1 \text{ and } R > r) = P_{H_0^i}(R_i > r_1) \text{ and } (R > r)$$

Here $R_i, R$ and $S$ denote the number of successes for endpoint 1 after the first and second stage and for endpoint 2 after the second stage, respectively. As the decision rule applied after the first stage does not make use of the results for endpoint 2, the binomial test of $H_0^2$ is a level-$\alpha$ test. Denoting the related decision boundary for the binomial test by $s$, one can search for boundaries $r_1$ and $r$ that provide together with $s$ a
level-α test for $H_0$. Then the following closed testing procedure controls the experimentwise type I error rate in the strong sense: If $H_0$ cannot be rejected at level α, $H_0^i$, $i = 1, 2$, are also not rejected. If $H_0$ can be rejected, the decision rule for $H_0$ is applied elementwise to the individual endpoint hypotheses $H_0^i$, $i = 1, 2$ to decide about rejection or acceptance. As the same decision boundaries are used for the tests of the global and the individual null hypotheses, this multiple test procedure assures that at least one of the null hypotheses for the single endpoints can be rejected after rejection of the global null hypothesis, i.e., consonance of the test procedure [14]. By construction of the closure test, it also guarantees that if an individual null hypothesis is rejected the global null hypothesis (that implies $H_0^i$, $i = 1, 2$) is rejected too, i.e., coherence of the test procedure [14]. Hence, the proposed multiple test procedure is both consonant and coherent which is a desirable property of multiple test procedures [15].

In the following section we show how this closed testing procedure is constructed in an example clinical trial and how it compares to optimal, minimax and admissible designs.

3. Results

3.1 Real Data Example of Lin et al.

Lin et al. [9] reported on a planned two-stage phase II trial where a new oncology medicine against a specific cancer type was evaluated with respect to its anti-tumor activity. Endpoint 1 was “objective response” determined by RECIST guidelines [10]. Additionally, “clinical response” was assessed as endpoint 2 which was defined by the occurrence of objective responder a prolonged disease stabilization of ≥24 weeks (median overall survival without treatment was 5.6 months for patients with this tumor). The success probabilities considered as too low or high enough, respectively, to justify further investigation of the drug were specified as $p_{10} = 0.05$, $p_{20} = 0.15$ and $p_{1a} = 0.15$, $p_{2a} = 0.30$, respectively. The type I error rate was $\alpha = 0.05$ and the desired power $1 - \beta = 0.85$. By simulations, Lin et al. determined the designs with $n_1 = 38$, $n = 63$, $r_1 = 1$, $r = 6$ and $s = 16$, or $r = 7$ and $s = 15$, respectively, to satisfy the requirements on significance level and power. The estimated type I error rate given by Lin et al. was 0.0496 or 0.0488, and the estimated power amounts to 0.9226 or 0.8994, respectively (5000 replications for each design). We obtained exact values for significance level and power of 0.0424 or 0.0284 and 0.8979 or 0.8812, respectively. The values for ESS (52.3) and PET under the null hypothesis (0.43) are the same for these designs. The optimal design subject to the restrictions on $\alpha$ and $\beta$ (which is uniquely defined in this case) is given by $n_1 = 25$, $n = 58$, $r_1 = 6$ and $s = 13$ with an ESS of 36.8 and a PET of 0.64. In contrast, the minimax design (which is also uniquely defined) has a maximum sample size of only $n = 48$ ($n_1 = 45$) but an expected sample size of 45.6 (decision rule $r_1 = 3$, $r = 5$, $s = 11$). Thus, the minimax design has a maximum sample size that is ten patients smaller than the optimal design but in return the expected sample size is about nine patients higher. In such a situation an admissible design may be a valuable compromise. If expected and maximum sample size are weighted equally, two admissible designs are obtained which are defined by $n_1 = 32$, $n = 49$, $r_1 = 1$ and $r = 6$, $s = 11$ or $r = 5$, $s = 12$, respectively. Figure 1 shows the various designs in the $n$ – ESS plane. It can be seen that increasing the maximum sample size of the minimax design by just one patient results in an admissible design with considerable smaller expected sample size (40.2) thus bringing the extremes of optimal and minimax closer together. As both admissible designs use the same rule for early stopping, the PET is the same for both (0.52). However, as the type II error rate is slightly smaller for the first design (0.1480 versus 0.1491), one may prefer this one.

If the decision rule for the global null hypothesis is also used for making inferences about the single endpoints, the rejection boundary for endpoint 2 has to be fixed such that a level-α binomial test for $H_0^2$ results. This holds true for the designs proposed by Lin et al. (actual level of the binomial test with parameters $n$ and $p_{20}$ and critical boundary $s$ is 0.0099 or 0.0215) as well as for the optimal and the minimax design (0.0450 or 0.0478, respectively). For the two admissible designs, only the second one features this characteristics (actual level of the binomial test 0.0550 or 0.0257, respectively). This may be a reason for favoring it despite its slightly lower power for rejecting of the global null hypothesis.

![Fig. 1](https://example.com/fig1.png)

**Fig. 1** Two-stage designs for testing the global null hypothesis $H_0^i$: $p_1 \leq p_{10} \leq p_2 \leq p_{20}$ at type I error rate $\alpha$ and with type II error rate $\beta$ for $p_1 = p_{1a}$ and $p_2 = p_{2a}$ for $(p_{10}, p_{20}, p_{1a}, p_{2a}, \alpha, \beta) = (0.05, 0.15, 0.15, 0.30, 0.05, 0.15)$
Table 1  Optimal and minimax design for testing the global null hypothesis $H_0: \pi_1 \leq \pi_{10}$ and $\pi_2 \leq \pi_{20}$ in the proposed two-stage design at type I error rate $\alpha$ and with type II error rate $\beta$ for $\pi_1 = \pi_{1a}$ and $\pi_2 = \pi_{2a}$. The maximum total sample size is $n$, the sample size of the first stage is $n_1$; $r_1$, $r$ and $s$ denote the number of successes for endpoint 1 after the first stage and the total number of successes for endpoint 1 and endpoint 2 after the second stage, respectively. For each set of values $(\pi_{10}, \pi_{1a}, \pi_{20}, \pi_{2a})$ the designs are given for $(\alpha, \beta) = (0.10, 0.10)$ (first line), $(0.05, 0.20)$ (second line) and $(0.05, 0.10)$ (third line). ESS($\pi_0$) and PET($\pi_0$) denote the expected sample size and the probability of early termination under $H_0$.

<table>
<thead>
<tr>
<th>$\pi_{10}$</th>
<th>$\pi_{1a}$</th>
<th>$\pi_{20}$</th>
<th>$\pi_{2a}$</th>
<th>Design</th>
<th>$s/r_1/n_1$</th>
<th>$s/r_1/s/n$</th>
<th>ESS($\pi_0$)</th>
<th>PET($\pi_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.30</td>
<td>0.20</td>
<td>0.40</td>
<td>Optimal</td>
<td>1/15</td>
<td>4.9/25*</td>
<td>19.5</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Minimax</td>
<td>1/19</td>
<td>4.8/24*</td>
<td>21.9</td>
<td>0.42</td>
</tr>
<tr>
<td>0.20</td>
<td>0.40</td>
<td>0.30</td>
<td>0.50</td>
<td>Optimal</td>
<td>3/16</td>
<td>11.15/39*</td>
<td>25.2</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Minimax</td>
<td>7/29</td>
<td>10.13/34</td>
<td>30.0</td>
<td>0.79</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.40</td>
<td>0.60</td>
<td>Optimal</td>
<td>7/22</td>
<td>16.21/42*</td>
<td>28.6</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Minimax</td>
<td>9/28</td>
<td>15.19/38*</td>
<td>31.2</td>
<td>0.68</td>
</tr>
<tr>
<td>0.40</td>
<td>0.60</td>
<td>0.50</td>
<td>0.70</td>
<td>Optimal</td>
<td>8/20</td>
<td>21.24/41</td>
<td>28.5</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Minimax</td>
<td>8/21</td>
<td>19.23/38*</td>
<td>29.1</td>
<td>0.52</td>
</tr>
<tr>
<td>0.50</td>
<td>0.70</td>
<td>0.60</td>
<td>0.80</td>
<td>Optimal</td>
<td>7/15</td>
<td>22.25/36*</td>
<td>25.5</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Minimax</td>
<td>7/17</td>
<td>21.23/33*</td>
<td>28.0</td>
<td>0.31</td>
</tr>
<tr>
<td>0.60</td>
<td>0.80</td>
<td>0.70</td>
<td>0.90</td>
<td>Optimal</td>
<td>6/11</td>
<td>22.22/28</td>
<td>20.1</td>
<td>0.47</td>
</tr>
</tbody>
</table>

For personal or educational use only. No other uses without permission. All rights reserved.
3.2 Tabulation of Various Designs

Table 1 provides the optimal and minimax designs for $(\alpha, \beta) = (0.10, 0.10), (0.05, 0.20)$ and $(0.05, 0.10)$, $p_{10} - p_{1a} = p_{20} - p_{2a} = 0.20$, $p_{20} - p_{10} = 0.10$, and $p_{10} = 0.10, 0.20, \ldots, 0.60$. For those situations for which the optimal design is not uniquely defined, the maximum sample size is the same for all solutions and hence the design with maximum power for rejecting $H_0$ was chosen. When there exists more than one minimax design with the same $n$, the one with minimal ESS under the null hypothesis was chosen. As can be seen from the table, for most of these designs the decision rule for $H_0$ can also be used to make inferences on the single endpoints within the proposed closed testing procedure.

C++ was used to find solutions for pre-specified values of $p_{10}, p_{20}, p_{1a}, p_{2a}, \alpha$, and $\beta$ by calculating type I and type II error rates for any given parameter set $(r_1, n_1, r, s, n)$ according to the algorithm described above. All solutions were saved to a txt file. The code is available from the authors upon request. Stata MP 10.0 was then used to sort the solutions within the txt file in order to find the minimax, optimal, and admissible design and to plot the results.

All calculations were done on a HP xw6600 workstation with eight processors with 2.5 GHz each and 3.25 GB RAM. The average runtime to find the minimax design was about seven seconds with a minimum of less than one second and a maximum of about 47 seconds. The average runtime to find the optimal design was about 286 seconds with a minimum of less than one second and a maximum of about 1216 seconds.

4. Discussion

We have extended Simon’s two-stage design [3] to two endpoints for which success in one of them implies success in the other. This approach allows a more comprehensive assessment of the overall picture of anti-tumor efficacy of a new treatment than restriction to a single outcome. For this design, optimal and minimax solutions were determined. As for two-stage designs with one endpoint, the characteristics of optimal and minimax designs with respect to expected and maximum sample size may be quite different. Hence, admissible designs as proposed by Jung et al. [6, 11] may be a valuable compromise as demonstrated by a clinical trial example. A multiple test procedure is presented that controls the experimentwise type I error rate in the strong sense and that allows to make inference not only for the global null hypothesis but additionally for the hypotheses concerning the single endpoints. Instead of using the multiple test procedure presented here, the two null hypotheses could be ordered and tested hierarchically. However, $H_{ij}^c$ can then only be tested if $H_{ij}^c$ is rejected. Therefore, this approach does not match the situation considered here, where it is essential to detect an effect for any of the two equally important outcomes.

For the proposed design, early stopping with acceptance of the null hypothesis is based on the number of successes observed for endpoint 1. This is motivated by the study example underlying this research where the results for one endpoint are available much earlier than for the other. Therefore, considering only the information on this endpoint after the first stage meets the requirement of a timely decision about termination or continuation of the study. However, the decision rule for the interim analysis can be extended to take into account both endpoints as done in other papers dealing with two outcomes in phase II oncology studies [16, 17]. In these papers, Lin and Chen [16] and Panagés et al. [17] considered complete and partial response as endpoints which are exclusive categories. Therefore, their methods do not match the problem we address. Lu et al. [18] proposed a design where the decision rules for both stages are based on the two endpoints “total response” and “complete response”. Therefore, the trial would proceed to the second stage when either enough responses regarding “total response” or “complete response” are observed. This approach is useful only when the information about the endpoints is available in time and both endpoints are similarly important. Furthermore, they used the marginal power functions for confirming one of the two alternative hypotheses instead of the joint alternative hypothesis.

We have given an analytical solution for the design proposed by Lin et al., which allows making inference about the two single endpoints. In addition to reporting the inferential decision and p-value it is helpful in phase II trials to report the success rate and associated confidence interval. Further research will investigate point estimates and confidence intervals for individual endpoints taking into account the particular sequential design and test problem.

Acknowledgments

The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG) for supporting this research by Grant Ki 708/1-1.

References

Appendix

Derivation of Type I and Type II Error Rate for Proposed Design

Let \( b(x; n, p) \) denote the binomial probability that there are \( x \) successes in \( n \) patients when the true success rate is \( p \), that is
\[
b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}.
\]
Correspondingly, \( m(x_1, x_2 - x_1; n, p_1, p_2 - p_1) \) denotes the respective trinomial probability, that is
\[
m(x_1, x_2 - x_1; n, p_1, p_2 - p_1) = \binom{n}{x_1} P_{p_1}^{x_1} (P_{p_2 - p_1})^{x_2 - x_1} (1-P_{p_2 - p_1})^{n-x_1},
\]
where \( p_2 > p_1 \) and \( x_2 \geq x_1 \). The probability of rejecting \( H_0 \) within the two-stage design described in Section 2.1 when the true success probability for endpoint \( i \) is \( p_i \), \( i = 1, 2, p_2 > p_1 \), is given by the equation shown in Figure 2. Within the brackets the first part of the sum refers to the first stage of the trial. Since the decision at the end of the first stage is based on only one endpoint, a binomial distribution is used. The second part of the sum refers to the second stage of the trial where the decision is based on two endpoints. This part of the sum, therefore, consists of trinomial distributions.

It follows that the type I error rate and the type II error rate at the point \( p_1 = p_{10} \) and \( p_2 = p_{20} \), respectively, for test problem (2) are
\[
\alpha = \text{rej}(n_1, r_1, r, s, p_{10}, p_{20}) \quad \text{and} \quad \beta = 1 - \text{rej}(n_1, r_1, r, s, p_{10}, p_{20}).
\]
The probability of early termination after the first stage (PET) is
\[
(\text{PET}) = \text{PET}(n_1, r_1, p_{10}) = \sum_{x_1 = 0}^{r_1} b(x_1; n_{10}, p_{10})
\]
and the expected sample size is
\[
\text{ESS}(n_1, n_1, r_1, p_{10}) = n_1 + (\text{PET}(n_1, r_1, p_{10}) \cdot (n - n_1)).
\]

Fig. 2 Formula for the probability of rejecting \( H_0 \) within the proposed two-stage design