

## Single arm two-stage studies: Improved designs for molecularly targeted agents

P. Dutton<sup>1</sup>, J. Holmes<sup>1</sup>

1. Centre for Statistics in Medicine (CSM), Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Botnar Research Centre, Windmill Road, Oxford, OX3 7LD, UK

Corresponding author

P Dutton

[peter.dutton@ndorms.ox.ac.uk](mailto:peter.dutton@ndorms.ox.ac.uk)

Abstract word count: 232

Paper word count: 3981

Keywords: biomarker, Bayesian, two-stage, stratified medicine, molecularly targeted agent

### Abstract

Mechanistic understanding of cancers and their potential interactions with molecularly targeted agents is driving the need for stratified medicine to ensure each participant receives the best possible care. This understanding, backed by scientific research, should be used to guide the design of clinical trials for these agents. The mechanism of action of a molecularly targeted agent often suggests that a biomarker can be used as a predictor of activity of the agent on the targeted disease. A biomarker driven trial is needed to confirm that the molecularly targeted agent stratifies the participant population with disease into high and low responder groups. We assume that the biomarker of interest can be dichotomised and propose a balanced parallel two-stage single-arm phase II trial that builds on existing two-stage single-arm designs. A single arm trial cannot distinguish between a marker being predictive in the population as a whole and the agent causing an increased response in the marker positive group, but it is a first step. We compare this approach to the existing single-arm approached, sequential enrichment, tandem two-stage, and parallel two-stage designs, and discuss the advantages and disadvantages of each design. We show that our design compares favourably to existing designs in the Bayesian framework, making a more efficient use of collected data. We recommend using the parallel two-stage balanced or sequential enrichment designs when randomisation is not practical in a phase II trial.

### 1. Introduction

New cancer treatments increasingly target the biology of cancers. As every participant has a different genetic makeup, so too does every tumour. Participants' response to treatment is thus likely to be heterogeneous. The underlying scientific understanding behind the mechanisms of molecularly targeted agents (MTA) suggests that there is a biomarker that can predict a participant's likelihood of response to each MTA. These continuous biomarkers can be used to dichotomise the participant population. When our understanding of an MTA's mechanism is sufficiently strong, we can assume the response in one group (marker-positive) will be greater than in the other (marker-negative). We therefore need to consider just three possible outcomes: the MTA can be effective in the entire eligible population, effective only in the marker-positive group, or ineffective in both groups. We do not need to be able to declare the treatment effective in the marker-negative group alone.

In an ideal world, all phase II and III clinical trials would randomise their participants between standard care (a comparator) and the new experimental treatment regime. As costs, lack of a comparator, and small populations can make this ideal impractical, phase II trials are often run without randomisation. Many such trials have traditionally used Simon's two-stage design<sup>1</sup>, in which a single arm two-stage trial looks for signs of drug activity using a binary outcome in a single target population. The trial is paused at a planned interim analysis (stage one) and, if there are less than a pre-specified number of responses, the trial stops for futility. Otherwise the trial continues to recruit participants and a final decision is made at the end of the second stage of the trial, based on all of the recruited participants. There are many other approaches that have built on Simon's designs<sup>2-8</sup>.

The literature contains some suggestions for trial designs when a population is dichotomised by a biomarker. An independent Simon's two-stage design could be used for each marker group. However, this design ignores the evidence that participants in the marker-positive group should benefit more than participants in the marker-negative group. Instead, we could first recruit from an unselected population (all comers), with the option to later enrich the population by restricting recruitment to a biomarker subgroup. Alternatively, we could start with an enriched population and open recruitment to an unselected population at a later stage. When we were faced with designing a single-arm trial for a biomarker, we found deficiencies in many of these published designs<sup>9-11</sup>, we discuss these deficiencies in section 2. We were also concerned that the most appropriate approach may depend on the prevalence of the biomarker subgroups in the population.

In this paper, we propose new designs for non-randomised MTA phase II trials when there is scientific rationale to expect marker-positive participants to do better than marker-negative participants. We assume a binary marker, which may mean that a continuous variable has been dichotomised with a pre-specified cut-off or similar simplification of more complex data. We compare our design to existing designs and discuss the merits and problems of each design.

### 2. Existing designs

We want to choose between the null and two alternative hypotheses:

$$H_0: \theta^- \leq \theta_0 ; \theta^+ \leq \theta_0$$

$$H_1: \theta^- \leq \theta_0 ; \theta^+ > \theta_0$$

$$H_2: \theta^- > \theta_0; \theta^+ > \theta_0$$

where  $\theta_0$  is some uninteresting probability of a response to the MTA and  $\theta^-$  and  $\theta^+$  are the probabilities of a response by participants in the marker-negative and marker-positive groups, respectively. If the treatment is effective then it may be effective in only the marker-positive group ( $H_1$ ) or it may be effective in both groups ( $H_2$ ). We define  $\theta_1$  to be the smallest probability of a response that warrants further investigation in a larger, possibly randomised trial, and design and evaluate the trial based on  $\theta_1$ . The methodology allows for a different value of  $\theta_1$  for each marker subgroup, but we considered the same value for each group for simplicity.

There are three existing designs that use two ordered subgroups to explore this problem: sequential enrichment<sup>12</sup>, tandem two-stage<sup>11</sup>, and parallel two-stage<sup>9,10</sup>. The sequential enrichment design first recruits only marker-positive participants, then switches to recruiting marker-negative participants if efficacy in the marker-positive group is indicated. The tandem two-stage initially recruits an unselected population, then decides whether to enrich to the marker-positive group at an interim analysis. The parallel two-stage recruits a predefined number of participants from each of the marker groups rather than a completely unselected cohort. This enriches the cohort with participants from the lower prevalence marker group and maximises the value of information from each participant across the two marker groups. A decision to drop the marker-negative group or halt the trial early is made at an interim analysis.

### Sequential enrichment design

The sequential enrichment design<sup>12</sup> is an amalgamation of two separate two-stage trials, proposed by Zang and Yuan. The first study is run in the marker-positive group. If and only if there is evidence of efficacy in the marker-positive group after completing the first two-stage study, then the trial recruits only the marker-negative group to the second study (see Figure 1).

*Figure 1 about here.*

### Tandem two-stage design

The tandem two-stage design<sup>11</sup> is a three-stage design that combines two two-stage designs. First, an unselected group is recruited until the first interim analysis. If the treatment is deemed futile in the unselected group, then a two-stage trial is completed from this point on, recruiting only marker-positive participants (stages 2b and 3 in Figure 2). Marker-positive participants from the first stage are included in this enriched population ( $n_1^+$ ). If the treatment is not deemed futile, the second stage of the trial continues to recruit an unselected group and the final analysis is completed on all recruited participants (stage 2a in Figure 2). The marker subgroups are not tested separately.

*Figure 2 about here.*

When the treatment is effective in the marker-positive group and the prevalence of this group is significant the trial is likely to succeed at the interim analysis and hence only test the unselected group. This then prevents identification of a positive effect in only the marker-positive group. The trial will struggle to enrich the population when it is correct to do so.

### Parallel two-stage design

The parallel two-stage design<sup>9,10</sup> is also an amalgamation of two separate two-stage trials, with the marker distinguishing between the two groups. A pre-specified number of participants are recruited to each marker group for the interim analysis, which enforces enrichment to ensure the desired amount of information is available for each group. When the target number of participants is reached in one group further recruitment to that group is paused to allow recruitment to the other group to complete. The marker-negative group is tested for futility first. If there is some evidence of efficacy, the trial continues to recruit an enriched cohort of participants from both marker groups. If there is sufficient evidence of futility in the marker-negative group, the marker-positive group is tested. If there is sufficient evidence of futility in the marker-positive group, the trial stops for futility. If there is some evidence of efficacy in the marker-positive group, recruitment continues in the marker-positive group alone (Figure 3). At the end of the trial, the marker-negative group is tested for efficacy. The marker-positive group is only tested if there is not sufficient evidence of efficacy in the marker-negative group.

This design uses a sequential testing approach, first testing the marker-negative group. We thus refer to this design as the parallel two-stage negative-first design.

*Figure 3 about here.*

The use of data at the interim in this design is inefficient. The information from the marker-positive group is discarded if there is sufficient evidence in the marker-negative group to continue the trial. If evidence in the marker positive group is negative then the decision to continue the trial should consider this.

### 3. Alternative design

#### Parallel two-stage balanced

We propose a more balanced approach to the parallel two-stage design negative first. First we test for futility in the unselected group. If the treatment is deemed futile in the unselected population, then we test the marker-positive group for futility. The trial stops if the treatment is futile in the marker-positive group, otherwise it continues recruiting only marker-positive participants. If the treatment is not deemed futile in the unselected population, then we test the marker-negative group for futility. The trial continues recruiting from the unselected population if the treatment is not deemed futile in the marker-negative group, otherwise it switches to recruiting only marker-positive participants (Figure 4). A similar testing approach is used at the end of the trial if both biomarker groups are still being recruited. Note that at the end of the trial the population will have been enriched but during the study recruitment is unselected unless recruitment to a group is stopped. Recruitment to a group will be stopped either due to evidence of futility or the pre-specified target number recruited for that group.

*Figure 4 about here.*

We use the posterior predictive probability to find the probability of the trial reaching the threshold for success if the trial were to continue to the end. If this predictive probability is high enough, we continue with the second stage of the trial. The posterior predictive probability combines the binomial distribution of recruiting the remaining  $n - n_{cur}$  with the current uncertainty of the probability of success. Our current estimate of the probability of success is the posterior probability at the time of the analysis. The amalgamation of these

probability distributions gives us a Beta-Binomial distribution for the number of successes at the end of the trial including currently un-recruited participants. The predictive probability is then the probability that we observe the predefined required number of successes at the end of the trial given the current history of the trial.

Under this approach, recruitment to a group only need pause once the target sample size for that group has been achieved. Interim analyses will occur once one of the groups has recruited the target number of participants for the interim. A further interim analysis will occur once one of the groups reaches the total target number.

For easy comparison to the parallel two-stage negative first design we repeat this design waiting for both groups to recruit the required number of participants, 17 at interim and 34 at the final analysis.

### 4. Design Scenario

This work was motivated by a trial in metastatic recurrent renal cancer investigating an MTA in the phase II setting. The MTA was thought to interact with a biomarker, which was used to split the participant population into three ordered groups. To keep this paper simple we have reduced this example to two levels.

All of the existing and proposed approaches can be used in either a frequentist or Bayesian framework. We use a Bayesian hypothesis testing framework with no prior data. We compared the methods using  $\theta_0 = 0.1$  and  $\theta_1 = 0.3$ . We assumed there is a biomarker that can be dichotomised and assumed a weakly informative Beta prior centred between  $\theta_0$  and  $\theta_1$  (Beta(0.05, 0.2)) on  $\theta^-$  and  $\theta^+$ . A maximum sample size of 34 participants per group (68 in total) was selected using the approach of Whitehead et al<sup>13</sup>, with an interim analysis halfway through recruitment (n=17). The two groups are likely to recruit at different rates due to random chance, and differences in the prevalence of the two marker-groups. For the parallel two-stage designs, when there are enough participants in one of the marker groups, recruitment to that group is paused while recruitment to the other group completes. Both groups are then analysed together. Consequently, only the number screened is affected by the marker prevalence for the parallel two-stage and sequential enrichment designs. We explore alternative approaches to interim analysis timing in a later section.

We defined our stopping rules as:

- At the interim analysis, treatment is deemed futile if the posterior probability of futility is greater than 0.98 –  $P(\theta < 0.3 \mid \mathbf{y}) > 0.98$
- At the final analysis, treatment is deemed efficacious if the posterior probability of success is greater than 0.95 –  $P(\theta > 0.1 \mid \mathbf{y}) > 0.95$

Under these stopping rules, a marker-group may stop for futility if one or fewer responses are seen at the interim (n=17) and the treatment is declared effective if seven or more responses are seen at the final analysis (n=34).

For the parallel 2-stage balanced design the interim stopping rule was:

- At the interim analysis, treatment is deemed futile if the posterior predictive probability of success is less than 0.05

We used an exact approach summing up the probabilities of all the possible ways each trial could complete to obtain the type I and type II error rates and the mean sample size. We also calculated the mean number of participants that need to be screened under each scenario using the negative binomial distribution to account for times when only one of the marker groups could be recruited.

*Table 1, about here.*

The results are shown in Table 1. For each method and scenario, we show the probabilities of claiming the treatment is futile, finding it is successful in both biomarker groups, and claiming it is successful in only the marker-positive group. The correct decision for each scenario is shown in bold. Also shown is the mean sample size required and the mean number of participants screened. The properties for two independent trials are given as a reference. We used a prevalence rate of 50% marker-positive participants for this table.

As there is no ordering principle in the two independent parallel trial design, it is possible to recommend further research in the marker-negative group alone. When the treatment was ineffective in both groups, the treatment was deemed futile 90.8% of the time, requiring on average 51.6 participants and screening 63.0 participants. When the treatment was only effective in the marker-positive group, this was correctly identified 87.1% of the time. When the treatment was effective in both groups, this was correctly identified 83.5% of the time. A further 7.9% of the time the wrong alternative hypothesis was selected, as only the marker-positive group was deemed effective.

The tandem two-stage design was not well suited to this problem when the proportion of participants in the marker-positive group was large. The marker-positive group pushed the trial through the first interim analysis most of the time, so could not be tested alone. As a result, when the treatment was only effective in the marker-positive group, this was correctly identified 9.0% of the time. Instead, the combined group was incorrectly recommended 53.1% of the time and the treatment was deemed ineffective in both groups 37.9% of the time.

The parallel two-stage designs fared better. They were all able to successfully identify the effective group with at least 87% power. The negative-first design outperformed the balanced design when the treatment was effective in at least one group.

Although the parallel two-stage negative-first design compared well to the parallel two-stage balanced design, this design made poor decisions in a number of scenarios. The parallel two-stage negative-first design could recommend continuing recruitment to both groups when 2 out of 17 participants in the marker-negative group had a response. This decision would be made regardless of the marker-positive group, which could have 0 out of 17 responses at this analysis time point. A total of 2 responses out of 34 participants is sufficient evidence to deem the treatment futile in the population as a whole and thus the decision to continue is poor.

The sequential enrichment design required a small mean sample size when the treatment was ineffective but required screening more participants than the other designs when the treatment was effective in one or more groups.

*Figure 5 about here.*

A biomarker is unlikely to equally split the population. For the independent parallel, parallel two-stage negative-first, and sequential enrichment designs this does not affect the type I or type II error since analysis occurs when sufficient participants are recruited for each group. For the parallel two-stage balanced design this is not the case. Figure 5 shows the effect of a number of prevalence levels on the probability of making the correct conclusion under the three hypotheses. There is a detrimental effect on the probability of correct conclusion when the prevalence is greater in the marker positive group and both treatments are effective. This is reversed when neither treatment is effective or only the marker-positive group is effective.

*Figure 6 about here.*

For each design and treatment scenario, Figure 6 shows the mean number of participants that would need to be screened to recruit enough participants for different biomarker prevalence levels. The tandem two-stage required the fewest screened participants. The enrichment designs were more efficient when the treatment did not work in either group, but were less efficient when the treatment was effective in at least one group.

## 5. Discussion

We have proposed and described the parallel two-stage balanced design as a method for evaluating activity in a single-arm phase II trial when it is believed marker-positive participants will respond better to a treatment than marker-negative participants. We compared the proposed design to several published methods and found superior operating characteristics as it addresses limitations in these methods.

Compared to the parallel two-stage negative-first design, the balanced design uses all data at each interim analysis. In the parallel two-stage negative-first design<sup>9,10</sup>, the marker-positive group may never be tested despite potentially containing invaluable information, particularly when the treatment is ineffective in both groups. There is a risk of ignoring evidence that the treatment does not work in the marker-positive group, resulting in an increased type I error. Our balanced approach addresses this issue. Both designs had superior running properties to running two independent parallel two-stage designs.

By carrying out the analysis when one group has reached the interim analysis the number of patients required at each analysis reduces the mean number of patients screened for the parallel two-staged balanced design. Consequently, the type I and type II errors are dependent on the prevalence of each marker-group.

The tandem two-stage design enriches the population at the interim analysis if the treatment is deemed futile in the unselected population. However, this method allows the marker-positive group to carry the trial through the futility check at the first interim analysis. The trial may continue with both biomarker groups even when there is no evidence in the marker-negative group to do so. When the trial finishes, the data in the marker-positive group may be sufficient to achieve significance for the unselected group without any positive information from the marker-negative group. Conversely, if the trial does not reach significance in the unselected group, the treatment is deemed futile in all groups. Therefore, we would not recommend using this design.



The sequential enrichment design performs well and exposes the fewest participants to a treatment that is ineffective under the null hypothesis. However, it is slower than the other designs when the treatment is effective because it only recruits one marker group at a time. This duration difference is greatest when the prevalence of each group is similar. This may be the most suitable design on ethical grounds.

We explored the effect of biomarker prevalence levels on the operating characteristics of our proposed balanced design (Figures 5 and 6). By allowing each analysis to be staggered, the mean number of participants required can be reduced. This is particularly prominent when one biomarker level is much more prevalent than the other. Pausing recruitment for interim analyses cost time but used fewer participants on average.

In the parallel two-stage designs, participants do not need to be tested for the biomarker prospectively from the outset. The logistical setup for prospective testing then does not need to be finalised before opening the trial. However, if sequential enrichment is used, marker-positive participants must be identified from the outset for inclusion in the trial.

Non-informative priors have been used throughout this paper. If relevant information exists, informative priors could be used instead.

## 6. Conclusions

The proposed parallel two-stage balanced design performs well under a range of biomarker prevalence levels and treatment effect scenarios. This design uses all available information at each analysis and makes a decision based on this information. We have shown that staggered interim analyses can be used to improve the efficiency of this design.

The proposed designs make small but significant gains above and beyond using two independent parallel two-stage trials. We recommend using either the parallel two-stage balanced or the sequential enrichment design.

## 7. Programs

The programs used to run the simulations and recreate the table and figures is available on github at: <https://github.com/csmoxford/BiomarkerTwoStage>.

## 8. Acknowledgements

We acknowledge English language editing by Dr Jennifer A. de Beyer of the Centre for Statistics in Medicine, University of Oxford.

## 9. References

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*. 1989;10(1):1-10.
2. Ye F, Shyr Y. Balanced two-stage designs for phase II clinical trials. *Clinical trials (London, England)*. 2007;4(5):514-524.
3. Chen K, Shan M. Optimal and minimax three-stage designs for phase II oncology clinical trials. *Contemporary clinical trials*. 2008;29(1):32-41.
4. Chen TT, Ng TH. Optimal flexible designs in phase II clinical trials. *Stat Med*. 1998;17(20):2301-2312.
5. Chen TT. Optimal three-stage designs for phase II cancer clinical trials. *Stat Med*. 1997;16(23):2701-2711.
6. Mander AP, Thompson SG. Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. *Contemporary clinical trials*. 2010;31(6):572-578.

7. Green SJ, Dahlberg S. Planned versus attained design in phase II clinical trials. *Statistics in medicine*. 1992;11(7):853-862.
8. Ensign LG, Gehan EA, Kamen DS, Thall PF. An optimal three-stage design for phase II clinical trials. *Statistics in Medicine*. 1994;13(17):1727-1736.
9. Jones CL, Holmgren E. An adaptive Simon Two-Stage Design for Phase 2 studies of targeted therapies. *Contemporary clinical trials*. 2007;28(5):654-661.
10. Parashar D, Bowden J, Starr C, Wernisch L, Mander A. An optimal stratified Simon two-stage design. *Pharm Stat*. 2016;15(4):333-340.
11. Puztai L, Anderson K, Hess KR. Pharmacogenomic Predictor Discovery in Phase II Clinical Trials for Breast Cancer. *American Association for Cancer Research*. 2007;13(20):6080-6086.
12. Zang Y, Yuan Y. Optimal sequential enrichment designs for phase II clinical trials. *Statistics in Medicine*. 2017;36(1):54-66.
13. Whitehead J, Valdés-Márquez E, Johnson P, Graham G. Bayesian sample size for exploratory clinical trials incorporating historical data. *Statistics in Medicine*. 2008;27(13):2307-2327.

### Figures

Figure 1: Sequential enrichment design

Figure 2: Tandem two-stage design

Figure 3: Interim decision rules for the parallel two-stage negative-first design

Figure 4: Interim decision rules for the parallel two-stage balanced design

Figure 5: Probability of correct conclusion for each trial design, at different biomarker prevalence levels, (left) ineffective in both groups, (middle) effective in the marker-positive group but ineffective in the marker-negative group, and (right) effective in both groups.

Figure 6: Mean number of participants that must be screened for each trial design to recruit enough participants, at different biomarker prevalence levels, when the treatment is (left) ineffective in both groups, (middle) effective in the marker-positive group but ineffective in the marker-negative group, and (right) effective in both groups.

Table 1: Comparison of the operating characteristics and final decisions of each design, computed using exact probabilities

Design	$\theta^-$	$\theta^+$	Futility	Success in marker-positive group	Success in both marker groups	Mean n	Mean screened
Two independent parallel trials	0.1	0.1	<b>0.908</b>	0.045	0.002	51.6	63.0
	0.1	0.3	0.082	<b>0.871</b>	0.043	59.5	71.1
	0.3	0.3	0.007	0.079	<b>0.835</b>	67.3	74.3
Tandem two-stage	0.1	0.1	<b>0.945</b>	0.008	0.047	32.6	39.4
	0.1	0.3	0.388	<b>0.089</b>	0.523	34.9	37.8
	0.3	0.3	0.073	0.013	<b>0.914</b>	34.1	34.6
Parallel two-stage negative-first	0.1	0.1	<b>0.908</b>	0.045	0.047	55.9	67.1
	0.1	0.3	0.078	<b>0.875</b>	0.047	59.7	74.7
	0.3	0.3	0.007	0.079	<b>0.914</b>	67.7	77.1
Parallel two-stage balanced (wait)	0.1	0.1	<b>0.943</b>	0.045	0.012	44.8	57.0
	0.1	0.3	0.083	<b>0.873</b>	0.044	58.3	74.0
	0.3	0.3	0.018	0.087	<b>0.895</b>	67.1	76.8
Parallel two-stage balanced	0.1	0.1	<b>0.926</b>	0.048	0.026	48.9	54.7
	0.1	0.3	0.084	<b>0.871</b>	0.045	57.8	68.8
	0.3	0.3	0.009	0.086	<b>0.905</b>	64.3	68.1
Sequential enrichment	0.1	0.1	<b>0.953</b>	0.044	0.002	27.0	54.0
	0.1	0.3	0.086	<b>0.872</b>	0.043	57.3	114.5
	0.3	0.3	0.086	0.079	<b>0.836</b>	64.5	128.9