Two-stage phase II oncology designs using short-term endpoints for early stopping

Cornelia U. Kunz[1], **James M. S. Wason**[2] **and Meinhard Kieser**[3]

Abstract

Phase II oncology trials are conducted to evaluate whether the tumour activity of a new treatment is promising enough to warrant further investigation. The most commonly used approach in this context is a two-stage single-arm design with binary endpoint. As for all designs with interim analysis, its efficiency strongly depends on the relation between recruitment rate and follow-up time required to measure the patients' outcome. Usually, recruitment is postponed after the sample size of the first stage is achieved up until the outcomes of all patients are available. This may lead to a considerable increase of the trial length and with it to a delay in the drug development process. We propose a design where an intermediate endpoint is used in the interim analysis to decide whether or not the study is continued with a second stage. Optimal and minimax versions of this design are derived. The characteristics of the proposed design in terms of type I error rate, power, maximum and expected sample size as well as trial duration are investigated. Guidance is given how to select the most appropriate design. Application is illustrated by a phase II oncology trial in patients with advanced angiosarcoma that motivated this research.

Keywords

two-stage designs, oncology, short-term endpoint, early stopping

[1]Warwick Medical School, University of Warwick, UK
[2]Hub for Trials Methodology Research, MRC Biostatistics Unit, Cambridge, U.K.
[3]Institute of Medical Biometry and Informatics, University of Heidelberg, Germany

Corresponding author:

Meinhard Kieser, Institut für Medizinische Biometrie und Informatik, University of Heidelberg, Im Neuenheimer Feld 305, D-69120 Heidelberg, e-mail: meinhard.kieser@imbi.uni-heidelberg.de, phone: +49 6221 564140

1 Introduction

Phase II oncology trials are carried out to determine if a new treatment is sufficiently effective to justify being tested in a larger trial. Typically endpoints used in phase II trials in solid tumours are based on the RECIST criteria [1], which categorises patients based on: (1) the change in their tumour size and (2) whether they experience some other indication that the treatment is ineffective, such as new lesions. Four categories are used, in order from best to worse: complete response (CR), partial response (PR), stable disease (SD) and progressive disease (PD). In trials of cytotoxic drugs, in which the mechanism of action is the destruction of tumour cells, the probability of CR or PR is typically used as a phase II endpoint. This is known as the response rate (RR). For cytostatic drugs, which aim to control the growth of the tumour, the RR is less suitable [2, 3]. Instead, the disease control rate (DCR), which is the probability of CR, PR or SD, is more appropriate. Both RR and DCR are typically analysed as binary outcomes.

Single-arm phase II designs are frequently applied in oncology. A recent review [4] found that that 60% of phase II trials in six different solid tumour types were single-arm. On the one hand, these trials are carried out as a precursor to larger, randomised, phase II or III trials. Alternatively, single-arm phase II trials may also stand on their own because drug approval may by based solely on data from these trials which highlights their relevance for the drug development process. According to the registration information for US Food and Drug Administration (FDA) oncology drug approvals for solid tumours 1981 through 2008 inclusive, 12% of cytotoxic drugs and 15% of targeted drugs were approved based on single-arm phase II data only [4], [5]. Furthermore, 52.7% of pivotal efficacy trials providing the basis for approval of novel therapeutic anti-cancer agents by the US FDA between 2005 and 2012 were performed without control [6]. Recently, the FDA's Director of Hematology and Oncology Products opined that due to the dramatic response rates observed for some targeted agents single-arm phase II trials will gain an even more important role in the approval process in future [7] Main attractions for using these designs are their simplicity and the small sample size, which allows the trial to be run in a single centre. Mostly these trials are conducted in a two-stage design with a futility threshold specified for the interim analysis. If the number of responders is above a critical value in the first stage, the trial continues to the second stage. Simon [8] proposed two designs for this type of trial: optimal and minimax. The optimal design has the lowest expected sample size (ESS) under the null hypothesis, whereas the minimax design has the lowest maximum sample size (MSS).

In designs with one or more interim analyses, such as Simon two-stage designs, the efficiency of the design is strongly impacted by delay between recruitment and assessment of patients [9]. The effect will depend on whether the trial suspends recruitment once all first stage patients are recruited. If the recruitment is suspended, the time taken to run the trial will increase as the delay length increases. If not, then second stage patients will be recruited until the first stage patients are followed up, with the result that the expected sample size of the trial will increase. In larger multi-centre trials, suspending the recruitment is generally not feasible, but it is possible in a smaller single-centre trial. We will henceforth consider trials where recruitment is suspended once the specified number of first-stage patients is recruited. One method to improve the efficiency of the trial in the presence of delay is to use an intermediate endpoint that is observed more quickly than the final endpoint. If there is a strong correlation between the intermediate and final endpoints, then testing the intermediate endpoint at the interim

analysis will provide much of the benefits of testing the definitive endpoint, but in a much quicker time. There are a number of papers that extend the Simon two-stage design to allow inclusion of more than one endpoint. Some examples are simultaneous testing of efficacy and toxicity [10], testing of two primary outcomes [11], testing of primary and secondary outcomes [12], considering complete and partial responders separately [13, 14, 15]. In this paper we propose a new design that evaluates a quickly observed intermediate endpoint at the first stage, and the definitive endpoint in the second stage. We show that this design has the potential to considerably reduce the time taken by the trial, and thereby improve the efficiency of the drug development process. Although we consider single-arm designs, methods we develop are easily applicable to randomised phase II trials as well.

2 Motivating example

Ray-Coquard et al. [16] report the outcomes of a phase II oncology trial of Sorafenib for patients with advanced angiosarcoma. They assumed that Sorafenib would not lead to a tumour response but could provide long-lasting disease stabilisation. Hence, the primary endpoint of the trial was 9-month progression free survival (PFS). Sample sizes were calculated using a Simon's minimax design [8] based on the assumption of a 9-month PFS rate of 0.127 under the null hypothesis and of 0.317 under the alternative, $\alpha = 10\%$, and $\beta = 5\%$. The statistical assumptions were based on a previous trial of Paclitaxel in the same patient population [17]. In the first stage $n_1 = 26$ patients were to be included. If at most $r_1 = 3$ successes in the first stage were observed, the trial would be stopped due to futility. Otherwise, the second stage of the trial would be opened and a further $n_2 = 17$ patients would be enrolled. The null hypothesis would be rejected if more than $r = 8$ successes were observed out of a total of $n = n_1 + n_2 = 44$ patients enrolled in the trial.

As the primary endpoint was 9-month PFS results for the first stage were not available up until 9 months after the last patient for the first stage was enrolled. Therefore, patient entry had to be postponed until the results for the first stage were available. Ray-Coquard et al. report that 26 patients were enrolled within 13 months (from June 2008 to June 2009). Assuming that the first patient was enrolled on day 0 and that on average another patient was enrolled every 15 days, the last patient was enrolled at day 375 and the outcome for the last patient was available 645 days after the first patient was enrolled.

In the following, we propose a design were the decision whether or not to continue the trial after the first stage is based on a short-term endpoint while the null hypothesis at the end of the trial is still tested based on the long-term endpoint. Ray-Coquard et al. mention that a 9-month PFS rate of 0.127 (0.317) correspondes to a 4-month PFS of 0.4 (0.6). The numbers are slightly rounded versions of the results given by Penel et al. [17] which are shown in Table 1. Based on these results we will investigate the effect of using a short-term endpoint for the first stage on the sample sizes, power, and trial length.

3 Methods

3.1 Notation and proposed design

Throughout the article, we use the following notation: p_1, p_2, and p_{12} are unknown parameters denoting the rates for a positive outcome for the long-term (primary), short-term (secondary), and both endpoints, respectively.

Table 1. Results from Penel et al. 2008 [17] (estimated rate ± standard error based on 27 assessable patients).

endpoint	2 months	4 months	6 months	9 months	12 months	18 months
OS	96% ± 3.8%	73% ± 8.5%	56% ± 9.6%	38% ± 9.3%	38% ± 9.3%	21%± 7.8%
PD	26% ± 8.4%	55% ± 9.6%	76% ± 8.2%	—	—	—
CR	0%	5% ± 4.2%	14% ± 6.7%	—	—	—
PR	19% ± 7.6%	14% ± 6.7%	5% ± 4.2%	—	—	—
CR+PR	19% ± 7.6%	19% ± 7.6%	19% ± 7.6%	—	—	—
SD	56% ± 9.6%	27% ± 8.5%	5% ± 4.2%	—	—	—
PFS	74% ± 8.4%	45% ± 9.6%	24% ± 8.2%	12.7% ± 6.4%	—	—

$p_1 - p_{12}$ gives the rate for the event of a positive outcome for the long-term endpoint and a negative outcome for the short-term endpoint, $p_2 - p_{12}$ gives the rate for a positive outcome for the short-term endpoint and a negative outcome for the long-term endpoint, and $1 - p_1 - p_2 + p_{12}$ gives the rate for a negative outcome for both endpoints. The rate for a positive outcome for the long-term endpoint under the null and alternative hypothesis will be denoted by $p_{1,0}$ and $p_{1,1}$, respectively.

We propose a design that uses the short-term endpoint at the end of the first stage to decide whether or not to continue to the second stage while at the end of the trial only the long-term endpoint, which is the outcome of primary interest, is tested (see Figure 1). Thus, the test problem assessed after stage two is given by $H_0 : p_1 \leq p_{1,0}$ versus $H_1 : p_1 \geq p_{1,1}$ with $p_{1,1} > p_{1,0}$. Note that the design is not restricted to using a short-term endpoint for decision making after the first stage but allows for application of arbitrary efficiency outcomes. Although we do not make any assumptions about the intermediate endpoint under the null hypothesis in order to be able to calculate the expected sample sizes (see Section 3.3), we denote by $p_{2,0}$ the assumed response rate for the secondary endpoint in the case that the treatment is inefficient. In order to calculate the power (see Section 3.2), we denote by $p_{2,1}$ the assumed response rate for the short-term endpoint in the case that the treatment is efficient. The same holds true for p_{12}, i.e no assumptions about p_{12} are made under the null hypothesis. However, in order to be able to calculate the power, we have to set a specific value for p_{12}. Section 4.4 shows how the power is influenced by the choice of this value. Furthermore, let n_1 be the sample size for the first stage and n the total sample size. The observed number of positive outcomes for the short-term endpoint at the end of the first stage (trial) will be denoted with x_1 (x) with $x_2 = x - x_1$. The observed number of positive outcomes for the long-term endpoint at the end of the first stage (trial) will be denoted with y_1 (y) with $y_2 = y - y_1$. The trial will proceed to the second stage if $x_1 > s_1$ where s_1 denotes the critical value for the first stage. At the end of the trial, the null hypothesis is rejected if $y > r$ where r denotes the critical value for the trial. The duration of the first stage in months is denoted by l_1 and the duration of the whole trial by l. Furthermore, the correlation between the two endpoints is denoted by Φ and is measured using Yule's correlation coefficient $\Phi = (p_{12} - p_1 p_2)/\sqrt{p_1(1-p_1)p_2(1-p_2)}$ [18].

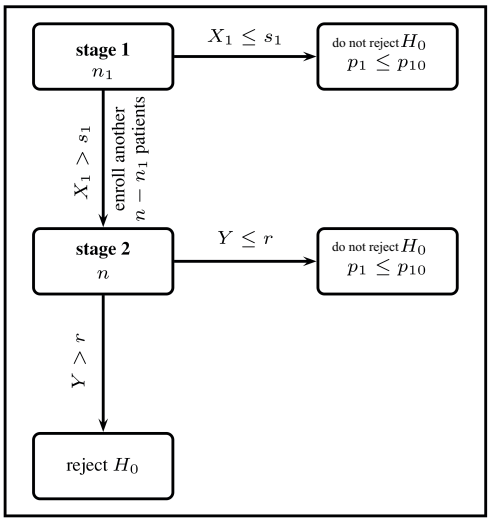

Figure 1. Decision path for the proposed design.

3.2 Type I error rate and power

For the above described design, the type I error rate is given by

$$
\begin{aligned}
1-\Bigg[& \sum_{x_1=0}^{s_1} \binom{n_1}{x_1} p_2^{x_1} (1-p_2)^{n_1-x_1} \\
& + \sum_{x_1=s_1+1}^{n_1} \sum_{z_1=0}^{\min[x_1,r]} \sum_{y_1=0}^{\min\limits_{r-z_1}{}_{n_1-x_1}} \Bigg\{ \binom{n_1}{x_1} \binom{x_1}{z_1} \binom{n_1-x_1}{y_1} (p_2-p_{12})^{x_1-z_1} (p_{1,0}-p_{12})^{y_1} p_{12}^{z_1} (1-p_{1,0}-p_2+p_{12})^{n_1-x_1-y_1} \\
& \qquad \cdot \sum_{x_2=0}^{n-n_1} \sum_{z_2=0}^{\min[x_2,r-z_1]} \sum_{y_2=0}^{\min\limits_{r-y_1-z_1-z_2}{}_{n-n_1-x_2}} \binom{n-n_1}{x_2} \binom{x_2}{z_2} \binom{n-n_1-x_2}{y_2} \\
& \qquad \cdot (p_2-p_{12})^{x_2-z_2} (p_{1,0}-p_{12})^{y_2} p_{12}^{z_2} (1-p_{1,0}-p_2+p_{12})^{n-n_1-x_2-y_2} \Bigg\} \Bigg]. \qquad (1)
\end{aligned}
$$

Within the brackets, the first part of the sum gives the probability to stop the trial after the first stage. Hence, we sum over $x_1 = 0$ to s_1. The second part of the sum gives the probability to proceed to the second stage (hence, $x_1 = s_1 + 1$ to n_1) but not rejecting the null hypothesis. In this case, we have to take into account that some of the patients entered in the first stage of the trial have a positive outcome for the primary endpoint. Furthermore, we have to take into account that some patients show a positive outcome for the short-term and long-term endpoint while others only show a positive outcome for the long-term endpoint. Therefore, we some over $z_1 = 0$ to $\min[x_1, r]$ and over $y_1 = 0$ to $\min[r - z_1, n_1 - x_1]$. The last three summation signs represent the patients entered in the second stage of the trial. If the trial proceeds to the second stage, we are only interested in the long-term endpoint only. Hence, we some over $x_2 = 0$ to $n - n_1$. The last two summation signs ensure that the total number of patients with a positive outcome for the long-term endpoint does not exceed r.

Note that under the null hypothesis only the rate for a positive outcome for the long-term endpoint is fixed by $p_1 = p_{1,0}$ while there is no restriction for p_2 and p_{12}. However, the type I error rate also depends on p_2 and p_{12}. In order to assure control of the type I error rate for all possible values of p_2 and p_{12}, it has to be worked out for which values of (p_2, p_{12}) the expression in formula (1) reaches its maximum. The maximum is attained if the trial

always proceeds to the second stage and the null hypothesis is invariably tested at the end. Choosing $p_2 = 1$ ensures that there is no early stopping and therefore the maximum type I error rate given by Equation (2) is achieved:

$$1 - \sum_{y=0}^{r} \binom{n}{y} p_{1,0}^{y} \left(1 - p_{1,0}\right)^{n-y}. \tag{2}$$

Hence, the maximum type I error rate of the proposed design equals that of a single-stage binomial trial with n patients and the same decision rule applied when testing the null hypothesis. In the following, the type I error rate is always calculated using Equation (2) (i.e. for the worst case scenario of $p_2 = 1$) to ensure control of the type I error rate irrespective of the values of p_2 and p_{12}.

The power of the proposed design can be calculated from Equation (1) by replacing $p_{1,0}$ by $p_{1,1}$, p_2 by $p_{2,1}$ and by employing assumptions about the value of p_{12}. The values of p_2 and p_{12} depend on what is known about the endpoints at the planning stage of the trial. Note that in the case of nested endpoints (for example, 4-month PFS and 9-month PFS), $p_{12} = p_1$ since in order for a patient to have a positive outcome at 9 months, the patient has to have a positive outcome at 4 months.

3.3 Optimisation criteria

Once values for the different response rates have been set, several solutions for s_1, n_1, r and n exist fulfilling the constraints with respect to type I error and power. In order to choose among these solutions, a number of optimisation criteria have been proposed in the literature. The most popular optimisation criteria are the minimax and the optimal criterion. The minimax criterion minimises the MSS; if this leads to more than one solution, the solution is selected which minimises the expected sample size under the null hypothesis (see [8]). The optimal criterion minimises the ESS under the null hypothesis. In general, the expected sample size is given by $ESS = n_1 + (n - n_1)(1 - PET)$, where PET is the probability of early termination after the first stage. Unlike for the "classical" Simon's design where the PET depends on the success rate for the primary endpoint p_1, the PET in our case depends solely on the rate for the short-term secondary endpoint p_2, i.e. the PET is independent of the values of p_1 and p_{12}. As there is no restriction for p_2 under the null hypothesis, the ESS to be minimised is not well-defined. However, if assumptions about p_2 are made that reflect the knowledge on the value of this parameter in the planning phase, the expected sample size can be calculated for the respective scenario. The two most extreme cases are (i) that we know the value for p_2, or (ii) that all values of p_2 are equally plausible (we assign a $U[0, 1]$ distribution to p_2). These situations are reflected in criteria A1 and A2a below. Less extreme assumptions are made for the criteria A2b to A4 where the knowledge about p_2 is modelled by alternative prior distributions. For the criteria A2b, A3a and A3b, c may, for example, be chosen as the point estimator obtained from a previous trial and a and b as the related lower and upper confidence interval limits. The triangular distribution is typically used if there are estimates for the minimum and maximum of p_2 and an estimate for p_2 itself.

Criterion A1: point assumption for p_2: for example $p_2 = p_{2,0}$

Criterion A2a: uniform distribution for p_2 over the complete interval $[0, 1]$: $p_2 \sim U[0, 1]$

Criterion A2b: uniform distribution for p_2 over a restricted interval $[a, b]$: $p_2 \sim U[a, b]$

Criterion A3a: triangular distribution for p_2 over the complete interval $[0, 1]$: $p_2 \sim Tri(0, 1, c)$

Criterion A3b: triangular distribution for p_2 over a restricted interval $[a, b]$: $p_2 \sim Tri(a, b, c)$

Criterion A4 normal distribution for p_2: for example $p_2 \sim N(p_{2,0}, p_{2,0}(1 - p_{2,0})/n)$

For the above criteria, the PET can be calculated as follows:

$$PET_{A1} = \sum_{x_1=0}^{s_1} \binom{n_1}{x_1} p_{2,0}^{x_1}(1 - p_{2,0})^{n_1 - x_1} \tag{3}$$

$$PET_{A2a} = \int_0^1 \sum_{i=0}^{s_1} \binom{n_1}{i} p_2^i (1 - p_2)^{n_1 - i} \mathrm{d}p_2 \tag{4}$$

$$PET_{A2b} = \int_a^b \tfrac{1}{b-a} \sum_{i=0}^{s_1} \binom{n_1}{i} p_2^i (1 - p_2)^{n_1 - i} \mathrm{d}p_2 \tag{5}$$

$$PET_{A3a} = \int_0^{p_{2,0}} \tfrac{2p_2}{p_{2,0}} \sum_{i=0}^{s_1} \binom{n_1}{i} p_2^i (1 - p_2)^{n_1 - i} \mathrm{d}p_2 + \int_{p_{2,0}}^1 \tfrac{2(1-p_2)}{1-p_{2,0}} \sum_{i=0}^{s_1} \binom{n_1}{i} p_2^i (1 - p_2)^{n_1 - i} \mathrm{d}p_2 \tag{6}$$

$$PET_{A3b} = \int_a^c \tfrac{2(p_2-a)}{(b-a)(c-a)} \sum_{i=0}^{s_1} \binom{n_1}{i} p_2^i (1 - p_2)^{n_1 - i} \mathrm{d}p_2 + \int_c^b \tfrac{2(b-p_2)}{(b-a)(b-c)} \sum_{i=0}^{s_1} \binom{n_1}{i} p_2^i (1 - p_2)^{n_1 - i} \mathrm{d}p_2 \tag{7}$$

$$PET_{A4} = \int_{-\infty}^{+\infty} \phi\left(\tfrac{p_2 - p_{2,0}}{\sqrt{p_{2,0}(1-p_{2,0})/n}}\right) \sum_{i=0}^{s_1} \binom{n_1}{i} p_2^i (1 - p_2)^{n_1 - i} \mathrm{d}p_2, \tag{8}$$

where ϕ denotes the density of the standard normal distribution. Employing the corresponding expression for PET, the expected sample size for the proposed design is given by $ESS = n_1 + (n - n_1)(1 - PET)$ while the expected trial length is given by $EL = l_1 + (l - l_1)(1 - PET)$. By this, solutions for s_1, n_1, r and n minimising ESS or EL can be identified. Let us denote by s the recruitment speed per time unit, by $n_{1,short}$ and $n_{1,long}$ (n_{short} and n_{long}) the sample size for the first stage (the total sample size) when applying the proposed design or the design using the results for the long-term endpoint for decision making in the interim analysis, respectively, and by FU_{short} and FU_{long} the follow-up time for the short- and the long-term endpoint. Then the difference in time until the decision can be made whether or not to continue to the second stage is given by $t_{1,long} - t_{1,short} = (n_{1,long} - n_{1,short})/s + (FU_{long} - FU_{short})$. If the trial proceeds to stage two, the difference in time until the final results are available is given by $t_{long} - t_{short} = (n_{long} - n_{short})/s + (FU_{long} - FU_{short})$.

4 Results

In the following, we report the results for the proposed design by considering three different scenarios for short-term endpoint (2-, 4-, and 6-month PFS) for decision making after the first stage together with the primary endpoint 9-month PFS. For the primary endpoint, we assume $p_{1,0} = 0.127$ and $p_{1,1} = 0.38$. As Ray-Coquard et al., we used slightly rounded values for $p_{2,0}$ and $p_{2,1}$. In order to calculate the power, we assumed $p_{2,1} = 0.9$ for

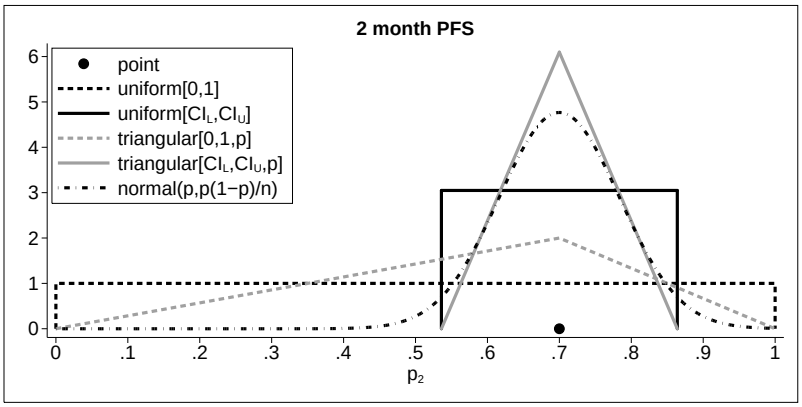

Figure 2. Density of prior distribution for different assumptions on the 2-month PFS rate.

2-month PFS, $p_{2,1} = 0.6$ for 4-month PFS, and $p_{2,1} = 0.4$ for 6-month PFS. For criterion A1, we assumed $p_{2,0} = 0.7$, $p_{2,0} = 0.4$, and $p_{2,0} = 0.2$ for 2-, 4-, and 6-month PFS, respectively. For criterion 2b and 3b, the limits of the confidence interval were calculated using the Wald-type method [19]: $p_{2,0} \pm c\sqrt{p_{2,0}(1-p_{2,0})/n_{\text{pilot}}}$ where $n_{\text{pilot}} = 27$ denotes the sample size from a previous trial (in this case, the sample size used by Penel et al. [17]). Note that in the example given here the endpoints are nested, hence, $p_{12} = p_{1,1} = 0.38$.

4.1 Critical values and samples sizes

In a first step, we searched for solutions for the critical values s_1 and r as well as the sample sizes n_1 and n. As the minimax and the optimal solution based on the 9-month PFS rate only had a maximum sample size of 43 and 45, respectively, we restricted our search to a maximum sample size of 60 or less assuming that by then we would have found the optimal solution. Hence, we searched over $n \in (2, 60)$, over $n_1 \in (1, n-1)$, over $s_1 \in (0, n_1 - 1)$, and over $r \in (0, n)$. For each combination of (s_1, n_1, r, n), we calculated the type I error rate and the power. If a combination met the criteria for α and β, the combination is stored until n reaches its maximum value. In a second step, we calculated the ESS for each solution under each assumption for $p_{2,0}$ in order to identify the minimax and the optimal solution. Two *Stata* functions are available from https://sites.google.com/site/jmswason/supplementary-material. The first function can be used to find the critical values and sample sizes and the second function can be used to calculate the expected sample size and trial length.

Table 2 lists the solutions for Simon's design based on 9-month PFS only as well as the solutions for using a secondary short-term endpoint at the first stage for different assumptions about $p_{2,0}$. The first column shows the endpoints that are used, for example 4-month PFS at the first stage and 9-month PFS at the second stage. Columns 2 to 4 list the values for $p_{1,0}$, $p_{1,1}$, and $p_{2,1}$. Columns 5 and 6 show the optimisation criteria. The correlation Φ between the short-term and the long-term endpoints is given in column 7. The critical values and sample sizes can be found in columns 8 to 11. The expected sample size and the probability of early termination are given in columns 12 and 13. Columns 14 and 15 give the type I error and the power. The last three columns give the length of the first stage, the length of the whole trial, and the expected length (measured in months). Numbers in bold highlight unique solutions with respect to s_1, n_1, r, and n.

The minimax solution used by Ray-Coquard et al. had a total sample size of $n = 43$. In order to reject the null

Table 2. Minimax and optimal solution for Simon's design based on 9-month PFS only as well as for using a short-term endpoint at the first stage for different assumptions about p_2.

endpoint	$p_{1,0}$	$p_{1,1}$	$p_{2,1}$	solution		Φ	s_1	n_1	r	n	ESS	PET	α	$1-\beta$	l_1	l	EL
9-month PFS	12.7%	31.7%	–	minimax		–	3	26	8	43	33.22	0.576	0.084	0.951	21.5	38.5	28.72
				optimal		–	3	24	8	45	31.64	0.636	0.097	0.953	20.5	39.5	27.41
6 + 9-month PFS	12.7%	31.7%	40%	minimax	A1	0.83	**4**	**24**	**8**	**43**	34.26	0.460	0.088	0.951	17.5	35.5	27.22
					A2a		**3**	**21**	**8**	**43**	39.00	0.182	0.088	0.952	16	35.5	31.95
					A2b		3	21	8	43	33.41	0.436	0.088	0.952	16	35.5	27.00
					A3a		3	21	8	43	39.04	0.180	0.088	0.952	16	35.5	31.99
					A3b		4	24	8	43	33.84	0.482	0.088	0.951	17.5	35.5	26.83
					A4		4	24	8	43	33.86	0.481	0.088	0.951	17.5	35.5	26.84
				optimal	A1		**4**	**23**	**8**	**44**	33.49	0.501	0.099	0.953	17	36	26.49
					A2a		3	21	8	43	39.00	0.182	0.088	0.952	16	35.5	31.95
					A2b		**2**	**16**	**8**	**44**	32.35	0.416	0.099	0.952	13.5	36	26.64
					A3a		3	21	8	43	39.04	0.180	0.088	0.952	16	35.5	31.99
					A3b		2	16	8	44	33.14	0.388	0.099	0.952	13.5	36	27.28
					A4		2	16	8	44	32.97	0.394	0.099	0.952	13.5	36	27.13
4 + 9-month PFS	12.7%	31.7%	60%	minimax	A1	0.56	**8**	**24**	**8**	**43**	36.77	0.328	0.078	0.953	15.5	33.5	27.60
					A2a		**4**	**15**	**8**	**43**	34.25	0.313	0.081	0.951	11	33.5	26.47
					A2b		4	15	8	43	34.90	0.289	0.081	0.951	11	33.5	26.99
					A3a		4	15	8	43	35.51	0.268	0.081	0.951	11	33.5	27.48
					A3b		4	15	8	43	35.84	0.256	0.081	0.951	11	33.5	27.74
					A4		4	15	8	43	35.42	0.271	0.081	0.951	11	33.5	27.41
				optimal	A1		**4**	**14**	**8**	**44**	35.62	0.279	0.086	0.951	10.5	34	27.44
					A2a		4	14	8	44	34.00	0.333	0.086	0.951	10.5	34	26.17
					A2b		4	14	8	44	33.86	0.338	0.086	0.951	10.5	34	26.06
					A3a		4	14	8	44	35.01	0.300	0.086	0.951	10.5	34	26.96
					A3b		4	14	8	44	34.67	0.311	0.086	0.951	10.5	34	26.69
					A4		4	14	8	44	34.31	0.323	0.086	0.951	10.5	34	26.41
2 + 9-month PFS	12.7%	31.7%	90%	minimax	A1	0.23	**13**	**19**	**8**	**43**	30.37	0.526	0.053	0.950	11	31.5	20.71
					A2a		**5**	**9**	**8**	**43**	22.60	0.600	0.071	0.951	6	31.5	16.20
					A2b		13	19	8	43	30.86	0.509	0.053	0.950	11	31.5	21.13
					A3a		5	9	8	43	24.86	0.534	0.071	0.951	6	31.5	17.89
					A3b		13	19	8	43	30.67	0.514	0.053	0.950	11	31.5	20.97
					A4		13	19	8	43	30.74	0.511	0.053	0.950	11	31.5	21.02
				optimal	A1		**13**	**18**	**9**	**50**	28.64	0.667	0.095	0.950	10.5	35	18.65
					A2a		**3**	**6**	**8**	**44**	22.29	0.571	0.079	0.951	4.5	32	16.29
					A2b		**10**	**15**	**8**	**44**	30.18	0.476	0.062	0.954	9	32	21.04
					A3a		**6**	**10**	**8**	**44**	23.96	0.589	0.073	0.954	6.5	32	16.97
					A3b		13	18	9	50	29.77	0.632	0.095	0.950	10.5	35	19.51
					A4		10	15	8	44	30.12	0.479	0.062	0.954	9	32	20.99

Note: Numbers in bold highlight unique solutions with respect to s_1, n_1, r, and n.

hypothesis more than $r = 8$ successes are needed. Comparing the minimax solutions for the proposed design with Simon's design, we see that in all cases the same critical value and sample size is used. Hence, in the case that the trial proceeds to the second stage the decision at the end of the trial would be the same and does not depend on the endpoint that is used for the first stage. However, the sample size and critical value for the first stage depend on the endpoint being used. While the original minimax solution had a sample size of $n_1 = 26$, the new designs all have a lower sample size which can be as small as 9. The ESS varies between 22.6 and 39.04 compared to an ESS of 33.22 for the original design. However, the ESS does depend on the assumptions about the short-term endpoint. When comparing the attained type I error rate, we see that some solutions can be highly conservative. Ray-Coquard et al. allowed the type I error rate to be up to 10% but some solutions show an actual type I error rate of only 5.3%.

When looking at the solutions based on the optimal criterion, we see that an optimal solution based on the long-term endpoint only would need a total sample size of $n = 45$ with a critical value of $r = 8$. Most optimal solutions that use a short-term endpoint at the first stage require a maximum sample size of only 44 or even only 43. However, the critical value is still 8. Hence, while a saving in sample size can be achieved, the same number of successes is still needed in order to reject the null hypothesis.

Overall, we see that the different assumptions about the distribution of $p_{2,0}$ may lead to different solution for the critical values and sample sizes but does not have to. For example, when using 4- and 9-month PFS, the same optimal solution results irrespective of the assumption about $p_{2,0}$, while four different optimal solutions result if 2-month PFS is used instead.

4.2 Power, probability of early termination and expected sample size

As under the null hypothesis we do not make any assumptions about the rate for the secondary endpoint, we investigated how much the characteristics of the solutions depend on the value for the response rate for the short-term endpoint. Figure 3 shows the power, probability of early termination, expected sample size, and expected length of the trial for all unique minimax solutions. An analogous figure for the optimal solutions can be found in the *Supplemental Material*. While the PET, the ESS, and the EL do not depend on the value for the response rate for the primary endpoint, the power does. In order to calculate the power, we assumed $p_1 = p_{1,1} = 0.317$. By definition of the endpoints, this means that for the power calculation the response rate for the short-term endpoint is at least 0.317.

For 2-month PFS, the power drops considerably if the actual response rate is lower than the value that has been used in order to obtain the design solution. Based on Table 1, we assumed that the 2-month PFS rate would be at least 0.9 if the treatment is working. The obtained power is indeed above 0.95 if the response rate is larger than 0.9 but may drop to nearly 0 if the response rate for the secondary endpoint is much less. The loss in power results from a high probability of early termination if the response rate is small. The PET also impacts on the expected sample size which is close to n_1 if the PET is high and closer to n if the PET is low. For the other short-term endpoints, the loss in power is less pronounced as less extreme assumptions about the response rate were made: For 4-month PFS, we assumed a response rate of 0.6 and for 6-month PFS, we assumed a response rate of 0.4. On the other hand, the PET is also smaller leading to a higher ESS.

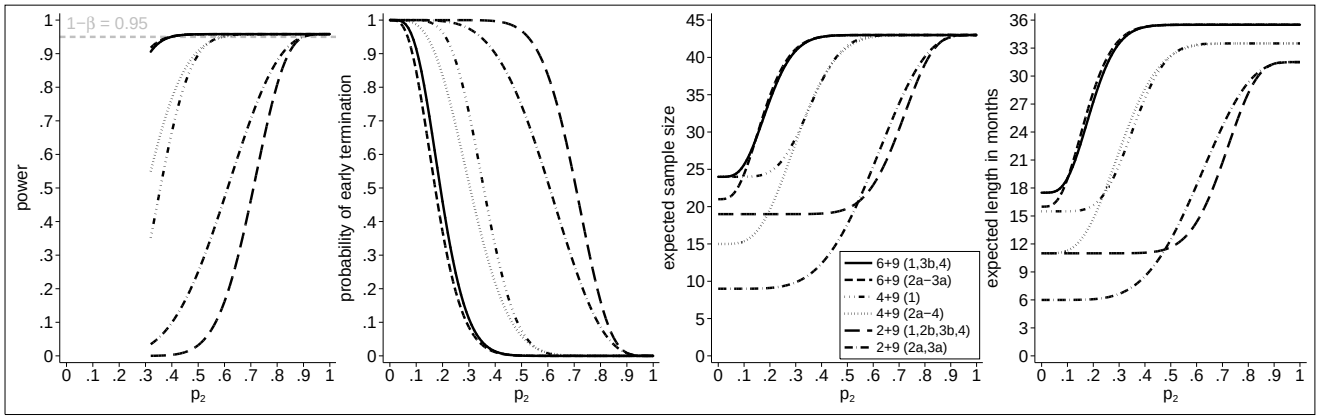

Figure 3. Power, probability of early termination, expected sample size, and expected trial length depending on the value for p_2 for all unique minimax solutions depicted in Table 2.

4.3 Length of the trial

The main reason to use a short-term endpoint at the end of the first stage is to shorten the length of the first stage and subsequently the length of the whole trial. Ray-Coquard et al. [16] report that between June 2008 and June 2009 $n_1 = 26$ patients were enrolled into the first stage of the trial. Enrollment was then halted until the endpoint was observed for all 26 patients. Assuming that there are 360 days in a year (30 days per month) a new patient was enrolled roughly every 15 days. Based on a follow-up time of 270 days (9 months) the outcome for the last patient was available about 645 days after the start of the trial. Hence, the decision whether or not to continue to the second stage of the trial could only be made 21.5 months after the trial was started. If the trial would proceed to the second stage, the final results are available after about 38.5 months after the trial started. Column 16, 17, and 18 of Table 2 list the length of the first stage, the length of the whole trial (if it continues to the second stage) and the expected length. For example, if a 4+9-month PFS minimax (A2a, A2b, A3a, A3b, A4) design would have been used instead of Simon's design, the results for the first stage would have been available after 330 days (11 months). Patient enrollment would have only been interrupted for 4 (instead of 9) months and the final results for the trial would have been available after 33.5 months. Hence, the gain in length for the first stage is 11.5 months and the gain in length for the second stage is 5 months. For the examples we investigated, all designs that use a short-term endpoint have a shorter length for the first and second stage and most designs have a shorter expected length compared to the design used by Ray-Coquard et al. However, the length of the trial depends on the sample sizes for the first and second stage as well as on the accrual rate. Hence, using a short-term endpoint for the first stage does not always result in a shorter trial length.

4.4 Influence of the correlation

For the examples we described above, the correlation between the endpoints is "fixed" as the endpoints are nested and thus $p_{12} = p_1$. Hence, once we have made an assumption about p_2, the correlation Φ can only take one value. However, combinations of short- and long-term endpoints exist where the correlation can vary. For example, let us consider a trial where the primary endpoint is the response rate (defined as either showing a complete or a partial response) after 6 months. At the first stage, the decision whether to continue or not will be based on the 2 month response rate. From Table 1, we see that the response rate after 2 and 6 months is roughly 20%. However,

in this case, endpoints are not nested. As we do not know how many of the patients showing a response after 2 months would show a response after 6 months, an assumption about the value of p_{12} has to be made. From a statistical point of view, p_{12} can vary between 0 and 0.2 and hence, the correlation Φ between the endpoints can vary between -0.25 and 1 (see definition of Yule's correlation).

The most extreme case would be that we *know* the value of the correlation to be $\Phi = 1$. In this case, we could simply use Simon's design. For $p_{1,0} = 0.20$, $p_{1,1} = 0.40$, $\alpha = 0.1$, and $\beta = 0.05$, the characteristics of the resulting minimax and optimal solutions can be found in Table 3.

Table 3. Minimax and optimal solutions for 6-month CR+PR

endpoint	$p_{1,0}$	$p_{1,1}$	solution	s_1	n_1	r	n	ESS	PET	α	$1-\beta$
6-month CR+PR	20%	40%	minimax	3	21	12	45	36.11	0.370	0.097	0.950
			optimal	5	25	13	50	34.58	0.617	0.097	0.953

However, let us now assume that we have *estimated* a correlation of $\hat{\phi} = 1$. If we want to use the short-term instead of the long-term endpoint for decision making after the first stage, we now have to take into account that the actual correlation might not be 1. Hence, in order to control the type I error rate in the strong sense, we have to allow for all possible values for the correlation and for the response rate for the secondary endpoint. In this case, using the critical values and sample sizes in Table 3 might lead to an inflation of the type I error rate. As explained in Section 3.2, the maximum type I error rate is given by Equation (2). For the minimax solution in Table 3, we get a maximum type I error rate of 0.099 which is still below the pre-specified significance level of $\alpha = 0.1$. However, for the optimal solution from Table 3, we get a maximum type I error rate of 0.111, which is larger than 0.1. In general, if the correlation is not known, application of Equation (2) ensures control of the type I error rate irrespective of the value of Φ. However, the power does depend on the assumption about the correlation which can be seen from Equation (1) by replacing p_{12} with $\Phi\sqrt{p_1(1-p_1)p_2(1-p_2)} + p_1p_2$.

We investigated the influence of the correlation on power for designs with 2-month and 6-month CR+PR as endpoints using four different values for $\Phi = -0.25, 0, 0.5, 1$. Based on Equations (1) and (2), we then searched for critical values s_1, r and sample sizes n_1, n so that the type I error is at most $\alpha = 0.1$ and the power is at least 0.95. A table showing all designs including their full characteristics can be found in the *Supplemental Material*. Figure 4 shows the power depending on the actual correlation between the two endpoints for all unique solutions for designs based on 2-month and 6-month CR+PR. Overall, it seemed that while the critical values and sample sizes can be different depending on the assumption about the correlation, once a solution for s_1, n_1, r, n has been found the power does not change much if the actual correlation between the endpoints differs from the correlation assumed at the planning stage of the trial.

5 Design guidance

The previous sections covered the mathematical background needed to assess the characteristics of different designs. In the following, we provide some guidance on how to select a design. We propose the following procedure:

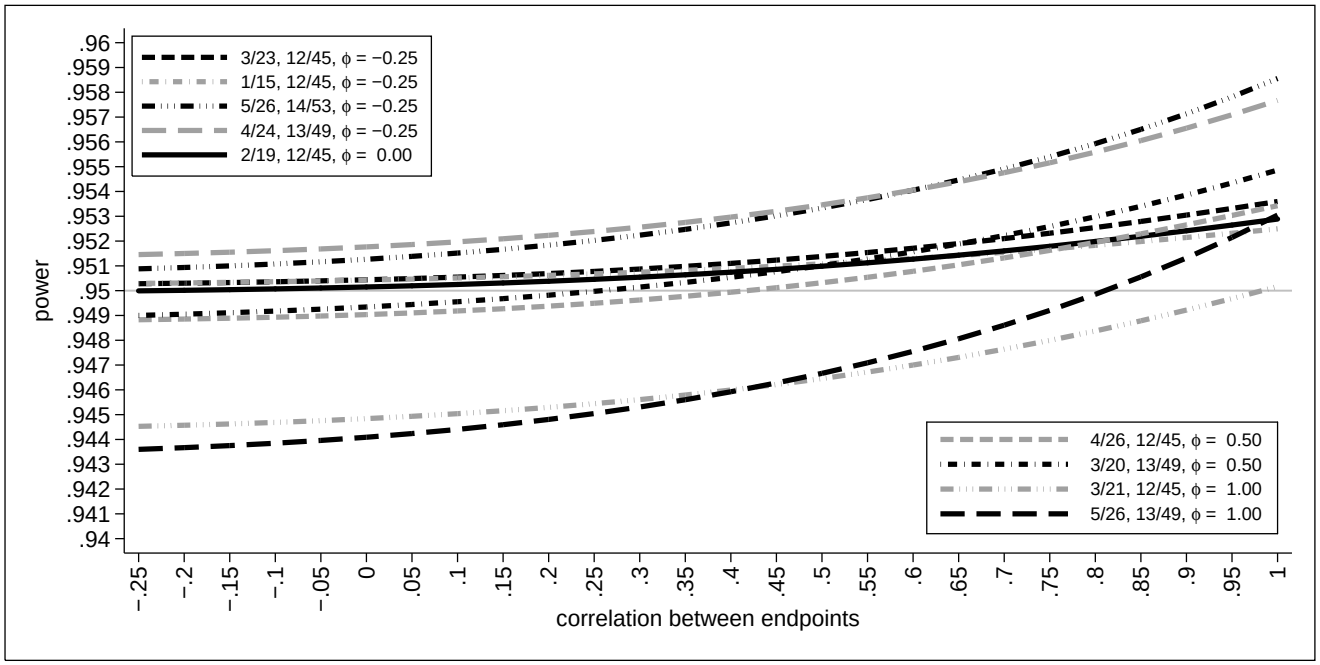

Figure 4. Power depending on the correlation Φ for all unique solutions depicted in Table 1S included in the *Supplemental Material*.

1. Choose an appropriate primary endpoint and specify $p_{1,0}$, $p_{1,1}$, α, and $1-\beta$
2. Decide which short-term endpoints will be available during the course of the trial and collect as much information about the response rate for the short-term endpoints from previous trials
3. Based on the information available for the short-term endpoints, choose an appropriate distribution (A1-A4)
   - A1 could be used if a point estimator based on a large sample size is available
   - A2a could be used if no information about the response rate is available
   - A2b could be used if (for example, based on the definition of the endpoint) a lower and upper bound of the response rate can be defined but a point estimator is not available
   - A3a could be used if a point estimator is available but neither a lower nor an upper bound can be defined (for example, because the sample size on which the point estimator was obtained is unknown)
   - A3b could be used if a point estimator as well as a lower and upper bound for the response rate exists (for example, the lower and upper confidence interval limits can be used as lower and upper bound for the integral)
   - A4 could be used instead of A3b. However, in most cases both assumptions lead to the same solutions for the critical values and sample sizes.
4. Set a value for the response rate for the short-term endpoint if the treatment is efficient and a value for the correlation between the two-endpoints.
5. Calculate critical values and sample sizes for Simon's design based on the long-term endpoint and for designs based on the short- and long-term endpoints

6. Compare the properties of the resulting designs with short- and long-term endpoint to Simon's design (focus especially on the length of the first stage and of the whole trial)

7. Choose a solution with a shorter length than Simon's design and investigate how much the power of this solution depends on the value for p_2

8. If the probable loss in power is acceptable, use the solution for your trial. If the loss in power is unacceptable as compared to the gain obtained by the shorter trial duration, go back to step 6.

Although the power does not seem to depend much on the assumed correlation between the endpoints, a considerable loss in power can occur if the response rate for the short-term endpoint is smaller than assumed. Hence, it is advisable to choose slightly lower values for both the correlation and especially the response rate for the short-term endpoint than the values observed in previous trials.

6 Discussion

In this paper we have investigated the use of intermediate endpoints in the Simon two-stage design, a very commonly used phase II oncology trial design. In oncology phase II trials, the endpoint used to assess efficacy is often observed with considerable delay after recruitment of a patient. For example the trial reported in Ray-Coquard et al. [16] used nine-month progression-free survival. This delay causes problems in trials with interim analyses, such as the Simon two-stage design. In this paper we assume that recruitment is paused once the planned number of first stage patients is reached, so that their response to treatment can be found and an interim decision made before recruitment of second stage patients begins. In this case, delay causes the pause in recruitment to be longer, and so increases the length of the trial. We show how a Simon two-stage design can be modified so that a correlated intermediate endpoint is used to make the interim decision. In all cases the type I error rate is controlled, although the power depends on the actual correlation between the intermediate and final endpoints. If the correlation is moderately high, this trial design can usefully reduce the length of a phase II trial whilst maintaining the power. Reducing the length of phase II trials is of practical interest, as it will increase efficiency of the drug development process.

An alternative effect of delay would occur if recruitment was not paused at the interim analysis - second stage patients would be recruited whilst the first stage patients are still being followed up. This phenomenon is called overrun and is a problem for two-stage designs as those second stage patients are recruited regardless of whether the trial terminates for futility or not. Overrun will therefore result in the expected sample size of the trial being higher. The impact of overrun can also be reduced by using an intermediate endpoint. It is possible that the optimal solution in the presence of delay will be somewhat different to the optimal solution when delay is not considered. This is a useful topic for further research, especially for larger multi-centre phase II trials where it is not easy to pause recruitment at an interim analysis.

An important consideration in two- or multi-stage designs is estimation of the treatment effect following the trial. It is well known that using the maximum likelihood estimator following a multi-stage trial is biased [20], but this is frequently ignored in practice. The bias is generally lower when the interim decision is based on a correlated endpoint [21], but gets higher as the correlation increases. Methods of estimation for secondary endpoints [22]

can be adapted for the trial design we have proposed in this paper by swapping the roles of the primary and secondary endpoint.

We have considered several configurations of the probability of success for the intermediate outcome in order to power the trial and find optimal solutions. Optimal solutions perform well under the configuration they are optimal for, but often perform poorly for other configurations [23]. An admissible solution extends an optimal one to consider more than one criteria. They were first proposed by Jung et al. [24] to balance the expected sample size under the null hypothesis and the maximum sample size. Since then they have been considered for a wide variety of other situations, including balancing several expected sample size criteria [25] and balancing the expected sample size and bias and mean squared error of estimators [26]. The ideas from admissible solutions could be used in the context of our work to allow consideration of several possible configurations for the treatment effect of the intermediate endpoint.

In this paper, we propose a design where information about an early endpoint is only used to allow for early stopping of the trial. However, in some cases, information about the short-term endpoint might also be used to decide whether a treatment is effective of not. Several designs exist that incorporate two endpoints (see, for example, the papers by Bryant and Day [10], Lin and Chen [13], Panageas et al. [14], Lu, Jin, and Lamborn [15], Lin, Allred, and Andrews [12], Kunz and Kieser [11]) and these methods may be applied in such situations.

Funding

CUK is funded by the Medical Research Council [grant number G1001344], JW is funded by the Medical Research Council [grant number G0800860] and the NIHR Cambridge Biomedical Research Centre and MK gratefully acknowledges funding support from the Deutsche Forschungsgesellschaft (DFG) [grant number KI 708/1-2].

References

[1] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). European Journal of Cancer. 2009;45(2):228–247.

[2] Adjei A, Christian M, Ivy P. Novel designs and end points for phase II clinical trials. Clinical Cancer Research. 2009;15:1866–1872.

[3] Sharma M, Maitland M, Ratain M. RECIST: no longer the sharpest tool in the oncology clinical trials toolbox. Cancer Research. 2012;72:5145–5149.

[4] Gan HK, Grothey A, Pond GR, Moore MJ, Siu LL, Sargent D. Randomized phase II trials: inevitable or inadvisable? Journal of Clinical Oncology. 2010;28(15):2641–2647.

[5] Administration UFaD. Drugs@FDA; 2014. Available from: http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm.

[6] Downing NS, Aminawung JA, Shah ND, Krumholz HM, Ross JS. Clinical trial evidence supporting FDA approval of novel therapeutic agents, 2005-2012. JAMA. 2014 Jan;311(4):368.

[7] Baghdadi R, Laffler MJ. The next phase in oncology: FDA's Pazdur has new vision for drug development. The Pink Sheet. 2014 Nov;.

[8] Simon R. Optimal two-stage designs for phase II clinical trials. Controlled Clinical Trials. 1989;10(1):1–10.

[9] Hampson LV, Jennison C. Group sequential tests for delayed responses. Journal of the Royal Statistical Society Series B. 2013;75:1–37.

[10] Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics. 1995;51(4):1372–1383.

[11] Kunz CU, Kieser M. Optimal two-stage designs for single arm Phase II oncology trials with two binary endpoints. Methods of Information in Medicine. 2011;50:372–377.

[12] Lin X, Allred R, Andrews G. A two-stage phase II trial design utilizing both primary and secondary endpoints. Pharmaceutical Statistics. 2008;7(2):88–92.

[13] Lin SP, Chen TT. Optimal two-stage designs for phase II clinical trials with differentiation of complete and partial responses. Communications in Statistics - Theory and Methods. 2000;29(5/6):923–940.

[14] Panageas KS, Smith A, Gönen M, Chapman PB. An optimal two-stage phase II design utilizing complete and partial response information seperately. Controlled Clinical Trials. 2002;23(4):367–379.

[15] Lu Y, Jin H, Lamborn KR. A design of phase II cancer trials using total and complete response endpoints. Statistics in Medicine. 2005;24(20):3155–3170.

[16] Ray-Coquard I, Italiano A, Bompas E, Le Cesne A, Robin YM, Chevreau C, et al. Sorafenib for patients with advanced angiosarcoma: a phase II trial from the French Sarcoma Group (GSF/GETO). The Oncologist. 2012;17(2):260–266.

[17] Penel N, Bui BN, Bay JO, Cupissol D, Ray-Coquard I, Piperno-Neumann S, et al. Phase II trial of weekly paclitaxel for unresectable angiosarcoma: the ANGIOTAX study. Journal of Clinical Oncology. 2008;26(32):5269–5274.

[18] Yule GU. On the methods of measuring association between two attributes. Journal of the Royal Statistical Society. 1912;75(6):579–652.

[19] Vollset SE. Confidence intervals for a binomial proportion. Statistics in Medicine. 1993;12(9):809–824.

[20] Koyama T, Chen H. Proper inference from Simon's two-stage designs. Statistics in Medicine. 2008;27(16):3145–3154.

[21] Choodari-Oskooei B, Parmar M, Royston P, Bowden J. Impact of lack-of-benefit stopping rules on treatment effect estimates of two-arm multi-stage (TAMS) trials with time to event outcome. Trials. 2013;14:1–15.

[22] Kunz CU, Kieser M. Estimation of secondary endpoints in two-stage phase II oncology trials. Statistics in Medicine. 2012 Dec;31(30):4352–4368.

[23] Wason JMS, Mander AP, Thompson SG. Optimal multi-stage designs for randomised clinical trials with continuous outcomes. Statistics in Medicine. 2012;31:301–312.

[24] Jung SH, Lee T, Kim KM, George SL. Admissible two-stage designs for phase II cancer clinical trials. Statistics in Medicine. 2004;23(4):561–569.

[25] Mander AP, Wason JMS, Sweeting MJ, Thompson SG. Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. Pharmaceutical Statistics. 2012;11:91–96.

[26] Bowden J, Wason J. Identifying combined design and analysis procedures in two-stage trials with a binary end point. Statistics in Medicine. 2012;31(29):3874–3884.