OPTIM-ARTS—An Adaptive Phase II Open Platform Trial Design With Application to a Metastatic Melanoma Study

Amy Racine-Poona, Anthony D’Amelio Jr., Oleksandr Sverdlov, and Tomas Haas

Novartis Pharma AG, Basel, Switzerland; Novartis Pharmaceuticals Corporation, East Hanover, NJ

ABSTRACT

Platform trials are increasingly popular in clinical research since these trials can evaluate multiple treatments in targeted subgroups of patients within a single trial infrastructure. In this article, we describe the development and implementation of a new clinical trial design OPTIM-ARTS (Open Platform Trial Investigating Multiple Compounds—Adaptive Randomized Design with Treatment Selection)—with application to a phase II study of an anti-PD-1 antibody in combination with other novel targeted therapies and immuno-agents in metastatic melanoma. The design consists of two main parts—the open platform (selection) phase, in which different treatments are initially assessed for efficacy in a randomized fashion, and the expansion phase, which will allow for a formal examination of efficacy with the selected promising treatment(s). The sample size for the expansion phase is chosen adaptively, using Bayesian shrinkage estimation, to mitigate selection bias due to potential random highs, and to ensure high predictive probability of obtaining significant final results. We performed calibration of the proposed design in a platform trial setting with four treatment arms and compared it with more standard designs. We found that our design is flexible and efficient, and it could be useful in an open platform trial setting with multiple investigational compounds.

1. Introduction

Modern clinical research has focused around the idea of precision medicine—finding the right treatment for the right patient at the right time. Scientific advances in biotechnology and genomics have enabled better understanding of the molecular basis of many fatal diseases such as cancer, which, in turn, led to discovery and development of many novel investigational compounds that hold promise to address high unmet medical needs. In modern clinical oncology, these scientific advances are especially important for patients who have limited still available treatment options (such as melanoma patients who have already failed anti-PD-1/PD-L1 therapy in their particular cancer). The development of these novel treatments requires special considerations in the design and analysis of clinical trials.

The number of potential treatment strategies may be large and the patient populations may be small. Therefore, there is a strong need for efficient research designs that would enable evaluation of multiple treatment options, testing multiple research hypotheses in well-defined patient groups. One approach to facilitate such innovative clinical research is the development of master protocols. A master protocol is defined as an overarching protocol designed to answer multiple research questions within the same overall trial structure (Woodcock and LaVange 2017; Sudhop et al. 2019). Recognizing importance of the topic, the FDA has recently issued a draft guidance on master protocols (FDA 2018). This guidance provides useful insights on scientific and operational aspects of these designs and highlights the importance of continuing discussions among the FDA, pharmaceutical industry, academia, and the public on the topic.

One can distinguish three types of master protocols: umbrella, basket, and platform trials, and the present article will focus on the design of an innovative platform trial. The platform trial is the most complex type of a master protocol. It may involve multiple treatment arms (either a fixed or a flexible number due to possible dropping/adding of the arms), multiple biomarker strata, and can use either fixed or adaptive randomization. With platform trials, we are trying to answer the following research question: which treatment(s) are the most promising for testing in subsequent confirmatory randomized clinical trials? If properly designed and implemented, platform trials can potentially be more efficient than a sequence of single-arm or two-arm trials (Berry, Connor, and Lewis 2015; Saville and Berry 2016; Hobbs, Chen, and Lee 2018).

Platform trials are very relevant in exploratory clinical settings, such as phase II, where the goal is to explore multiple candidate treatments and generate data that could be used to support viable research hypotheses to be tested in confirmatory phase III trials. Some examples of platform trials are I-SPY2 (Barker et al. 2009; Park et al. 2016; Rugo et al. 2016), BATTLE (Zhou et al. 2008; Kim et al. 2011), LUNG-MAP (Ferrarotto et al. 2015; Steuer et al. 2015), to name a few.

Our motivating example is an ongoing randomized, open-label, adaptive phase II open platform trial evaluating efficacy and safety of a novel anti-PD-1 antibody in combination with other novel targeted therapies and immuno-agents in metastatic melanoma for previously treated patients. The main rationale for using an open platform design was that a high number of promising compounds that are combinable with this novel...
anti-PD1 antibody are becoming available at different times, and assessing these compounds within the same trial infrastructure could potentially enable efficient selection of the most promising treatment combinations. Given potential complexity of the study, several design options were considered and evaluated at the planning stage. The implemented platform trial was deemed overall most appropriate; however, other viable options were considered as well, and their statistical properties were evaluated over a range of plausible experimental scenarios.

Recently, several authors proposed statistical designs for platform trials (Berry et al. 2016; Jacob et al. 2016; Yuan et al. 2016; Hobbs, Chen, and Lee 2018; Kaizer, Hobbs, and Koopmeiners 2018; Ventz et al. 2018; Tang, Shen, and Yuan 2019). These methods were developed in particular contexts, including phase II multi-arm trials with or without a common control group, using either fixed (equal) or response-adaptive randomization. Our OPTIM-ARTS design shares some important features of the aforementioned open platform designs, and it also has several distinct elements that make our design flexible and, at the same time, statistically rigorous.

The current article is organized as follows. In Section 2, we describe the statistical methodology of OPTIM-ARTS design and give an illustrative example of a single trial run using this design. In Section 3, we present results of a simulation study under a range of experimental scenarios that are plausible in a clinical research setting for the considered indication. In Section 4, we present additional simulation studies to compare some variants of our proposed OPTIM-ARTS design with a more standard design. Section 5 concludes with a discussion and outlines some important future work.

2. Methodology

2.1. Design Concept

The OPTIM-ARTS design was developed in the context of a randomized phase II clinical trial to evaluate early efficacy of different treatments (combinations of a novel antibody with other targeted therapies and immuno-agents) in patients with unresectable or metastatic melanoma. It can be also adopted in other clinical development programs for different indications. A combination of continuous monitoring with decision criteria based on predictive probability for a multi-arm platform trial was explored by Hobbs, Chen, and Lee (2018), where the authors showed that such an approach could result in considerable efficiency gains. This approach merits investigation but it is beyond the scope of our current work.

Several other approaches to facilitate data monitoring and decision making during Part 1 could be considered. If feasible, a continuous monitoring scheme where decisions are made sequentially, rather than after groups of patients, can be used. Furthermore, go/no-go decision criteria can be cast based on Bayesian predictive probability of success for individual treatment arms rather than posterior probabilities of the treatment effects. A combination of continuous monitoring with decision criteria based on predictive probability for a multi-arm platform trial was explored by Hobbs, Chen, and Lee (2018), where the authors showed that such an approach could result in considerable efficiency gains. This approach merits investigation but it is beyond the scope of our current work.

In the expansion phase (Part 2), the most promising (“winner”) treatment arms from Part 1 will be further evaluated. Additional patients will be enrolled into each “winner” treatment arm to ensure sufficient sample size for statistical hypothesis testing concerning ORR of this treatment at the end of the study. There are several ways to determine the number of additional patients. One possibility is to expand each arm or response-adaptive (e.g., allocation skewed adaptively in favor of empirically better treatment arms). In principle, the design could be also modified to include a control arm (e.g., if the standard of care is available). In the latter case, the allocation ratio and the randomization procedure would have to be specified to account for comparisons of experimental treatments versus control; see Sverdlov and Rosenberger (2013) for a review of available optimal allocation designs for multi-arm trials that can be useful in this setting. However, in the current article we just focus on the case of multiple experimental treatments without the control arm.

The primary efficacy endpoint is the objective response rate (ORR), defined as the proportion of subjects with best overall response of either confirmed complete response or confirmed partial response, as per local review and according to the RECIST v1.1 criteria (National Cancer Institute 2017). In addition, safety, pharmacokinetics (PK), and important biomarker data will be collected and analyzed.

The purpose of Part 1 is to identify promising efficacious arm(s) and terminate futile arm(s). This can be done in a variety of ways. One approach is to apply a two-stage Simon’s design (Simon 1989) for each treatment arm tested in Part 1; this would require prespecification of treatment sample sizes for the interim and final analyses. Another, more flexible approach, is to use a Bayesian design with several interim analyses (IAs) either after a prespecified number of patients, or at some predefined calendar times. With this approach, the observed ORR will be modeled using Bayesian methods, and at each IA a statistical decision rule will be applied to evaluate each treatment arm with respect to its potential: (i) to be expanded into Part 2 (provided there is strong evidence of treatment efficacy); (ii) to be terminated for futility (provided there is strong evidence of lack of efficacy); or (iii) to be continued in Part 1 (if efficacy results are intermediate). The decision rule will be quantified using Bayesian posterior probabilities to ensure prespecified, acceptable rate of decision errors. We also assume that the usual safety monitoring rules are in place, and the trial can potentially be stopped if there are safety findings. However, main design adaptations are based on the efficacy endpoint (ORR).

Several other approaches to facilitate data monitoring and decision making during Part 1 could be considered. If feasible, a continuous monitoring scheme where decisions are made sequentially, rather than after groups of patients, can be used. Furthermore, go/no-go decision criteria can be cast based on Bayesian predictive probability of success for individual treatment arms rather than posterior probabilities of the treatment effects. A combination of continuous monitoring with decision criteria based on predictive probability for a multi-arm platform trial was explored by Hobbs, Chen, and Lee (2018), where the authors showed that such an approach could result in considerable efficiency gains. This approach merits investigation but it is beyond the scope of our current work.

In the expansion phase (Part 2), the most promising (“winner”) treatment arms from Part 1 will be further evaluated. Additional patients will be enrolled into each “winner” treatment arm to ensure sufficient sample size for statistical hypothesis testing concerning ORR of this treatment at the end of the study. There are several ways to determine the number of additional patients. One possibility is to expand each arm
to some fixed, prespecified number. Alternatively, for any arm promoted into Part 2, the sample size can be chosen adaptively, to ensure sufficient predictive power (probability of obtaining a statistically significant result based on the combined data from Part 1 and Part 2). In case of a statistically and clinically significant result for the ORR based on combined data, and provided that the safety profile is acceptable, a given treatment would be taken forward into a randomized confirmatory active-controlled phase III trial (outside of the current master protocol).

2.2. Design Details

A schematic of the design considered for the actual melanoma trial is displayed in Figure 1.

The open platform part starts with 3 treatment arms (TRT1, TRT2, and TRT3). During the trial, new combination treatments may be included into the master protocol. We require that a maximum of 10 treatment arms may ever be open at any stage of the trial.

2.2.1. Part 1 (Open Platform)

In Part 1, initial \( m = 30 \) patients will be randomized in a 1:1:1 ratio (10 subjects per arm are viewed as the minimal number for the purpose of declaring futility as per the decision rules presented later in this section). The first IA is planned after approximately 10 subjects in each treatment arm have been enrolled, treated, and followed-up for 20 weeks to perform efficacy evaluation (tumor response). Subsequent IAs will be conducted approximately 5 months thereafter. The maximum number of patients per arm in Part 1 is capped at \( n^{(1)}_{max} \). In our example we set \( n^{(1)}_{max} = 29 \). This is consistent with Simon’s optimal two-stage design with a Type I error rate of 0.05 (when the true ORR = 0.10) and power of 0.80 (when the true ORR = 0.30) (Simon 1989). The choice of the ORR = 0.30 for the alternative hypothesis will be discussed momentarily. With these parameters, Simon’s optimal two-stage design would initially accrue 10 patients and stop the arm for futility, if the observed ORR is 1/10 or less. Otherwise, 19 additional patients would be accrued for a total of 29, and the null hypothesis would be rejected if the final observed ORR for this arm is 6/29 or more.

To facilitate decision making in Part 1, the ORR is modeled using a Bayesian approach as follows. For the \( i \)th treatment arm, let \( p_i \) denote the true ORR, and let \( X_i \) denote the number of confirmed responses among \( m_i \) patients, such that \( X_i \sim \text{Binomial}(m_i, p_i) \). Assuming \( p_i \sim \text{Beta}(\alpha_i, \beta_i) \) (where \( \alpha_i > 0 \) and \( \beta_i > 0 \) are the prior parameters), the posterior \( p_i | (x_i, m_i) \sim \text{Beta}(\alpha_i + x_i, \beta_i + m_i - x_i) \). For instance, a practical choice is to set \( \alpha_i = \beta_i = 1 \) for each treatment.

Based on the input from an advisory board of key opinion leaders, and taking into consideration expected efficacy for the target population, the following clinical activity thresholds were elicited for the true ORR in any treatment group:

- “Winner”: ORR \( \geq 25\% \) (the value of 25\% is viewed as the minimally clinically relevant effect, whereas the value of 30\% is viewed as a clinically desirable effect which is used to plan for \( \geq 80\% \) power for hypothesis testing of this arm);
- “Futility”: ORR < 10\%;
- “Interesting activity”: ORR = 10–25\%.

For the \( i \)th treatment arm (assuming it is open at the time of an IA), the following decision rules were posited:

1. **Expand**: Declare the treatment a “winner” and promote it into Part 2, if \( Pr \left( p_i > \pi_E | \text{data} \right) > c_E \), where \( \pi_E \) and \( c_E \) are user-defined efficacy thresholds. In our study we set \( \pi_E = 0.20 \) (slightly less than 0.25, which is deemed to be a lower bound for an efficacious treatment) and \( c_E = 0.70 \) (a reasonably large value for a Bayesian posterior probability).

2. **Terminate**: Declare the treatment “futile” and stop enrollment into this arm, if \( Pr \left( p_i < \pi_F | \text{data} \right) > c_F \), where \( \pi_F \) and \( c_F \) are user-defined futility thresholds. In our study we set \( \pi_F = 0.15 \) (slightly greater than 0.10, the futility threshold) and \( c_F = 0.70 \).

3. **Continue**: Consider the treatment as “interesting” and continue enrollment into this arm, if neither efficacy nor
futility criteria are met, that is, if \( \Pr \left( p_i > \pi_E \mid \text{data} \right) \leq c_E \) and \( \Pr \left( p_i < \pi_F \mid \text{data} \right) \leq c_F \). Here, as in i) and ii) above, we set \( \pi_E = 0.20, \pi_F = 0.15, \) and \( c_E = c_F = 0.70 \).

In the described decision rule, the parameter values \( \pi_E, \pi_F, c_E, \) and \( c_F \) were chosen empirically through simulation, to ensure low, acceptable rate of decision errors to support the clinical activity thresholds that would allow a combination arm to enter the expansion phase (Expand rule), continue in the selection phase (Continue rule), or stop enrollment into the arm (Terminate rule). In general, the choice of the parameters \( \pi_E, \pi_F, c_E, \) and \( c_F \) will depend on the clinical trial context and the design calibration would have to be done by the statistician with input from the clinician. See Xie, Ji, and Tremmel (2012) and Jacob et al. (2016) for some good examples of how to perform such calibrations in practice.

2.2.2. Part 2 (Expansion)

In Part 2, the “winner” arms from Part 1 (if any) will be further evaluated. A particular “winner” arm may complete Part 1 with the number of subjects between 10 and \( n_{\text{max}} \), depending on the underlying ORR, observed data, and the number of patients enrolled in Part 1. A given arm is declared a “winner” and is promoted into Part 2 either if this arm reaches the “Expand” decision any time during Part 1, or if this arm reaches the “Continue” decision at the end of Part 1.

Alternatively, one may require that a given arm is expanded into Part 2 only if it reaches the “Expand” decision during Part 1 (i.e., the “Continue” decision at the end of Part 1 implies that the arm should not be further pursued). Such an option results in a more conservative design which has higher probability of stopping efficacious arms, but also has lower probability of expending truly efficacious treatment arms into Part 2, and hence, lower statistical power for testing these arms. We explored this option through simulation and found it to be too conservative for use in practice.

For each “winner” arm selected for Part 2, additional patients will be enrolled to ensure sufficient total sample size for statistical hypothesis testing at the end of the study. The main rationales for including extra subjects in Part 2 is to ensure control of the Type I error rate in the final analysis for a given treatment arm (i.e., without multiplicity adjustment), and also to have sufficient safety data. The primary objective for each “winner” arm in Part 2 is to test the hypotheses \( H_0 : \text{ORR} \leq 0.10 \) versus \( H_1 : \text{ORR} > 0.10 \). The true ORR \( \geq 0.25 \) (e.g., \( \text{ORR} = 0.30 \)) is viewed as warranting further investigation of a treatment in pivotal studies.

Data combined from Part 1 and Part 2 will be used for the purpose of hypothesis testing. Statistical significance will be declared if the lower bound of a 2-sided 95% confidence interval using Clopper–Pearson’s exact method for ORR is >0.10. Importantly, any arm that enters Part 2 is treated separately, as it would be treated if this arm was in its own standalone single-arm study. In other words, there is no in-between comparison if multiple arms enter the expansion phase.

It is also possible that at a particular IA during Part 1, two or more treatment arms meet the “winner” criteria. In the current article, we do not limit the number of “winner” arms that can be expanded into Part 2. However, in practice an investigator may wish that only one experimental treatment should actually be moved forward. In this case, some additional criteria to facilitate selection among “winner” arms would be required. There are several methodologies for ranking and selection of new therapies in randomized phase II trials; see, for instance, Liu, Moon, and LeBlanc (2012) for a review. These methods can be also used in open platform trial designs (such as ours), after some modification. For instance, suppose we have two “winner” treatments at a given IA during Part 1. One can require that a treatment arm with the highest posterior mean of the shrinkage-estimated ORR is taken into Part 2. Or alternatively, one can cast a selection rule using a posterior probability of one arm superior to the other given accrued data. It may turn out that both “winner” arms are very similar in terms of efficacy, in which case it may be prudent to include some additional criteria, such as safety, to aid the selection. These useful design modifications deserve further investigation, but they are beyond the scope of our current work.

There are two possibilities to determine the number of additional patients per winner arm in Part 2. A simple approach is to expand each arm to some fixed, prespecified number \( n_{\text{dep}} \) based on the desired Type I error/power consideration. Alternatively, for any arm promoted into Part 2, the sample size can be chosen adaptively, to provide sufficient Bayesian predictive power (probability of obtaining a statistically significant result based on the combined data from Part 1 and Part 2). The details of the latter approach are delineated below.

2.2.3. Adaptive Choice of the Sample Size for Part 2

At the time when a decision to expand a “winner” arm is made, Bayesian predictive power will be calculated for several choices of the sample size. For illustrative use in this article, we considered the values from 20 to 70 in increments of 5, that is, 20, 25, . . . , 70. The Part 2 sample size will be chosen such that the mean predictive power is at least 100\( \gamma \)% (where 0 < \( \gamma < 1 \) is a user-defined parameter), and Type I error is maintained at a one-sided 2.5% level. In our example we set \( \gamma = 0.7 \).

The calculation of predictive power can be based on a Bayesian posterior predictive distribution of the ORR of a “winner” arm. However, ORR estimates based on Part 1 data are potentially biased due to treatment selection in Part 1, and, therefore, the arm(s) to be expanded can potentially be selected based on a random-high ORR observed in Part 1. Therefore, the estimated sample size and power are potentially biased as well. To mitigate this problem, we used a shrinkage estimator of the ORR of a “winner” arm, which is derived based on a hierarchical model that takes into account the ORR from all available treatment arms. The use of the shrinkage estimator can be viewed as a conservative approach, since it takes into account Part 1 ORR results from all treatment arms at the time when the “winner” treatment is identified. The shrinkage estimator of ORR was determined based on a hierarchical model

\[
\log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \delta_i,
\]

where \( \beta_0 \) is the mean logit ORR among the treatment arms, and \( \delta_i \) is the random effect of the \( i \)th treatment. We assumed \( \beta_0 \sim \text{Normal} \) with mean \( -2.2 \) and variance 0.25, which is equivalent to the median ORR of \( -0.10 \). For the
random effect term, we assumed $\delta_i \sim \text{Normal}(0, \sigma^2)$, and $\sigma^2 \sim \text{IG}(\alpha = 2, \beta = 1, \text{upper} = 2)$ (the inverse gamma distribution with shape parameter = 2, scale parameter = 1, and truncation when the distribution reaches the value of 2). The hyper-prior for $\sigma^2$ was chosen such that it is conservative and allow shrinkage. Our chosen prior was calibrated such that the median for ORR is $\sim 0.1$, 90% prior interval is (0.02 and 0.35) and its upper 99th percentile is 0.54.

With these priors, the shrinkage estimator model is developed as follows: $p_i = \frac{\exp(\beta_0+\delta_i)}{1+\exp(\beta_0+\delta_i)}$; $r_i \mid (m_i, p_i) \sim \text{Binomial}(m_i, p_i)$. Thus, $r_i$ is the modeled number of responders for the $i$th treatment arm. The model parameters to be used in the shrinkage estimator were assessed using a Markov chain Monte Carlo (MCMC) method (PROC MCMC in SAS). A total of 2000 MCMC samples were used, with 1,000 burn-in samples, and afterward taking every 10th sample (to avoid highly correlated samples) until there are 2000 samples.

As soon as at least one “winner” arm is identified in Part 1, the calculation of $n_2$ (additional number of patients in Part 2) for this arm is done as follows:

1. Calculate the observed ORR for the “winner” arm, as well as for the other arms that are available at the time of the IA.
2. Calculate the shrinkage estimator for the “winner” arm (as well as for the other arms) based on 2000 iterations of the MCMC. Use the shrinkage estimator for the “winner” arm for the purpose of conditioning of the predictive power calculation.
3. The predictive power (PP) distribution is formed based on:
   - Distribution of the observed ORR, adjusted by the shrinkage estimator evaluated via MCMC.
   - From this distribution, given data from Part 1, calculate the number of responses needed in Part 2 to reject the null hypothesis for a predefined $\alpha$ at the final analysis.
   - For each sample of MCMC, the corresponding power is calculated. PP is then the average power of the MCMC samples.

For any “winner” arm, PP is calculated for several choices of the sample size for Part 2 (e.g., for the values from 20 to 70 in increments of 5, that is, 20, 25, …, 70), and $n_2$ is chosen to be the smallest number for which PP is at least 70%.

### 2.3. A Single Trial Illustration

To illustrate utility of the design described in Section 2.2, we present a simulation for a single trial. The trial starts with 3 arms (TRT1, TRT2, TRT3), and the fourth arm (TRT4) is introduced after IA2. Individual responses are generated using Bernoulli distributions with success probabilities 0.07, 0.10, 0.25, and 0.30 for TRT1, TRT2, TRT3, and TRT4, respectively. Given the clinical activity thresholds defined in Section 2.2, TRT3 and TRT4 are efficacious, whereas TRT1 and TRT2 are not. We assume an enrollment rate of 8 patients per month and ~5 months between each interim analysis.

Figure 2 shows accumulating trial data and the corresponding decisions.

The first 30 patients are randomized equally (10 per arm). At IA1, the observed treatment response rates are, respectively, 0, 2/10, and 2/10. The estimated posterior Pr(Efficacy) and Pr(Futility) are shown in Table 1. Based on these results, the following decisions are made: Terminate TRT1 for futility (since
Table 1. Interim decisions during Part 1 for a single trial run of the platform trial design (cf. Section 2.3).

| Interim analysis | n  | x  | ORR | Pr(Efficacy) | Pr(Futility) | Decision  |
|------------------|----|----|-----|--------------|--------------|-----------|
| TRT1             | 10 | 0  | 0   | 0.086        | 0.833        | Terminate |
| TRT2             | 10 | 2  | 0.2 | 0.617        | 0.221        | Continue  |
| TRT3             | 10 | 2  | 0.2 | 0.617        | 0.221        | Continue  |

Interim analysis 2

| n  | ORR | Pr(Efficacy) | Pr(Futility) | Decision |
|----|-----|--------------|--------------|----------|
| TRT1 | 10  | N/A          | N/A          | N/A      |
| TRT2 | 29  | 0.069        | 0.044        | 0.849    | Terminate |
| TRT3 | 29  | 0.241        | 0.761        | 0.070    | Expand    |

Interim analysis 3

| n  | ORR | Pr(Efficacy) | Pr(Futility) | Decision |
|----|-----|--------------|--------------|----------|
| TRT1 | 10  | N/A          | N/A          | N/A      |
| TRT2 | 29  | 0.069        | N/A          | N/A      |
| TRT3 | 29  | 0.241        | N/A          | N/A      |
| TRT4 | 10  | 0.4          | 0.950        | 0.016    | Expand    |

Note: Individual responses are generated using Bernoulli distributions with success probabilities 0.07, 0.10, 0.25, and 0.30 for TRT1, TRT2, TRT3, and TRT4, respectively. It is assumed that enrollment rate is 8 patients per month, and there are 5 months between each interim analysis.

After IA1, additional 38 patients are randomized in Part 1 (19 patients to TRT2, and 19 patients to TRT3). At IA2, the observed response rates for TRT2 and TRT3 are 2/29 = 0.069 and 7/29 = 0.241, respectively (Table 1). This calls for the “Terminate” decision for TRT2 (since Pr(Futility) = 0.849 > 0.70) and the “Expand” decision for TRT3 (since Pr(Efficacy) = 0.761 > 0.70).

For TRT3, the posterior distribution of a shrinkage estimator of ORR is displayed in Figure 3 (left plot); its mean is 0.196, which is lower than the observed response rate of 0.241 due to the shrinkage estimator incorporating data from the non-encircaus arms TRT1 and TRT2. The sample size for the expansion cohort for TRT3 is chosen based on the predictive power calculations; the smallest $n_2$ for which PP is at least 70% is $n_2 = 45$. It is projected that $R = 7$ additional responses in Part 2 would be needed to obtain a statistically significant result for this treatment.

After IA2, the study continues with TRT3 in Part 2, and TRT4 is introduced in Part 1. At IA3, the observed ORR for TRT4 is 4/10 = 0.4, which calls for the “Expand” decision for this treatment (since Pr(Efficacy) = 0.950 > 0.70). Data from additional $n_2 = 20$ patients on TRT4 would provide 70% predictive power, and $R = 4$ additional responses would be needed to obtain a statistically significant result in Part 2.

The results of the final analyses for the “winner” arms (TRT3 and TRT4) at the end of Part 2 are as follows (bottom part of Figure 2). For TRT3, we have 15 out of 74 responders (ORR = 0.203; 95% CI: 0.111–0.294). For TRT4, we have 11 out of 30 responders (ORR = 0.367; 95% CI: 0.194–0.539).

In summary, for the above considered example, the design has reached the following decisions:

- TRT1 (with true ORR = 0.07) was terminated for futility at IA1 based on data from 10 patients;
- TRT2 (with true ORR = 0.10) was terminated for futility at IA2 based on data from 29 patients;
- TRT3 (with true ORR = 0.25) was expanded into Part 2 at IA2, and the final result was significant with the total sample size of 74; and
- TRT4 (with true ORR = 0.30) was expanded into Part 2 at IA3, and the final result was significant with the total sample size of 30.

The grand total sample size of the trial in this example was $10 + 29 + 74 + 30 = 143$. Note that this number will not always be the same across simulations due to randomness in the data.

3. Design Operating Characteristics

3.1. Simulation Framework

Extensive simulation studies were performed to examine operating characteristics of the design described in Section 2.2. In Part 1, we assume that patients are enrolled in cohorts of size $n_2$ for which PP is at least 70% is $n_2 = 45$. It is projected that $R = 7$ additional responses in Part 2 would be needed to obtain a statistically significant result for this treatment.

After IA2, the study continues with TRT3 in Part 2, and TRT4 is introduced in Part 1. At IA3, the observed ORR for TRT4 is 4/10 = 0.4, which calls for the “Expand” decision for this treatment (since Pr(Efficacy) = 0.950 > 0.70). Data from additional $n_2 = 20$ patients on TRT4 would provide 70% predictive power, and $R = 4$ additional responses would be needed to obtain a statistically significant result in Part 2.

The results of the final analyses for the “winner” arms (TRT3 and TRT4) at the end of Part 2 are as follows (bottom part of Figure 2). For TRT3, we have 15 out of 74 responders (ORR = 0.203; 95% CI: 0.111–0.294). For TRT4, we have 11 out of 30 responders (ORR = 0.367; 95% CI: 0.194–0.539).

In summary, for the above considered example, the design has reached the following decisions:

- TRT1 (with true ORR = 0.07) was terminated for futility at IA1 based on data from 10 patients;
- TRT2 (with true ORR = 0.10) was terminated for futility at IA2 based on data from 29 patients;
- TRT3 (with true ORR = 0.25) was expanded into Part 2 at IA2, and the final result was significant with the total sample size of 74; and
- TRT4 (with true ORR = 0.30) was expanded into Part 2 at IA3, and the final result was significant with the total sample size of 30.

The grand total sample size of the trial in this example was $10 + 29 + 74 + 30 = 143$. Note that this number will not always be the same across simulations due to randomness in the data.
Table 2. Operating characteristics of the platform trial design (cf. Section 2.2): Decision probabilities (%) for the four treatment arms at different IAs during Part 1.

| Interim Analysis | Decision | Scenario A | Scenario B |
|------------------|----------|------------|------------|
|                   | TRT1     | TRT2       | TRT3       | TRT4       | TRT1     | TRT2       | TRT3       | TRT4       |
|                   | (0.07)   | (0.10)     | (0.25)     | (0.30)     | (0.07)   | (0.10)     | (0.15)     | (0.25)     |
| Interim Analysis 1| Expand   | 2.82       | 7.4        | 47.91      | N/A      | 6.9        | 18.65      | 47.91      |
|                   | Terminate| 85.04      | 73.47      | 24.02      | N/A      | 73.42      | 55.22      | 24.02      |
|                   | Continue | 12.14      | 19.13      | 28.07      | N/A      | 19.68      | 26.13      | 28.07      |
|                   |          |            |            |            |          |            |            |            |
| Interim Analysis 2| Expand   | 4.1        | 9.54       | 25.76      | N/A      | 9.34       | 18.03      | 24.85      |
|                   | Terminate| 7.3        | 8.21       | 1.0        | N/A      | 8.38       | 5.39       | 0.99       |
|                   | Continue | 0.74       | 1.38       | 1.31       | N/A      | 1.96       | 2.71       | 2.23       |
|                   |          |            |            |            |          |            |            |            |
| Interim Analysis 3| Expand   | 0.68       | 1.28       | 1.3        | 61.97    | N/A        | 1.75       | 2.59       |
|                   | Terminate| 0.06       | 0.1        | 0.01       | 14.29    | N/A        | 0.16       | 0.1        |
|                   | Continue | 0          | 0          | 23.74      | N/A      | 0.05       | 0.02       | 0.03       |
|                   |          |            |            |            |          |            |            |            |
| Interim Analysis 4| Expand   | N/A        | N/A        | N/A        | 23.48    | N/A        | 0.05       | 0.02       |
|                   | Terminate| N/A        | N/A        | N/A        | 0.26     | N/A        | 0          | 0          |
|                   | Continue | N/A        | N/A        | N/A        | 0        | N/A        | 0          | 0          |

Note: Probabilities are not cumulative. TRT4 is introduced into the study after IA2. In Scenario A, the true ORRs = 0.07, 0.10, 0.25, and 0.30. In Scenario B, the true ORRs = 0.10, 0.15, 0.25, and 0.35.

### 3.2. Simulation Results

Table 2 shows simulated decision probabilities for the four treatment arms at different IAs during Part 1. In both scenarios A and B, the chances for making correct decisions early on (i.e., based on data from 10 to 20 patients per arm) are reasonably high. For instance, in scenario A for TRT1 with true ORR = 0.07, the probability of terminating this arm for futility at IA1 is 85.04%, and a similar quantity at either IA1 or IA2 is ∼92% (85.04 + 7.3). For TRT4 with true ORR = 0.30 (which is introduced into the study after IA2), the correct decision of promoting this arm into Part 2 is made with probability 61.97% at IA3, and with probability ∼85% (61.97 + 23.48) at either IA3 or IA4. In scenario B, the most efficacious TRT4 with ORR = 0.35 is promoted to Part 2 with probability ∼74% at IA3, and with probability ∼91% (73.89 + 17.53) at either IA3 or IA4.

Table 3 displays additional operating characteristics during Part 1 (overall decision probabilities for the treatment arms and treatment sample sizes for Part 1), and the operating characteristics for Part 2 (treatment sample sizes for Part 2 and total, and power of the final analysis). As regards Part 1, treatment arms with a given value of the true ORR exhibit consistent results across the experimental scenarios, regardless of the other treatments within the same scenario. For instance, for a treatment arm with true ORR = 0.10 (TRT2 in scenario A and TRT1 in scenario B), there is ∼18% chance of being expanded into Part 2 and ∼82% chance of being terminated for futility in Part 1. This is achieved with the average sample size of 14 patients. Likewise, for a treatment with true ORR = 0.25 (TRT3 in both scenarios A and B), the overall probabilities of “Expand” and “Terminate” are ∼75% and ∼25%, respectively, with the average sample size of 15. Overall, the decision probabilities and the sample size in Part 1 are reflective of the underlying treatment efficacy: very poor (or very good) treatments are identified with high probability and small sample size.

As regards Part 2, it is important to consider both unconditional and conditional characteristics in Table 3. The unconditional characteristics were derived across all 10,000 simulation runs. For instance, unconditional power \( P_{1-2} \) is the probability that a treatment arm is promoted into Part 2 and
subsequently demonstrates a statistically significant result based on the pooled Part 1 and Part 2 data for this arm. The conditional results were derived across those simulation runs for which a decision to expand the arm into Part 2 was made. In this case, \( P_{1 \rightarrow 2} \) is the conditional probability that a treatment yields a statistically significant result given that it has passed Part 1. The number of simulation runs for which a given arm is expanded into Part 2 can be obtained by multiplying the \( \text{Pr(Expand)} \) of the final analysis using pooled data from Parts 1 and 2. In Scenario A, the true ORRs = 0.07, 0.10, 0.25, and 0.30. In Scenario B, the true ORRs = 0.10, 0.15, 0.25, and 0.35. SD = standard deviation; IQR = interquartile range [Q25–Q75].

### Table 3. Operating characteristics of the platform trial design (see Section 2.2): Part 1 decision probabilities and sample sizes for the four treatment arms, and Part 2 characteristics.

| Scenario A | Scenario B |
|------------|------------|
| **Part 1 characteristics** | **Part 1 characteristics** |
| Decision probabilities (%) |  |
| \( \text{Pr(Expand)} \) | 7.6 | 18.04 |
| \( \text{Pr(Terminate)} \) | 92.4 | 81.96 |
| **Part 1 sample size \( n_1 \)** |  |
| Mean (SD) [IQR] | 12 (6) | 14 (7) | 15 (8) | 15 (8) | 14 (7) | 15 (8) | 15 (8) | 13 (7) |
| **Part 2 sample size \( n_2 \)** |  |
| Mean (SD) [IQR] | 5 (18) | 12 (26) | 37 (29) | 36 (26) | 12 (26) | 24 (32) | 36 (29) | 34 (24) |
| Total sample size \( n_{tot} \) |  |
| Mean (SD) [IQR] | 17 (22) | 26 (30) | 52 (32) | 51 (27) | 25 (30) | 39 (37) | 52 (32) | 47 (24) |
| Power \( (P_{1 \rightarrow 2}) \) (%) | <0.1 | 0.8 | 69.7 | 83.6 | 0.8 | 13.1 | 69.4 | 90.4 |

| Scenario A | Scenario B |
|------------|------------|
| **Part 2 - unconditional characteristics** | **Part 2 - unconditional characteristics** |
| **Part 2 sample size \( n_2 \)** |  |
| Mean (SD) [IQR] | 68 (9) | 66 (13) | 49 (22) | 42 (23) | 65 (14) | 61 (18) | 49 (23) | 37 (22) |
| Total sample size \( n_{tot} \) |  |
| Mean (SD) [IQR] | 90 (15) | 87 (17) | 66 (24) | 56 (24) | 87 (18) | 80 (22) | 65 (24) | 51 (22) |
| Power \( (P_{1 \rightarrow 2}) \) (%) | 0.4 | 4.4 | 93.0 | 97.4 | 4.3 | 33.3 | 92.5 | 98.9 |

Note: Part 1 characteristics: Decision probabilities (%) for the four arms, and Sample size for Part 1 \( n_1 \)—mean (SD) [IQR]. Part 2 characteristics (Unconditional—derived across 10,000 simulation runs, and Conditional—derived across simulation runs for which a decision to expand a treatment into Part 2 is made) for the four arms: Sample size for Part 2 \( n_2 \)—mean (SD) [IQR]; Total sample size \( n_{tot} \)—mean (SD) [IQR]; Power \( (P_{1 \rightarrow 2}) \) (%) of the final analysis using pooled data from Parts 1 and 2. In Scenario A, the true ORRs = 0.07, 0.10, 0.25, and 0.30. In Scenario B, the true ORRs = 0.10, 0.15, 0.25, and 0.35.

### Table 4. Designs compared in Section 4.

| Design | Part 1 | Part 2 (only for the “winner” arms from Part 1) |
|--------|--------|-----------------------------------------------|
| 1      | Platforma | Adaptive sample size, to ensure \( \geq 70\% \) predictive power in the final analysis |
| 2      | Platformb | Expand to a fixed \( n_{tot} = 59\) |
| 3      | Simon’s 2-stage for each armb | Expand to a fixed \( n_{tot} = 59\) |
| 4      | \( n_1 = 17 \) patients for each arm, with futility analysis | Expand to a fixed \( n_{tot} = 59\) |

a Bayesian decision rules at IAs; maximum number of patients per arm is capped at \( n_{max}^{(1)} = 29\).
bFutility analysis after first 10 patients; if not futile, enrol 19 more patients and perform an additional analysis based on data from 29 patients.
c \( n_{tot} = 59 \) comes from Simon’s optimal two-stage design with Type I error of 0.01 \( (H_0 : \text{ORR}=0.10) \) and power of 0.9 \( (H_1 : \text{ORR}=0.30) \).
sample size, which is derived based on all simulation runs. These numbers are quite variable, because if a treatment is not expanded into Part 2, then its sample size for Part 2 is zero. As can be seen from Table 3, the unconditional average total sample size per arm ranges from 17 for ORR = 0.07 to about 47–52 for treatments with ORR ≥ 0.25. One should also be mindful of variability in the unconditional total sample size, as captured by the IQR.

4. A Comparison of Different Designs

We conducted additional simulation studies to compare three variants of our OPTIM-ARTS design with a more standard design. Due to a platform nature of the trial, it is difficult to find a single design that would serve as a “reference” in this setting. Since the study goal is to formally test efficacy of selected promising treatments, we use Simon’s optimal two-stage design (Simon 1989) for calibrating the total sample size per treatment arm. Assuming \( H_0 : \text{ORR} = 0.10 \) and \( H_1 : \text{ORR} = 0.30 \), the Type I error rate of 0.01 and the power of 0.9, Simon’s optimal two-stage design initially evaluates 17 patients, and if the observed ORR is 2/17 or fewer, the arm is stopped for futility; otherwise additional 42 patients are evaluated, and in the final analysis \( H_0 \) is rejected if the observed ORR is 12/59 or more. In our trial context, such a design can be viewed as one that has the fixed sample sizes for both Part 1 (context, such a design can be viewed as one that has the fixed sample sizes for both Part 1 (\( n_1 = 17 \)) and Part 2 (\( n_2 = 42 \)), such that \( n_{\text{tot}} = 17 + 42 = 59 \) for any arm.

Table 4 summarizes the designs to be compared. Designs \( \mathcal{D}_1 \) and \( \mathcal{D}_2 \) use the same building elements for Part 1: Bayesian interim decision rules as described in Section 2.2.1, with the maximum number of patients per arm capped at \( n_{\text{max}}^{(1)} = 29 \). For Part 2, \( \mathcal{D}_1 \) uses adaptive sample size calculation to ensure ≥70 predictive power in the final analysis for a given arm, whereas \( \mathcal{D}_2 \) simply expands the arm to a fixed \( n_{\text{tot}} = 59 \). Design \( \mathcal{D}_3 \) for each arm in Part 1 uses Simon’s optimal two-stage design with one IA for futility after 10 patients and a total sample size of 29. This corresponds to a Type I error of 0.05 (under \( H_0 : \text{ORR} = 0.10 \)) and power of 0.80 (under \( H_1 : \text{ORR} = 0.30 \)). For Part 2, \( \mathcal{D}_3 \) expands the arm (provided \( H_0 \) is rejected based on data from 29 patients in Part 1) to a fixed \( n_{\text{tot}} = 59 \). Design \( \mathcal{D}_4 \) for each arm uses Simon’s optimal two-stage design with one IA for futility based on 17 patients in Part 1, and a final analysis with \( n_{\text{tot}} = 59 \).

Figures 4 and 5 show the results for the four designs under scenarios A and B, respectively. Similar outputs were generated for six other scenarios, and they are available in the Supplemental Appendix. In Figures 4 and 5, displayed for each treatment arm are the power of the final test based on combined data from Parts 1 and 2 (left-hand plots) and the average total sample size (ATSS), right-hand plots.

The following important observations can be made for both scenarios A and B:

- Among the four designs, \( \mathcal{D}_4 \) has highest power and, in most cases, highest ATSS per arm, whereas \( \mathcal{D}_3 \) has lowest power and lowest ATSS per arm.
- For futile treatments with ORR ≤ 0.10 (TRT1 and TRT2 in scenario A and TRT1 in scenario B), all four designs have power (Type I error rate) ≤ 1%, but different values of ATSS; for example, for a treatment with ORR=0.10, the ATSS is 25–26 for \( \mathcal{D}_1 \), 20 for \( \mathcal{D}_2 \), 16 for \( \mathcal{D}_3 \), and 27 for \( \mathcal{D}_4 \).
- For ORR = 0.15 (TRT2 in scenario B), \( \mathcal{D}_1 \) has power = 13.1% and ATSS = 39, \( \mathcal{D}_4 \) has power = 13.7% and ATSS = 37, whereas \( \mathcal{D}_2 \) and \( \mathcal{D}_3 \) have both lower power (11.4% and 9.4%, respectively) and lower ATSS (30 and 25, respectively).
- For ORR = 0.25 (TRT3 in both scenarios A and B), \( \mathcal{D}_1 \) and \( \mathcal{D}_4 \) have the same ATSS = 52, but different power: 69–70% for \( \mathcal{D}_1 \), and 74.3% for \( \mathcal{D}_4 \). At the same time, \( \mathcal{D}_2 \) and \( \mathcal{D}_3 \) have both lower power (66.9% and 60.5%, respectively) and lower ATSS (47 and 44, respectively).
- For ORR = 0.30 (TRT4 in scenario A), \( \mathcal{D}_1 \), \( \mathcal{D}_2 \), and \( \mathcal{D}_3 \) have about the same ATSS (50–52), but \( \mathcal{D}_1 \) and \( \mathcal{D}_2 \) have somewhat higher power (~83%) than \( \mathcal{D}_3 \) (79%).
- For the most successful treatment with ORR = 0.35 (TRT4 in scenario B), \( \mathcal{D}_1 \), \( \mathcal{D}_2 \), and \( \mathcal{D}_3 \) have ~90% power, but quite different values of ATSS: 47 for \( \mathcal{D}_1 \), 55 for \( \mathcal{D}_2 \), and 54 for \( \mathcal{D}_3 \). At the same time, \( \mathcal{D}_4 \) has ~97% power and the largest ATSS=58.

![Figure 4](image-url)  
*Figure 4.* Power of the final analysis (left-hand plot) and average total sample size (right-hand plot) for the four designs considered in Section 4, for Scenario A with true ORRs = 0.07, 0.10, 0.25, and 0.30.
Overall, no design seems to be “uniformly best” in terms of cost-efficiency (ATSS/power tradeoff). Higher power naturally comes at the expense of a larger ATSS. Design D1 is most advantageous for successful treatment arms (e.g., with ORR = 0.30 and 0.35). Note that D1 and D2 examine accumulating trial data and apply decision rules more frequently than D3 or D4. This added flexibility may be deemed very important from a clinician’s perspective in the open platform trial context.

5. Discussion

The significance of this article is that it provides a framework for open platform randomized phase II trials where Part 1 is cast for exploration of investigational compounds, and Part 2 is used to formally test efficacy of selected compounds. The framework is very flexible and can be used to construct various designs by different combinations of building blocks (e.g., data monitoring/decision rules, randomization procedures, treatment selection rules, sample size reassessment after Part 1, etc.) Actually, even some existing designs still fall under the described framework, including Simon’s optimal two-stage design with one futility analysis and fixed predetermined sample sizes for Parts 1 and 2.

The OPTIM-ARTS design studied in detail in Section 3 (using Bayesian monitoring in Part 1, shrinkage estimation for treatment selection, and sample size reassessment using Bayesian predictive probability) is an example that was implemented in the actual melanoma trial. We described our first-hand experience/thinking process that actually took place at the study planning stage.

We investigated utility of our design through simulations. We found that our design performs very sensibly—in the platform phase it has high probabilities of terminating futile arms and selecting promising efficacious arms, and in the expansion phase it adaptively adds more patients to the “winner” arms to ensure sufficiently high statistical power while protecting the Type I error rate. For each treatment arm, we reported two sets of operating characteristics: unconditional (regardless of whether the arm is taken forward into the expansion phase), and conditional on that the arm is declared a “winner” during the platform phase and is taken further into the expansion phase. These results should help investigators to better understand the statistical properties of our design. One important feature of the proposed methodology is its flexibility—the design can be tailored to a given trial by fine-tuning its parameters for desirable operating characteristics. We explored three variants of the OPTIM-ARTS design by considering different building elements for Parts 1 and Part 2 and also compared the resulting designs with a standard (Simon’s optimal two-stage) design. In practice, this would be an important step for evaluating design options, and it will require close collaborative efforts between the statistician, the clinician, and other relevant study team stakeholders.

We would like to highlight several additional items that merit further investigation. The performance of a Bayesian design may be sensitive to the choice of prior. As the base case, we assumed Beta(1,1) prior for the ORR of any treatment arm at the beginning of Part 1. However, an investigator may wish to use a more diffuse prior reflecting greater uncertainty in response rates, or a prior that is centered around the value of ORR = 0.10, to be more aligned with the null hypothesis. We ran additional simulations of design D1 (cf. Section 4) under scenario A with true ORRs = 0.07, 0.10, 0.25, and 0.30 using the following priors: Beta(1,1) (base case), Beta(1/2, 1/2) (Jeffreys’ prior); Beta(1/3, 1/3) (neutral noninformative prior (Kerman 2011) and Beta(0.1, 0.9) (a prior consistent with the null hypothesis). The results (not shown here) were, overall, similar for the first three priors. However, with Beta(0.1, 0.9), futile arms (TRT1 with ORR = 0.07 and TRT2 with ORR = 0.10) were terminated during Part 1 with higher probability and lower sample size than for the other priors, whereas efficacious arms (TRT3 with ORR = 0.25 and TRT4 with ORR = 0.30) had lower Pr(Expand) and hence, lower unconditional power—compared to the other priors. Based on these simulations, Beta(1,1) seems to be a very reasonable choice for practical purposes.

In our study, we assumed that individual outcomes are generated assuming constant true ORR in each treatment group. In
real clinical research setting, trial patients are likely to vary in terms of known and unknown covariates, and the assumption of constant ORR (even within a given treatment group) may not be feasible. Additional simulations assuming some kind of heterogeneity in the ORR are warranted. This may also call for some more elaborate Bayesian modeling of ORR during Part 1 (instead of a simple beta-binomial model), to borrow information across different treatment arms.

The choice of a randomization procedure is another important consideration. In our study we used equal randomization during Part 1. Alternatively, an investigator may wish to use response-adaptive randomization (RAR) to skew allocation toward empirically better treatment(s) such that, on average, a greater number of study patients are assigned to more efficacious treatment arms, while controlling the Type I error rate and maintaining power. Implementation of RAR in clinical trials requires careful considerations. Several authors highlighted limitations of Bayesian RAR in two-arm trials (Korn and Freidlin 2011; Thall, Fox, and Wathen 2015). For instance, treatment group imbalances induced by RAR in a two-arm comparative setting may result in loss in power and biased estimation of a treatment difference (Thall, Fox, and Wathen 2015). For multi-arm trials when all treatment arms are available at the enrollment of the first patient, RAR can be advantageous over equal randomization under some circumstances (Trippa et al. 2012; Wason and Trippa 2014). For instance, if an allocation ratio for a control arm is fixed and RAR is applied among the experimental arms to skew allocation to the empirically best treatment (when it exists), then the design can have higher statistical power for a comparison of the best experimental treatment versus control. On the other hand, in a multi-arm trial without a control group, RAR may be of limited value, especially in situations when differences between treatment success probabilities are small (Wathen and Thall 2017). In multi-arm controlled platform trials, where experimental arms can be added/dropped adaptively during the course of the study, utility of ORR has been less well researched. Some authors have shown that in this case utilizing RAR to target information balance by randomizing more study participants to novel emerging experimental therapies in the presence of exchangeable information from the control groups at different stages of the platform trial can enhance design performance from both ethical and statistical perspectives (Kaizer, Hobbs, and Koopmeiners 2018; Normington et al. 2020). However, in multi-arm platform trials without a control group (as in the present article), the use of RAR instead of equal randomization warrants further investigation and we defer this task to the future work.

Finally, we would like to note that the OPTIM-ARTS design was developed for binary outcome trials, which are ubiquitous in phase II cancer settings. In principle, the methodology can be extended to more elaborate settings of joint efficacy–toxicity outcomes or time-to-event outcomes (e.g., progression-free survival) for diseases where events are expected to occur relatively quickly. Other possible useful extensions include incorporation of a control arm, modification of treatment selection rules during the selection phase, and introduction of a formal comparison of selected experimental treatment(s) versus control once the expansion phase is completed. We hope to pursue these important problems in our future work.

Supplementary Materials

Supplemental Appendix contains simulation results for six additional experimental scenarios of treatment objective response rates.

Acknowledgments

The authors would like to thank an associate editor and two anonymous referees for their constructive comments on the original version of the article. Special thanks are also to our Novartis colleagues: Frank Bretz, David James, and Brian P. Smith for their thorough reviews and feedback that led to a substantially improved version of the article.

References

Barker, A. D., Sigman, C. C., Kellof, G. J., Hylton, N. M., Berry, D. A., and Esserman, L. (2009), “I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy,” Clinical Pharmacology & Therapeutics, 86, 97–100. [283]

Berry, S. M., Connor, J. T., and Lewis, R. J. (2015), “The Platform Trial: An Efficient Strategy for Evaluating Multiple Treatments,” JAMA, 313, 1619–1620. [283]

Berry, S. M., Petzold, E. A., Dull, P., Thielman, N. M., Cunningham, C. K., Corey, G. R., McClain, M. T., Hoover, D. L., Russell, J., McLeod Griffiss, J., and Woods, C. W. (2016), “A Response-Adaptive Randomization Platform Trial for Efficient Evaluation of Ebola Virus Treatments: A Model for Pandemic Response,” Clinical Trials, 13, 22–30. [284]

FDA (2018), “Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics,” Guidance for Industry (Draft Guidance), available at https://www.fda.gov/media/120721/download. [283]

Ferrarotto, R., Redman, M. W., Gandara, D. R., Herbst, R. S., and Papadimitrakopoulos, V. A. (2015), “Lung-MAP—Framework, Overview, and Design Principles,” Chinese Clinical Oncology, 4, 36. [283]

Hobbs, B. P., Chen, N., and Lee, J. J. (2018), “Controlled Multi-Arm Platform Design Using Predictive Probability,” Statistical Methods in Medical Research, 27, 65–78. [283,284]

Jacob, L., Uvarova, M., Boulet, S., Begai, I., and Chevret, S. (2016), “Evaluation of a Multi-Arm Multi-Stage Bayesian Design for Phase II Drug Selection Trials—An Example in Hemato-Oncology,” BMC Medical Research Methodology, 16, 67. [284,286]

Kaizer, A. M., Hobbs, B. P., and Koopmeiners, J. S. (2018), “A Multi-Source Adaptive Platform Design for Testing Sequential Combinatorial Therapeutic Strategies,” Biometrics, 74, 1082–1094. [284,293]

Kerman, J. (2011), “Neutral Noninformative and Informative Conjugate Beta and Gamma Prior Distributions,” Electronic Journal of Statistics, 5, 1450–1470. [292]

Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein, G. R., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus, J., Gupta, S., and Alden, C. M. (2011), “The BATTLE Trial: Personalizing Therapy for Lung Cancer,” Cancer Discovery, 1, 44–53. [283]

Korn, E. L., and Freidlin, B. (2011), “Outcome-Adaptive Randomization: Is It Useful?,” Journal of Clinical Oncology, 29, 771–776. [293]

Liu, Y.Y., Moon, J., and LeBlanc, M. (2012), “Phase II Selection Designs,” in Handbook of Statistics in Clinical Oncology (3rd ed.), eds. J. Crowley and A. Hoering, Boca Raton, FL: Chapman & Hall/CRC Press, pp. 151–161. [286]

National Cancer Institute (2017), “Common Toxicity Criteria,” available at https://ctep.cancer.gov/protocolDevelopment/ [284]

Normington, J., Zhu, J., Mattiello, F., Sarkar, S., and Carlin, B. (2020), “An Efficient Bayesian Platform Trial Design for Borrowing Adaptively From Historical Control Data in Lymphoma,” Contemporary Clinical Trials, 89, 105890. [293]

Park, J. W., Liu, M. C., Yee, D., Yau, C., van’t Veer, L. J., Symmans, W. F., Paoloni, M., Perlmutter, J., Hylton, N. M., Hogarth, M., and DeMichele, A. (2016), “Adaptive Randomization of Neratinib in Early Breast Cancer,” The New England Journal of Medicine, 375, 11–22. [283]
Rugo, H. S., Olopade, O. I., DeMichele, A., Yau, C., van’t Veer, L. J., Buxton, M. B., Hogarth, M., Hylton, N. M., Paolini, M., Perlmutter, J., and Symmans, W. F. (2016), "Adaptive Randomization of Veliparib-Carboplatin Treatment in Breast Cancer," *The New England Journal of Medicine*, 375, 23–34. [283]

Saville, B. R., and Berry, S. M. (2016), "Efficiencies of Platform Clinical Trials: A Vision of the Future," *Clinical Trials*, 13, 358–366. [283]

Simon, R. (1989), "Optimal Two-Stage Designs for Phase II Clinical Trials," *Controlled Clinical Trials*, 10, 1–10. [284, 285, 291]

Steuer, C. E., Papadimitrakopoulou, V., Herbst, R. S., Redman, M. W., Hirsch, F. R., Mack, P. C., Ramalingam, S. S., and Gandara, D. R. (2015), "Innovative Clinical Trials: The LUNG-MAP Study," *Clinical Pharmacology & Therapeutics*, 97, 488–491. [283]

Sudhop, T., Brun, N. C., Riedel, C., Rosso, A., Broich, K., and Senderovitz, T. (2019), "Master Protocols in Clinical Trials: A Universal Swiss Army Knife?", *Lancet Oncology*, 20, e336–e342. [283]

Sverdlov, O., and Rosenberger, W. F. (2013), "On Recent Advances in Optimal Allocation Designs in Clinical Trials," *Journal of Statistical Theory and Practice*, 7, 753–773. [284]

Tang, R., Shen, J., and Yuan, Y. (2019), "ComPAS: A Bayesian Drug Combination Platform Trial Design With Adaptive Shrinkage," *Statistics in Medicine*, 38, 1120–1134. [284]

Thall, P. F., Fox, P., and Wathen, J. K. (2015), "Statistical Controversies in Clinical Research: Scientific and Ethical Problems With Adaptive Randomization in Comparative Clinical Trials," *Annals of Oncology*, 26, 1621–1628. [293]

Trippa, L., Lee, E. Q., Wen, P. Y., Batchelor, T. T., Cloughesy, T., Parmigiani, G., and Alexander, B. M. (2012), "Bayesian Adaptive Randomized Trial Design for Patients With Recurrent Glioblastoma," *Journal of Clinical Oncology*, 30, 3258–3263. [293]

Ventz, S., Cellamare, M., Parmigiani, G., and Trippa, L. (2018), "Adding Experimental Arms to Platform Clinical Trials: Randomization Procedures and Interim Analyses," *Biostatistics*, 19, 199–215. [284]

Wason, J. M. S., and Trippa, L. (2014), "A Comparison of Bayesian Adaptive Randomization and Multi-Stage Designs for Multi-Arm Clinical Trials," *Statistics in Medicine*, 33, 2206–2221. [293]

Wathen, J. K., and Thall, P. F. (2017), "A Simulation Study of Outcome Adaptive Randomization in Multi-Arm Clinical Trials," *Clinical Trials*, 14, 432–440. [293]

Woodcock, J., and LaVange, L. (2017), "Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both," *The New England Journal of Medicine*, 377, 62–70. [283]

Xie, F., Ji, Y., and Tremmel, L. (2012), "A Bayesian Adaptive Design for Multi-Dose, Randomized, Placebo-Controlled Phase I/II Trials," *Contemporary Clinical Trials*, 33, 739–748. [286]

Yuan, Y., Guo, B., Munsell, M., Lu, K., and Jazaeri, A. (2016), "MIDAS: A Practical Bayesian Design for Platform Trials With Molecularly Targeted Agents," *Statistics in Medicine*, 35, 3892–3906. [284]

Zhou, X., Liu, S., Kim, E. S., Herbst, R. S., and Lee, J. J. (2008), "Bayesian Adaptive Design for Targeted Therapy Development in Lung Cancer—A Step Towards Personalized Medicine," *Clinical Trials*, 5, 181–193. [283]