Therapeutic Innovation & Regulatory Science (2023) 57:823–838
https://doi.org/10.1007/s43441-023-00500-w

DIA

REVIEW

Evolution of Phase II Oncology Trial Design: from Single Arm to Master Protocol

Ziji Yu[1,2] · Liwen Wu[2] · Veronica Bunn[2] · Qing Li[3] · Jianchang Lin[2]

Received: 30 August 2022 / Accepted: 10 February 2023 / Published online: 4 March 2023

Abstract

The recent development of novel anticancer treatments with diverse mechanisms of action has accelerated the detection of treatment candidates tremendously. The rapidly changing drug development landscapes and the high failure rates in Phase III trials both underscore the importance of more efficient and robust phase II designs. The goals of phase II oncology studies are to explore the preliminary efficacy and toxicity of the investigational product and to inform future drug development strategies such as go/no-go decisions for phase III development, or dose/indication selection. These complex purposes of phase II oncology designs call for efficient, flexible, and easy-to-implement clinical trial designs. Therefore, innovative adaptive study designs with the potential of improving the efficiency of the study, protecting patients, and improving the quality of information gained from trials have been commonly used in Phase II oncology studies. Although the value of adaptive clinical trial methods in early phase drug development is generally well accepted, there is no comprehensive review and guidance on adaptive design methods and their best practice for phase II oncology trials. In this paper, we review the recent development and evolution of phase II oncology design, including frequentist multistage design, Bayesian continuous monitoring, master protocol design, and innovative design methods for randomized phase II studies. The practical considerations and the implementation of these complex design methods are also discussed.

Keywords Adaptive design · Phase II oncology trials · Simon's two-stage design · Bayesian continuous monitoring · Master protocol · Dose optimization

Introduction

The primary goals of phase II oncology trials are to establish the anti-cancer activity of the investigational agent and to recommend its further clinical trial evaluation strategy. Although inherently comparative, phase II trials are usually open-label, single treatment arm designs comparing the investigational agent with historical data using short-term binary efficacy endpoints such as ORR as the primary efficacy endpoint. It is desired to screen out ineffective investigational drugs with a minimal number of patients being exposed. On one hand, substantial numbers of new drug candidates and combination therapies bring tremendous opportunities for sponsors to conduct proof-of-concept (POC) trials. On the other hand, the high competition also requires early and fast decisions based on the results of phase II clinical studies [1–3]. Therefore, the competitive landscape of oncology drug development brings unique challenges to study design, such as combination treatment strategies, multiple endpoints, multiple objectives in a single master protocol, and ongoing study adaptations in phase II trials [4]. An adaptive design, defined as a clinical trial design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in the trial, is more flexible and efficient than conventional designs [4], and therefore is becoming more common in Phase II oncology studies. Frequentist adaptive designs provide flexibility in terms of sample size, study duration, early futility stopping rules, and target product profile. Bayesian adaptive design in phase II trials allows continuous monitoring of the efficacy and safety results and helps the study team to make a timely decision. Master protocol design which allows multiple patient populations and drug regimens to be compared in the same study is also critical to increasing

✉ Ziji Yu
ziji.yu@takeda.com

[1] 95 Hayden Ave, Lexington, MA 02421, USA

[2] Takeda Pharmaceuticals, Lexington, USA

[3] Morphosis AG, Boston, USA

the efficiency of phase II oncology trial design [5–7]. Since the primary purpose of phase II is not seeking regulatory approval, phase II exploratory trials do not generally have the same regulatory expectations as confirmatory trials intended to provide substantial evidence of effectiveness in terms of statistical rigor and operating characteristics. However, it is still important to follow the good principles of adaptive trial design to avoid erroneous conclusions [4]. Although many adaptive design methods have been proposed, there is no comprehensive review and guidance on adaptive design methods and their best practice in phase II oncology trials. This paper intends to bridge this gap.

The structure of the paper is summarized as follows: In Sect. "Two-stage Adaptive Design Methods", we will review the frequentist adaptive design methods; in Sect. "Bayesian Adaptive Design", Bayesian adaptive design and its implementation in phase II oncology studies will be discussed; in Sect. "Master Protocol", we will review the master protocol design and its application in early phase oncology trials, as well as the advanced methods proposed to overcome statistical challenges in master protocol design, including information borrowing, adaptive randomization, and multiplicity adjustment; In Sect. "Randomized Phase II Oncology Trials", we will discuss the importance of randomized Phase II studies, including randomized dose optimization studies and randomized control proof-of-concept studies, and the application of adaptive design in such studies.

Two-Stage Adaptive Design Methods

In phase II oncology trials, it is desirable to reject an inefficacious treatment with a minimal number of patients exposed to the investigational drug. The majority of such trials are open-label single-arm studies with short-term binary efficacy endpoints based on the following general hypothesis testing:

$$H_0 : p \le p_0 \text{vs} H_0 : p \ge p_1$$

Where p_0 is the maximum 'unacceptable' response rate, and p_1 is the minimum 'acceptable' response rate. Due to the ethical and economic considerations, an adaptive two-stage design, which early terminates the study based on the unpromising interim analysis result, is often preferable, given the benefits of sample size savings and patient protection.

Simon's Two-Stage Design

The most commonly used adaptive two-stage design is Simon's two-stage design [8]. A Simon two-stage design has the following form:

- In the first stage, N_1 patients are accrued, treated, and observed for clinical response. If C_{R1} or fewer responses are observed, the trial is terminated and the treatment is not recommended for further investigation;
- In the second stage, additional $(N_2 - N_1)$ patients are accrued if this study is not stopped after the first stage. In the final analysis, if C_{R2} or fewer responses are observed among all N_2 patients, then the treatment is not recommended for further investigation; if more than C_{R2} responses are observed, and treatment is recommended.

The design parameters $\boldsymbol{Q} = (N_1, N_2, C_{R1}, C_{R2})$ will be selected to satisfy the following error constraints:

$$\Pr\left(RecommendTreatment | p = p_0\right) \le \alpha,$$

$$\Pr\left(RecommendTreatment | p = p_1\right) \ge 1 - \beta,$$

and one of the following 'optimal' criteria:

- Simon's 'Optimal' design: to minimize the expected sample size when $p = p_0$,
- Simon's 'Minimax' design: to minimize the maximum sample size N_2.

The design parameters $(N_1, N_2, C_{R1}, C_{R2})$ will be determined by enumeration using exact binomial probabilities.

Simon's optimal design is the most commonly used adaptive design method for phase II oncology studies. According to a survey conducted by Ivanova et al. [2], more than 40% of the phase II oncology trials with results published in leading oncology journals between 2010 and 2015 used this design. When the difference in expected sample sizes is small between the 'Optimal' design and 'Minimax' design and the enrollment is slow, Simon's 'minimax' design may be more attractive because it reduce the maximum sample size at stage 2. The optimization criterion is not unique and the design strategy should be determined based on study specific-assumptions.

Adaptive Two-Stage Design with Flexible Stage II Sample Size

In the traditional Simon's two-stage design, the second stage sample size is fixed regardless of the number of responses from the first stage as long as it is over the early stopping threshold. It is counterintuitive because if the overwhelming efficacy of the investigational drug is observed from stage 1, the trial may not need to enroll as many sample sizes as planned for stage 2. The adaptive two-stage design, proposed by Lin and Shih [9], is an extension of Simon's two-stage design, which defines the alternative hypothesis based on the interim readout and adapts the sample size and

decision rules for stage 2. Specifically, if the interim readout is overwhelmingly positive, the testing strategy will be more aggressive at stage 2 – the study will be powered at a higher target response rate which requires a smaller sample size. On the other hand, if the interim data is not overwhelmingly positive but still deemed promising, at stage 2 we will test for a lower target response rate which requires a larger sample size.

The scheme of the adaptive two-stage design is illustrated in Fig. 1. At stage 1, N_1 patients are enrolled and evaluated. If the number of responders observed from stage 1 is larger than C_{R2}, then at the second stage $N_3 - N_1$ patients will be enrolled and evaluated, and the study will be powered with a higher response rate for the alternative hypothesis. At the end of stage 2, if more than C_{R4} responders are observed, the drug is considered promising. On the other hand, if the responder rate observed from stage 1 is lower than C_{R2} and greater than C_{R1}, at the second stage $N_2 - N_1$ patients will be enrolled and evaluated. The study will test for a lower responder rate for the alternative hypothesis and if more than C_{R3} responders are observed at the end of stage 2, the drug is promising. If the number of responders is lower than C_{R1}, the drug will be early terminated at interim analysis.

Similar to Simon's two-$\boldsymbol{Q} = (N_1, N_2, N_3, C_{R1}, C_{R2}, C_{R3}, C_{R4})$ will be selected by enumerations with exact binomial probabilities, and will need to satisfy the type I and type II error constraints as below:

$$\Pr\left(RecommendTreatment|p = p_0\right) \le \alpha,$$

$$\Pr\left(RecommendTreatment|p = p_1\right) \ge 1 - \beta_1,$$

$$\Pr\left(RecommendTreatment|p = p_2\right) \ge 1 - \beta_2,$$

and one of the four optimality criteria:

- O1: $E(N|p = p_0)$ is minimized,
- O2: $\max\left(E\left(N|p = p_0\right), E\left(N|p = p_1\right), E(N|p = p_2)\right)$ is minimized,
- O3: $\max(N_2, N_3)$ is minimized, and if multiple $\boldsymbol{Q}$ s are identified, pick the one with the smallest $E\left(N|p = p_0\right)$,
- O4: $\max(N_2, N_3)$ is minimized, and if multiple $\boldsymbol{Q}$ s are identified, pick the one with the smallest $\max\left(E\left(N|p = p_0\right), E\left(N|p = p_1\right), E(N|p = p_2)\right)$.

The adaptive two-stage design reduces the risk of rejecting a potentially promising therapy due to the overly optimistic expectation, and also allows the adaptability to reduce the sample size in stage 2 when the assumption for drug efficacy is too conservative. This method has been further discussed by Banerjee and Tsiatis [10], and Englert and Kieser [11], who introduced the designs which allow the sample size in the second stage as a function of the efficacy results at the first stage. Shan et al. [12] further enhanced

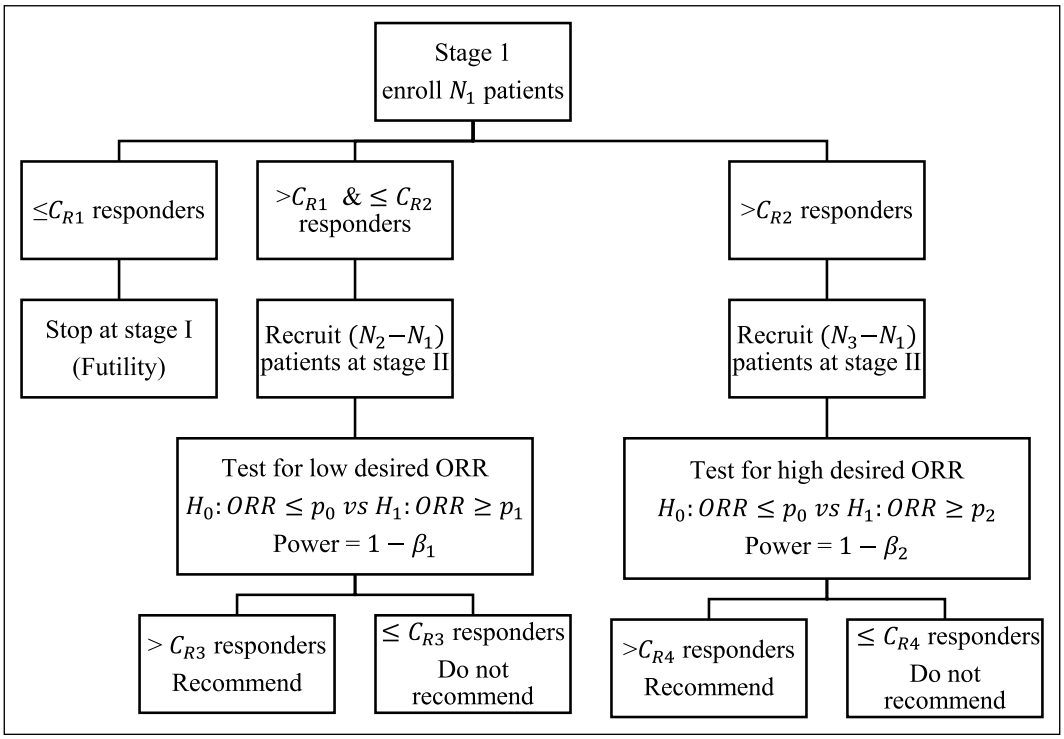

Figure 1 Scheme of adaptive two-stage design with flexible stage II sample size.

the efficiency of the approach by defining the second stage sample size as a monotonically decreasing function of the response rate of the first stage.

Curtailed Two-Stage Design

Simon's two-stage design, which is proposed to early terminate ineffective agents, will still require lengthier time if the observational period is long, or if the enrollment is slow. The curtailed adaptive design method is proposed to early terminate the trial once the accumulative data have crossed the critical point, and the go/no-go decision does not have to wait until all predetermined numbers of patients have been treated or evaluated.

For example, Simon's two-stage design, with a predetermined sample size N_1 at stage 1 and N_2 at stage 2, has the following decision rules.

Stage 1: $\leq C_{R1}$ responders then early terminate the study, otherwise, move on to stage 2;

Stage 2: $\leq C_{R2}$ responders then do not recommend the agent, otherwise, recommend the agent.

With the curtailed design, we can early terminate the trial as soon as $(N_1 - C_{R1}+1)$ failures are observed at stage 1. The early termination decision can be made before the N_1^{th} patient is treated and observed. Similarly, if the trial continues to stage 2, the trial can be terminated as soon as $(N_2 - C_{R2}+1)$ failures, or C_{R2} responders are observed.

The curtailed two-stage design described above will early terminate a study as soon as the go or no-go decision is certain, which is referred to as non-stochastic curtailment [13]. The curtailed design can also be extended to scenarios where patients can be monitored by any cohort size, or continuously. For example, Law et al. [14] proposed stochastic curtailment methods, which early terminates a study in a continuous fashion not only when a 'go' decision is either certain or no longer possible, as in non-stochastic curtailment above, but also when a 'go' decision is either likely or unlikely. They used conditional power in conjunction with stochastic curtailment to make the go/no-go decision. If the conditional power is lower than a specific threshold or exceeds a certain threshold, the trial will be early terminated. Similar methods have also been discussed by Ayanlowo and Redden [15], and Kunz and Kieser [16].

Two-Stage Design with Complex Endpoints

Simon's two-stage design is proposed for single-arm Phase II oncology trials with a single binary primary endpoint. However, the assessment of a single primary endpoint may not capture the full impact of the treatment. For example, Bryant and Day [17], and Conaway et al. [18] among others developed a two-stage adaptive design that monitors both efficacy and toxicity, and early terminates the trial at interim analysis if either unpromising efficacy or unacceptable toxicity result is observed at interim. Similar to the efficacy evaluation, toxicity is evaluated as a dichotomous event, as patients either experiencing or not experiencing unacceptable levels of toxicity. At stage 1, the study will be early terminated if either the number of observed responses is inadequate or the number of observed toxicities is excessive. Otherwise, the study will move on to stage 2, and the agent will be recommended only if there are both a sufficient number of responses and an acceptable small number of toxicities.

The designs developed are based on the following hypothesis testing:

$H_{00} : p_r \leq p_{r0}, p_t \leq p_{t0}$ Unacceptable toxicity and efficacy.

$H_{01} : p_r \leq p_{r0}, p_t > p_{t0}$ Acceptable toxicity but unacceptable efficacy.

$H_{10} : p_r > p_{r1}, p_t \leq p_{t0}$ Acceptable efficacy and unacceptable toxicity.

$H_{11} : p_r > p_{r1}, p_t > p_{t1}$ Acceptable efficacy and toxicity.where p_r is the response rate and p_t is the nontoxicity rate, p_{t0} is the maximum unacceptable nontoxicity rate and p_{t1} is the minimum acceptable nontoxicity rate, p_{r0} is the maximum unacceptable response rate and p_{r1} is the minimum acceptable response rate.

The design is specified by a vector of six parameters $\left(N_1, N_2, C_{R1}, C_{R2}, C_{T1}, C_{T2}\right)$ with the following decision rule:

At stage 1, N_1 patients are accrued, treated, and evaluated; if C_{R1} or fewer responses are observed, or C_{T1} or more toxicity events are observed, the trial is terminated and the treatment is not recommended for further investigation. Otherwise, an additional $\left(N_2 - N_1\right)$ patients are accrued at stage 2. If more than C_{R2} responses are observed and fewer than C_{T2} toxicity events are observed, the treatment is recommended. Otherwise, the treatment will not be recommended. Similar to Simon's two-stage design, the final optimal design will be selected such that both optimality criteria and error rate constraints are satisfied.

Such design is useful when the toxicity of an agent is poorly understood and incorporating toxicity endpoints as part of the early termination decision-making could protect patients against excessive toxicity. Other authors like Kocherginsky et al. [19] and Tan et al. [20] discussed the adaptive two-stage design in Phase II studies that evaluate the anti-cancer agent with both Cytostatic and Cytotoxic effects. A such investigational agent will be recommended based on either a high response rate or a high stable disease rate (i.e. low progression rate). Other authors like Chang et al. [21] proposed similar two-stage designs where the new agent is considered promising only if it has both a sufficiently high response rate and a low early progression rate. Such a design could be useful for newly developed

anti-cancer agents when stronger evidence of efficacy is required to establish an agent is promising.

Bayesian Adaptive Design

In Sect. "Two-stage Adaptive Design Methods", we reviewed the commonly used adaptive design for two-stage studies, in which the efficacy outcome will be examined at a fixed number of stages during the study (for example, Simon's two-stage design inspects data at interim or final analysis). The two-stage design can be extended to three or more stages, which allows the investigators to monitor the clinical trial and make early decisions. However, such designs require the investigators to predetermine the number of interim analyses and their operational milestones before the initiation of a clinical study. Such strict sample size guidelines in each stage could be difficult to adhere to, especially for multi-site clinical studies with highly complex patient enrollment coordination and cross-site communication process. When the actual trial conduct deviates from the original plan, the stopping boundaries will be undefined, and the operating characteristics cannot be controlled [22]. For example, if Simon's two-stage design requires the investigator to conduct interim analysis at the sample size of 15, however, 2 additional patients are enrolled and evaluated at the time of interim analysis data cut, then the decision rule will be left undetermined.

The Bayesian adaptive design is a more flexible alternative option offering more frequent or continuous trial monitoring. Bayesian adaptive design is a continuous learning process as the new data can be naturally synthesized into the Bayesian posterior distribution and therefore facilitate a more frequent and flexible trial monitoring process. It allows the investigator to monitor the trial either continuously, or by any cohort size. Trial enrollment deviations such as accidentally enrolling more patents than needed at interim analysis won't jeopardize the operating characteristics considering the interim decision is made based on the Bayesian predictive/posterior probability of an event of interest, which can be calculated at any time point during the trial. Moreover, Bayesian adaptive design is proven to be robust towards protocol deviation scenarios such as skipping one or more interim analyses or conducting interim analyses by different cohort sizes. It is shown by Lin and Lee [23] through simulation studies that the inflation of the type I error is small and usually under 10%. It also allows for the incorporation of relevant prior information. By nature, the Bayesian method is adaptive and provides an ideal framework for adaptive trial designs [23]. The design parameters of Bayesian methods such as early termination threshold parameters and sample size are usually calibrated through simulation such that the desired operating characteristics are retained.

Many Bayesian designs have been proposed in literature focusing on Phase II single-arm studies with binary endpoints. Thall and Simon [24] proposed to use the Bayesian posterior probability to monitor the result of phase II studies; Lee and Liu [25] on the other hand proposed to use of Bayesian Predictive Probability to inform the go/no-go decisions; Cai et al. [26] discussed the Bayesian interim decision rules for studies with the delayed outcome. The Bayesian designs with complex endpoints are discussed by Thall et al. [27], Zhou et al. [28], Guo and Liu [29], and Zhao et al. [30].

Bayesian Posterior Design vs BAYESIAN Predictive Design

Generally, there are two types of Bayesian go/no-go decision rules: (1) Go/no-go decision based on the Bayesian posterior distribution of the parameters; (2) Go/no-go decision based on the Bayesian predictive probability of success at the end of the trial.

The Bayesian posterior design is easy to understand and implement. During the trial monitoring process, the investigator can terminate the study early based on the Bayesian posterior distribution of the parameter of interest. For example, if we have the following hypothesis of interest:

$$H_0 : p \le p_0 vs H_a : p \ge p_1$$

where p_0 is the maximum 'unacceptable' response rate, and p_1 is the minimum 'acceptable' response rate. Assuming the prior distribution of the binary response rate follows a beta distribution,

$$p \sim Beta(a_0, b_0),$$

and x responders out of n patients are observed at interim analysis, then the posterior distribution of the response rate will also follow a beta distribution

$$p|(X = x) \sim Beta(a_0 + x, b_0 + n - x) \quad (1)$$

And the go/no-go decision at interim analysis will be made based on the Bayesian posterior probability $\Pr(p \ge p_0|data)$. Specifically, if $\Pr(p \ge p_0|data) \le \theta$, the study will be early terminated due to lack of efficacy; if $\Pr(p \ge p_0|\text{data}) > \theta$, the enrollment will be continued till the next interim analysis or final analysis. The threshold parameter θ will be calibrated via simulation to retain the optimal operating characteristics. Specifically, the 'optimal' θ will be selected such that $\Pr(recommendthedrug|\text{H}_\text{a})$ is maximized while $\Pr(recommendthedrug|\text{H}_0)$ is controlled at the desired level.

The Bayesian predictive design, on the other hand, uses the predictive probability of rejecting the null hypothesis at the end of the trial (with maximum enrollment) to support

interim decision-making. Hence it will incorporate both interim data and future data into the process of decision-making. Given the posterior probability of response rate p follows the beta distribution described in (1), the number of future responders Y, in $m = N - n$ future patients will follow a beta-binomial distribution

$$Y|(X = x) \sim Beta - Binomial\left(m, a_0 + x, b_0 + n - x\right) \quad (2)$$

Specifically, if $Y = i$ future responders are observed, then the posterior probability of p is

$$p|(X = x, Y = i) \sim Beta\left(a_0 + x + i, b_0 + N - x - i\right).i = 1, \ldots, m. \quad (3)$$

Then the predictive probability of trial success at full enrollment is calculated as

$$PP = \sum_{i=0}^{m} I_i \times Pr(Y = i|X = x)$$

with

$$I_i = I\left(Pr(p > p_0|X = x, Y = i) > \theta|X = x, Y = i\right).$$

I_i can be understood as the indicator for treatment being efficacious at the end of the trial given the current data (x responders out of$n$) and the potential outcome (i responders out of$m$). With p follows the beta distribution in (3), I_i is a function of$i$. Therefore the predictive probability (PP) can be calculated as the weighted sum of the indicator I_i over all possible observations of$Y = i$, which follows the beta-binomial distribution in (2). PP is used to determine whether the trial should be stopped early due to futility. Similar to Bayesian posterior designs, the 'go' decision will be made at interim analysis if $PP \geq \eta$, and the no-go decision will be made if$PP < \eta$. The design parameters (θ, η) will be calibrated through simulation such that type I/II errors are controlled with the smallest maximum sample size$N$.

Both Bayesian posterior design and Bayesian predictive design allow the investigator to monitor the trial continuously or by any cohort size. Both methods retain good operating characteristics by selecting the optimal design parameters to control type I error and maximize power. Another important feature of both designs is that the stopping rules can be defined before the execution of the trial, which makes such designs operationally easy to implement. In other words, one does not need to calculate the actual Bayesian posterior/predictive probabilities to make the go/no-go decision in the interim analysis. Instead, the interim decision can be made by comparing the interim efficacy readout with the defined critical values.

The predicted chance of success by Bayesian predictive design and the posterior probability of $\Pr\left(p \geq p_0|\text{data}\right)$ is the same when there is an infinite number of future patients to enroll in the interim analysis, regardless of the selection of threshold parameters. When the total sample size is finite and fixed, the Bayesian predictive probability of trial success will be close to the posterior estimates of $\Pr\left(p \geq p_0|data\right)$ at the early stage of the study. As more patients are enrolled, the Bayesian predictive probability will move closer to 0 or 1 [31]. Emerson et al. [32] conducted a simulation and showed that with the same desired operating characteristics, the Bayesian predictive design will be less likely to reject a drug at the early stage of the study, but more likely to reject it at a later stage than the Bayesian posterior probability design. Therefore for the investigational drug that is inefficacious, the Bayesian posterior design will more likely early terminate the study with a smaller sample size. On the other hand, if the clinical benefit of a drug is arbitrarily efficacious, the Bayesian predictive probability of trial success will vary more dramatically across interim analysis than Bayesian posterior probability design, and therefore is the more sensitive and informative metric that could guide the interim decision-making.

Although acknowledging the statistical differences between Bayesian predictive design and Bayesian posterior design, we recommend the researchers select the metric for interim monitoring based on the specific questions that need to be answered during the interim analyses. The Bayesian posterior design can be considered as the metric to answer the question 'is there convincing evidence in favor of null or alternative hypothesis with data at interim analysis', while the Bayesian predictive design is best when the research question is 'is the trial likely to show convincing evidence in favor of the alternative hypothesis if additional data are collected' [31]. Lin and Lee [23] described the predictive probability monitoring as 'conceptually appealing' and 'better mimic the decision-making process of claiming the drug promising or non-promising at the end of the trial'. However, the computation burden could limit the feasibility of predictive probability in some clinical trial settings. Another limitation of the Bayesian predictive design is it is less straightforward to fit into the design with an uncertain final sample size. For example, for trials with Bayesian adaptive randomization, Bayesian predictive design need to take an additional step of computing the expected predictive probability of success and enumerating all possible future sample size [33].

Bayesian Design with Complex Endpoints

In Sect. "Two-stage design with complex endpoints" of this paper, we have reviewed frequentist adaptive designs for phase II trials with complex endpoints. Similarly, Bayesian adaptive design methods that simultaneously monitor multiple types of events and endpoints under a unified framework have been discussed by several authors [27–30]. In these

papers, the Dirichlet-multinomial model is used to embrace complex endpoints, and at each interim, the go/no-go decision is made based on either posterior probability or predictive probability of events of interest.

Let Y be the joint endpoint variable that follows a multinomial distribution

$$Y \sim Multinom\left(\theta_1, \ldots, \theta_k\right),$$

where θ_k is the probability of Y being in the k^{th} category. Assume a Dirichlet prior for $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_k\right)^T$,

$$\left(\theta_1, \ldots, \theta_k\right) \sim \mathrm{Dir}(a_1, \ldots, a_k)$$

and the posterior distribution of $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta}|x_1, \ldots, x_k \sim \mathrm{Dir}\left(a_1 + x_1, \ldots, a_k + x_k\right) \quad (1)$$

where x_k is the number of patients in k^{th} category at interim analysis. The decision of go/no-go will be made based on the posterior distribution calculated from the cumulative data

$$\Pr\left(\boldsymbol{b\theta} \le \phi|x_1, \ldots, x_k\right) > C$$

where $\boldsymbol{b\theta}$ a linear combination of $\boldsymbol{\theta}$ that characterize the event of clinical interest, and the cutoff C will be selected such that error constraints are satisfied with the minimum sample size. C can be a fixed constant or a monotonic decreasing function of the interim sample size · The rationale is a more stringent stopping rule is desired at the beginning of the trial when information is limited. As the trial proceeds and information accumulates, we have less uncertainty regarding the endpoints of interest, and thus, it is desirable to have a more relaxed stopping rule with smaller$C$.

Bayesian adaptive design with complex endpoints is practically useful for phase II trials with different purposes. For example, in a phase II dose optimization study where the investigator wishes to recommend a certain dose level with both promising efficacy and acceptable toxicity, then the hypothesis of interest is

$$H_0 : ORR < \delta_{orr} \mathrm{or} TOX > \delta_{tox} \mathrm{vs} H_a : ORR \ge \delta'_{orr} \mathrm{and} TOX \le \delta'_{tox}.$$

Four categories of Y will be 1 = (ORR, TOX), 2 = (ORR, no TOX), 3 = (no ORR, TOX), and 4 = (no ORR, no TOX). Therefore the trial will be early terminated if either ORR is unpromising or TOX is overwhelming. Therefore the interim decision rule can be guided by the Bayesian posterior probability of the events of interest. Specifically, if $\Pr\left(\theta_1 + \theta_2 \le \phi|x_1, \ldots, x_k\right) > C$ or $\Pr\left(\theta_1 + \theta_3 \ge \phi'|x_1, \ldots, x_k\right) > C'$, the trial will be early terminated interim, otherwise more patients will be enrolled until the next interim analysis or final analysis.

Similar to Bayesian adaptive design with a single endpoint, the design methods for complex endpoints possess good operating characteristics, and stopping rules can be obtained before the execution of clinical trials. At the same time, it offers a flexible framework that continuously monitors trials with complex endpoints which can be practically useful for Phase II trials to answer a broader range of clinical questions.

Master Protocol

Recent advances in drug discovery and biotechnology have accelerated the detection of treatment candidates tremendously. In addition, medical diagnostics have become more refined, leading to more precisely defined disease descriptions and hence smaller patient populations for targeted therapies. Such development in precision medicine creates challenges in recruiting patients with rare genetic subtypes of diseases for classical clinical development programs, which study one or two treatments in a single disease. The master protocol is an innovative adaptive design that could potentially overcome the limitations of traditional study designs [6]. Such a design is defined as an overall trial structure that is designed to answer multiple research questions with multiple substudies with different objectives. Several types of master protocol trials can be distinguished, including basket trials (the single investigational drug is evaluated in the context of multiple diseases or histologic features), umbrella trials (multiple investigational drugs are evaluated in the context of single or multiple diseases), and platform trials (an adaptive umbrella trial within which existing or new drugs may enter or leave the platform while the trial runs perpetually)[6, 34].

The most important advantage of a master protocol is, instead of setting up separate new trials for each research question, it offers a common infrastructure that studies multiple research questions in one study and therefore enhances operational efficiency. It offers a trial network with infrastructure in place to streamline trial logistics, improve data quality, enhance coordination, and facilitate data collection and sharing. A second important benefit of a master protocol is its enrollment efficiency. It accrues patients for multiple research substudies in parallel and therefore increases patients' chance of meeting the inclusion criteria for one of the substudies. It also allows the researchers to evaluate orphan molecular aberrations that are too rare to study in the traditional clinical trial setting. The third benefit of the master protocol design is it offers opportunities to incorporate innovative statistical designs and modeling strategies which could further increase flexibility and improve efficiency. Although it can be used for the registrational purpose, most of the master protocol clinical trials are exploratory. In this section, we will review the innovative statistical methods available for Phase II master protocols with a few examples.

Information Borrowing in Basket Trials

Phase II basket trials intend to identify the patient subpopulations demonstrating favorable responses to the investigational therapy with a small number of subjects and a short duration of drug exposure. Therefore, how to improve the efficiency of basket trial design is an important but challenging question. The innovative dynamic information borrowing methods have been developed by many authors, to acknowledge the heterogeneity while at the same time promoting model efficiency. Berry et al. [35] proposed a Bayesian hierarchical modeling (BHM) method which allows information borrowing across patient subpopulations by assuming the treatment effect for each cohort are statistically exchangeable and follows an underlying common distribution characterized by the hyperparameters μ and σ^2. A non-informative weak prior is recommended for these hyperparameters, and the Bayesian framework enables continuous learning about (μ, σ^2) as the trial unfolds. One limitation of Berry's BHM method is the degree of borrowing can be sensitive to the selection of the prior distribution for the precision σ^2 even when a weak prior is used, especially for cases where the number of cohorts is relatively small and information on between-cohort heterogeneity from clinical data is limited [36]. Another limitation of the BHM method is that it assumes between-cohort exchangeability using a common μ for all cohorts, which is not always optimal [37].

Several robust BHM methods have been proposed to overcome these limitations. The exchangeable non-exchangeable mode (EXNEX) proposed by Neuenschwander et al. [38] allows the exchangeability assumption not to hold for all cohorts. The prior of the hyperparameters will be a mixture of distributions consisting of two parts: a common parent distribution and a cohort-specific distribution. By adjusting the weight parameters of the mixture distribution, the EXNEX method allows more information sharing among similar cohorts, and less information sharing for distinct cohorts. A calibrated Bayesian hierarchical method (CBHM) was proposed by Chu and Yuan [39]. The CBHM determines the value of σ^2 as a monotonically increasing function of the Chi-square test statistics that measures the between-cohort heterogeneity to ensure that more information will be borrowed across cohorts when data shows strong evidence of homogeneity and less between-cohort borrowing and vice versa. The CBHM does not require prior distribution assumptions, and the method yields better type I error control than BHM when the treatment effect is heterogeneous across cohorts. Another innovative method named Bayesian latent subgroup design (BLAST) is proposed by Chu and Yuan [40] in which the authors introduced a latent subgroup membership variable that aggregates different patient cohorts into subgroups (for example, 'sensitive baskets' vs 'insensitive baskets') for information borrowing. The latent membership variable is jointly modeled by the binary treatment response rate and the longitudinal biological activity measurements of the investigational drug.

Other recent advances in this field include the Bayesian model averaging (BMA) method proposed by Psioda and Xu [41], a two-stage method called hierarchical Bayesian clustering design proposed by Kang et al. [42]; a multiple cohort expansion design (MUCE) proposed by Lyu et al. [43], and the Bayesian semi-parameter design (BSD) proposed by Li et al. [44].

Adaptive Randomization in Umbrella and Platform Trials

Umbrella trials and platform trials with multiple investigational agents or agent combinations usually require the randomization of patients with common diseases. In the traditional designs, patients are randomized to each treatment arm with a fixed ratio. However, there is also a strong desire of minimizing the exposure of patients to less effective treatment. Especially in the field of oncology target therapy development, where patients with distinct biomarker profiles may respond differently to different target agents, innovative flexible randomization is desired to increase patients' chance of being assigned to the customized optimal treatments.

Response adaptive randomization methods (RAR) are proposed to adjust future patients' allocation based on the past patients' responses to each treatment option in multi-arm trials. The application of RAR has been extensively discussed in both frequentist and Bayesian contexts, for both exploratory and confirmatory trial settings. In this paper, we will focus our discussion on the application of Bayesian response adaptive randomization (BRAR) for exploratory multi-arm umbrella/platform trials.

Most BRAR allocation rules are calculated based on the posterior distributions of the response rates. Let π_j denote the chance of arm j being the best of all treatment arms

$$\pi_j = P\left(p_j \geq p_k \forall j \neq k | data\right)$$

where p_j denotes the response rate for arm j based on interim data. Then the randomization allocation rates will be determined as a function of π_j. In Table 1, we summarized some commonly used methods that have been proposed to adjust the randomization ratio based on π_j. The most attractive feature of BRAR is from the ethical perspective as it can potentially increase patients' chances of being randomized to a more promising treatment arm. It is especially appealing for the fast-moving early stage oncology trials in which the goal is to efficiently rule out unpromising treatment options/identify promising treatment options, with limited available prior information. It also allows highly effective treatment options to proceed quickly through the exploratory platform

Table 1 Summary of commonly used BRAR algorithms

Reference	Method	Comment
Thompson et al. [72]	$r_j = \frac{\sqrt{\pi_j}}{\sum_j \sqrt{\pi_j}}$	Increase the randomization allocation ratio for more promising arms with a higher chance of success
Thall et al. [73]	$r_j = \frac{\left(\sqrt{\pi_j}\right)^{n/N}}{\sum_j \left(\sqrt{\pi_j}\right)^{n/N}}$	n is the current sample size and N is the trial's maximum sample size. It tends to reduce the allocation variability at the beginning stage of the trial
Connor et al. [74]	$r_j = \frac{\sqrt{\pi_j Var(p_j)/n_j}}{\sum_j \sqrt{\pi_j Var(p_j)/n_j}}$	The allocation ratio will be directed toward the arm with a higher success proportion but lower precision
Wason and Trippa [75]	For the control arm: $r_{jk} = \frac{1}{J}\exp\left(\max\left(n_1^k, \ldots, n_J^k\right) - n_0^k\right)^{\eta\left(\frac{\sum_j n_j^k}{N}\right)}$ For treatment arms: $r_{jk} \propto \frac{\pi_j^{\gamma k/K}}{\sum \pi_j^{\gamma k/K}}$	r_{jk}= treatment allocation ratio for arm j at stage k γ, η = tuning parameters n_j^k = current sample size at stage k for arm j Wason's method is specifically proposed for multi-arm designs with the common control arm. The randomization assignment for treatment arms is very similar to Thall et al. [73]. The assignment of the control arm will retain the sample size when the success rate is low for the control arm and therefore preserves the statistical power

trials and graduate sooner for further confirmatory evaluations. Trials with BRAR features may also be appealing to patients and as a result, may increase their willingness of participating.

Given that the randomization of later enrolled patients is based on the response data of early enrolled patients, the BRAR methods may not be feasible for trials with outcomes observed slower than the enrollment speed or trials with efficacy endpoints that require a long-term follow-up (such as trials in cardiovascular disease with time-to-event endpoints). However, it is less of a concern for exploratory platform/umbrella trials in which the primary efficacy outcome is usually short-term binary responses. Survival trials with moderately delayed endpoints can also benefit from RAR, and their effects on RAR have been studied in the literature, see Zhang and Rosenberger [45], Huang et al. [46], Nowacki et al. [47], and Lin et al. [48].

Table 2 summarizes some recent real-world phase II master protocol trials in oncology and their innovative design features. Some trials are used for exploratory purposes with the primary goal to identify promising signals of activity in individual tumor types that could be pursued in subsequent studies. Some trials are used to support accelerated regulatory approval of specific indications. Many of these recent master protocol trials incorporate advanced statistical methodologies discussed in this section. For example, the BATTLE-1 trial is an umbrella trial that utilized both adaptive randomization and Bayesian hierarchical borrowing features [49, 50]. Patients were characterized into biomarker-based subgroups and the treatment allocation ratio was adjusted based on a hierarchical Bayesian probit model so that patients are more likely to be assigned to the target therapies that match their biomarker portfolio. Bayesian continuous monitoring method was also implemented to suspend a particular combination of biomarker group and treatment arm if the treatment is found not to be promising for the biomarker group. The I-SPY 2 and GBM AGILE studies are platform trials with adaptive randomization [51, 52]. For basket trials, recent examples include VE-Basket and NEGIVATE, both of which employ methods extended from Simon's two-stage design [53, 54].

Multiplicity Adjustment in Master Protocol Designs

Multiplicity is an issue that arises from testing multiple hypotheses in the context of master protocol design. Generally, the multiplicity adjustment for family wise type I error (FWER) is needed if there is more than one 'win' criteria to claim trial success. For example, trials with multiple endpoints, multiple time points of analysis, or multiple subpopulations may require FWER. There are also different degrees of FWER control. For example, the strong control of FWER is used if one wishes to control error rates under all possible configurations of true and false null hypotheses. In other words, it is of interest to control the probability of falsely rejecting any true null hypothesis regardless of which and how many other hypotheses be true. Strong FWER control is usually required for regulatory purposes. The weak control of FWER, on the other hand, only controls for a specific configuration in which all null hypotheses are true. Therefore, it's important to understand the scientific objectives and regulatory objectives before designing the multiplicity adjustment.

The topic of multiplicity for master protocol designs has been extensively discussed by many authors in the past decades [55]. A wide range of multiplicity adjustment

Table 2 Examples of Phase II studies using master protocol design and advanced statistical methods

Study	Master Protocol Type	Phase	Disease area	Sponsor	ClinicalTrials.gov Identifier	Advanced statistical design/method used
TAK-117	Umbrella trial	II	Gastric cancer	Takeda	NCT02551055	Adaptive randomization
I-SPY2 [51]	Platform trial	II	Breast cancer	Quantum Leap Healthcare Collaborative, Charitable org	NCT01042379	Adaptive randomization; Futility stopping with Bayesian predictive probability design
VE-Basket [53]	Basket trial	II	V600 mutation-positive solid tumor	Hoffmann-La Roche	NCT01524978	Adaptive Simon's two-stage design
BATTLE-1 [50]	Umbrella trial	II	Lung cancer	M.D. Anderson Cancer Center	NCT00409968	Adaptive randomization; Bayesian hierarchical information borrowing; Futility stopping with Bayesian posterior probability design
BATTLE-2 [49]	Umbrella trial	II	Lung cancer	M.D. Anderson Cancer Center	NCT01248247	Adaptive randomization; Bayesian hierarchical information borrowing; Futility stopping with Bayesian posterior probability design; Bayesian Lasso model to select prognostic biomarker
NAVIGATE [54]	Basket trial	II	NTRK fusion positive solid tumors	Bayer	NCT02576431	Simon's two-stage design; *Larotrectinib was approved as tumor-agnostic therapy based on pooled data from 3 clinical studies
GBM AGILE [52]	Platform trial	II/III	Glioblastoma	Global Coalition for Adaptive Research	NCT03970447	Adaptive randomization; Superiority or futility stopping with Bayesian Predictive Probability Design; Seamless design
PRECISION PROMISE [76]	Platform trial	II/III	Pancreatic Cancer	Pancreatic Cancer Action Network	NCT04229004	Adaptive randomization; Futility stopping with Bayesian predictive probability design; Seamless design

techniques are available and can be selected based on the study objectives and features [56, 57]. These methods can be applied not only to master protocol designs but also to other types of study designs, such as adaptive and seamless trials. Given the complex nature of master protocols, the key challenge is to determine whether and when multiplicity adjustment is needed, especially for confirmatory trials. Howard et al. [58] discussed under different scenarios that, based on whether the hypotheses inform a single claim of effectiveness and whether all the hypotheses are required to be superior, the necessity of multiplicity considerations could be different. Stallard et al. [55] offered a similar discussion from the perspective of the design features, i.e. whether subpopulations are nested under different treatment arms or vice versa. The rationale for the need for multiplicity adjustment is similar in both papers. When an investigational treatment has multiple chances to be claimed effective, for example in different subgroups, FWER control is recommended. Some master protocols also involve planned interim analyses where homogeneity is evaluated, or decisions are made on whether to "prune away" specific substudies. Under such a scenario, multiplicity adjustment is required since repeatedly looking at the trial data poses the risk of inflated FWER. All these considerations are summarized in Table 3.

Randomized Phase II Oncology Trials

For Phase II studies in oncology, the most prevalent design historically has been the single arm design with a binary variable, e.g. tumor response, as the outcome measure. The size of such a single-arm design is justified using hypothesis testing comparing the tumor response rate to an assumed known historical response rate. However, it has been increasingly acknowledged that such a development path may not be exactly followed for different scientific and logistical reasons. Specific scientific questions could be better addressed with a randomized design rather than a single-arm design in Phase II. In this session, we will discuss the opportunities and limitations of randomized Phase II oncology trials.

Randomized Phase II Design for Dose Optimization

One important application of randomized Phase II study is in oncology dose optimization. It has been increasingly acknowledged that the current maximum tolerated dose (MTD) dose selection paradigm based on Phase I dose escalation studies, which is initially defined for cytotoxic chemotherapeutics, might be suboptimal for modern target therapies or immunotherapies where a higher dose does not necessarily improve the anti-tumor activity but may result in a higher rate of long-term toxicity [49]. For example, Cabozantinib was shown to be equally effective at a lower dose than the approved dose level with a lower dose reduction rate [50, 51]. To reform the dose optimization and dose selection paradigm in oncology drug development, the Oncology Center of Excellence (OCE) of the FDA initiated 'Project Optimus' [52]. More sponsors of oncology drugs have been required by the FDA to conduct dose optimization studies before late-phase development, in which patients will be randomized to multiple dose options and the final dose will be recommended jointly based on efficacy and safety outcome. For such randomized Phase II trials, it is not necessary to formally identify a superior dose or the order of dose options using the stringent criteria employed for hypothesis testing. The purpose is to make sure if one dose level is inferior to the other, there is a small probability that the inferior dose level is recommended for future evaluation. In another word, the goal is to make a rational choice, not to establish statistical superiority [53].

Despite the need to improve dose-finding for oncology drug development, it is challenging to design the right dose optimization study because of the limited sample size, the heterogeneity of patient's responses to the drug, and the urgency of the development timeline. Therefore, the adoption of innovative adaptive design has been advocated for such studies, considering its potential to reduce clinical development costs, shorten drug development time, and ultimately increase the likelihood of meaningful benefit to patients with cancer. Depending on the adaption feature, a variety of adaptive design approaches have been developed to resolve the limitations of Phase II dose-finding studies from a different perspective. Simon et al. [54] described a method for treatment selection for randomized Phase II trials. Their method aims at selecting a superior option from multiple possible arms by comparing the response rate through a ranking and selection procedure. Thall et al. [55, 56] proposed a two-stage design with unpromising treatment options screened out at stage 1 and definite between-arm comparisons made at stage 2. Both methods require the selection of treatment options determined solely based on the efficacy endpoint. Sargent and Goldberg [53] further generalized these methods by allowing the decision to be ambiguous when the observed difference between arms is small, and allowing the investigator to include other information in addition to the primary endpoint to make the final dose selection. Steinberg and Venzon [57] described an adaptive design approach allowing one to potentially early terminate the suboptimal dose option at the interim analysis if an adequate gap in the number of responses between the dose levels has been observed.

The proper dose selection method in oncology should follow a multifactorial decision process and the assessment of toxicity could be equally important to the assessment of efficacy. To better evaluate the benefit-risk portfolio of multiple-dose options of the investigation product from both

Table 3 Summary of the necessity of multiplicity adjustment

Design feature	Effectiveness claimed by	Example	FWER adjustment	Reason
Treatment arms nested under subpopulations	Individual hypothesis testing for each arm	Umbrella trials	Not necessary	When the claim of treatment effectiveness is only informed by a single individual hypothecating, it is similar to conducting multiple independent trials, and the effectiveness for each arm is only assessed once
Subpopulations nested under treatment arms	Any hypothesis testing in a participating subgroup	Basket trials with common treatments for different subpopulations	Recommended	For the same treatment, any success in a particular subgroup can be the efficacy claim of the entire study. FWER control could prevent erroneously claiming the effectiveness of the experimental treatment in different subgroups
Subpopulations nested under treatment arms within treatment-determined subpopulations	Any hypothesis testing in a participating subgroup nested under the experimental treatment	Basket trials with multiple experimental treatments	Recommended	Multiplicity adjustment is recommended for multiple hypothesis tests of the same experimental treatment in different subpopulations
Interim analyses are planned to either decide homogeneity or prune unpromising cohorts	Individual hypothesis testing or combined conclusions from multiple hypotheses	Any master protocol trials with planned interim analyses	Required	Interim analyses that involve repeatedly evaluating accumulating data and/or making hypothesis selections may inflate FWER and require multiplicity adjustment

efficacy and safety perspectives, the adaptive design using complex endpoints reviewed in Sect. "Two-stage design with complex endpoints"of this paper could be used. Another option is the BOPII design reviewed in Sect. "Bayesian design with complex endpoints" where the dose level with unpromising efficacy or overwhelming toxicity level will be early terminated based on the Bayesian posterior probability. Other authors including [58, 59] defined a utility function to measure the efficacy-toxicity trade-off and use its posterior estimation to direct dose selection. To improve the efficiency of multiple-dose expansion studies, Bayesian Hierarchical Methods that facilitate cohort information borrowing (reviewed in Sect. "Information borrowing in basket trials" can be used. For example, the MUCE design [41] is specifically proposed for multiple indication multiple dose expansion studies and used Bayesian Hierarchical modeling to facilitate adaptively information borrowing across dose levels and indications. Unlike the traditional BHM design, MUCE allows for different degrees of borrowing across doses and indications and can be useful for Phase II dose optimization studies with multiple indications. PMED [60] is another innovative Bayesian adaptive framework proposed for simultaneous indication and dose selection. Specifically, PMED allows dose and indication simultaneous selection based on both efficacy and toxicity evaluation, incorporated between cohort dynamic information borrowing to improve efficiency, and adopts Bayesian early termination rules to protect patients from treatment options that are either not efficacious or toxic.

Randomized Phase II Design for Proof-of-Concept

Phase III confirmatory studies in oncology drug development are usually large-scale randomized control studies following a phase I dose-finding study and a single-arm phase II proof-of-concept (POC) study. The Phase III confirmatory trials are designed to confirm the efficacy of the drug and provide definite information to guide drug labeling and clinical practice, of which statistical hypothesis testing will be used. However, there has been a high rate of failure to achieve statistical significance on the primary efficacy endpoint among Phase III oncology trials. Many factors are contributing to the high failure rate and one of the important factors is the efficacy data generated from the early phase is not sufficient.

A common assumption used for single-arm phase II design with binary response endpoint is the response rate from the standard of care (SOC) is known based on historical data. However, for rare or new indication/patient populations with limited historical data, it might be hard to obtain a reliable estimation of the response rate of SOC. Also, the evaluation of response for the drug is usually heavily confounded by the baseline disease severity and other relevant features of the patient cohort, and a well-controlled comparison is hard to achieve without a randomized-control design. Moreover, changes in the standards of care in terms of therapy, disease assessment, and ancillary and supportive care may also make the historical data less useful for comparison purposes. A randomized control or hybrid design might be a solution to these limitations associated with historical control for single-arm design. Another major limitation of single-arm Phase II studies is the use of the binary tumor response as the primary endpoint. On one hand, the therapeutic efficacy of certain oncology agents may manifest through mechanisms other than tumor response, and the clinical efficacy may not be fully captured by the tumor response rate. On the other hand, the clinical benefit in overall survival in certain disease areas cannot be predicted by or correlated with the tumor response. For such cases, the sponsor's prediction of the probability of success (POS) for phase III based on tumor response could be greatly compromised, and time-to-event endpoints such as progression-free survival (PFS) should be used for POC. Based on the FDA's guidance 'Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics', the single-arm design does not adequately characterize time-to-event endpoints, and a randomized control design is needed [59].

Despite the known benefits of the randomized-control design for phase II POC studies, the major criticism against it is that the randomized Phase II trials are an 'under-powered' version of Phase III trials, which may result in a high chance of false negative outcomes, especially for cases where the outcome of Phase II POC is the major evidence to support the go/no-go decision for Phase III development. The phase II POC studies, although may use a randomized control design, should be exploratory and therefore the goal is not to establish statistical superiority with a calculated p-value. Therefore the go/no-go decision should not be advised by the p-value, but by the preliminary estimation of efficacy. For more discussions on the pros and cons of the single-arm Phase II design vs randomized phase II design, see [60–64]

Beyond the simple compartmentalization of Phase I, II, and III trials, the community of oncology clinical trials is also exploring many innovative approaches such as seamless Phase II/III designs and 2-in-1 designs that start with a smaller scale phase II randomized studies with the potential to be expanded to a large-scale phase III trial [65–69]. Such a design eliminates the white space between phase II and phase III development and reduces trial-to-trial variability. Inferential seamless design, where data collected across both phase II and phase III stages are used together at the end of the study, further improves the efficiency in drug development and POS of the trial. Seamless designs can also be combined with other adaptive features, such as treatment arm/dose level and subgroup selection.

Discussion

In this paper, we reviewed and summarized the commonly used adaptive design methodologies for phase II oncology studies. Similar to late-phase registrational studies, the planning of adaptive design for early phase exploratory trials should be treated with care and rigor, to avoid suboptimal development decisions such as choice of dose, population, or study endpoints in subsequent studies [4]. To maintain the data integrity for adaptive design, the sponsor needs to ensure that the protocol and statistical analysis plan (SAP) have a clear description of adaptation rules before the study initiation. Examples of the data integrity issue include the inflation of the type I error with multiple interim analyses, the multiplicity issue that is associated with testing multiple hypotheses, and the bias in estimation due to the adaption features of the design.

Phase II trials are mostly exploratory to guide late-phase development on questions such as choice of dose, selection of indication, or endpoints determination. However, for investigational agents that are indicated to treat serious conditions or fill the unmet medical need, the FDA may grant accelerated approval based on the readout of phase II studies, when the surrogate endpoint (e.g. ORR) is reasonably likely to predict the actual clinical benefit. A phase III confirmatory study is usually required after the accelerated approval to confirm the actual clinical benefit. Depending on the adaption features, the adaptive design method reviewed in this paper can provide a variety of advantages over conventional designs for Phase II pivotal trials that are used to support accelerated approval. For example, the FDA approved vemurafenib for Erdheim-Chester disease based on the readout of VE-BASEKT in 2017 [53, 70], where the basket trial design was used to simultaneously evaluate the clinical benefit of vemurafenib in 7 histologic cohorts. Lin and Shih's adaptive two-stage design [9] (reviewed in Sect. "Adaptive two-stage design with flexible stage II sample size") was used to guide the go/no-go decision at futility interim analysis for each cohort. Under the recent FDA initiative, Project Frontrunner, seamless phase II/III designs have the potential to support accelerated approval with randomized phase II readouts and conversion to standard approval for the same study [71].

Despite the statistics challenge of complex adaptive phase II designs, it also requires extensive planning for the study team and creates additional trial operational complications. From the data generation perspective, the study team should pay attention to the data entry, data cleaning, and data extraction since the beginning of the trial because the ongoing efficacy data and safety data will be used for guiding the adaptive decisions. The focus of different adaptive design might be different. The two-stage adaptive design methods using the frequentist approach allow early stopping due to futility and help to reduce the sample size. These frequentist approaches have clear go/no-go criteria at the end of stage 1 which makes it easier for the study team to plan the interim analysis readout timing. However, these types of designs will also face some operational challenges. Due to the limited sample size or slower enrollment, the interim analysis timing at the end of the first stage might be late. Therefore, it may delay the second stage enrollment and overall study timeline. In addition, a study may have to be paused the enrollment to wait for the interim analysis, which may delay the overall clinical development timeline as well. Therefore, planning the enrollment strategy and coordinating the enrollment pause after finishing the enrollment after stage 1 is critical. Bayesian adaptive design on the other hand allows continuous monitoring of the efficacy and safety signals of the trials. Therefore, no long enrollment pause is needed since there can be no formal interim analysis. This approach adds flexibility for the interim analysis timing and sample size determination. Nevertheless, it also requires the study team to collect, clean, and analyze data promptly. Otherwise, one cannot have enough useful information to make an informed decision on time. A master protocol allows simultaneous enrollment of patients from multiple treatments and disease types and is an efficient way to accelerate phase II oncology drug development. With many treatment arms and populations in one trial, it is by nature more difficult for trial operations. Statisticians need to work closely with the study team to adapt to potential changes during the trial implementation and optimize the operating characteristics of the trial. For example, (i) the enrollment for different treatment arms might be quite disparate, so the study team might have to prepare the interim analysis for one or several specific arms but not for the entire study; (ii) the study closure criteria or timings for different arms might be different because of different enrollment speeds thus the predication of database lock and final analysis timing is crucial for the study success; (iii) the strategy and priority for each arm might also change depending on the interim results, as a consequence a protocol amendment might be needed. In summary, the adaptive features of phase II oncology clinical trials require careful planning of the study operations and close collaborations among various functions in a study team.

Author Contributions

All authors have contributed to this review article. All authors read and approved the final manuscript for submission for publication.

Data Availability

Not applicable.

Declarations

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Renfro LA, An M-W, Mandrekar SJ. Precision oncology: a new era of cancer clinical trials. Cancer Lett. 2017;387:121–6.
2. Ivanova A, Paul B, Marchenko O, Song G, Patel N, Moschos SJ. Nine-year change in statistical design, profile, and success rates of phase II oncology trials. J Biopharm Stat. 2016;26(1):141–9.
3. Thezenas S, Duffour J, Culine S, Kramar A. Five-year change in statistical designs of phase II trials published in leading cancer journals. Eur J Cancer. 2004;40(8):1244–9.
4. FDA, Guidance Document: Adaptive designs for clinical trials of drugs and biologics, 2019.
5. Simon R. Designs for efficient clinical trials. Oncology. 1989;3(7):43–9.
6. Woodcock J, LaVange LM. Master protocols to study multiple therapies, multiple diseases, or both. N Engl J Med. 2017;377(1):62–70.
7. Berry SM, Carlin BP, Lee JJ, Muller P. Bayesian adaptive methods for clinical trials. Boca Raton: CRC Press; 2010.
8. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989;10(1):1–10.
9. Lin Y, Shih WJ. Adaptive two-stage designs for single-arm phase IIA cancer clinical trials. Biometrics. 2004;60(2):482–90.
10. Banerjee A, Tsiatis AA. Adaptive two-stage designs in phase II clinical trials. Stat Med. 2006;25(19):3382–95.
11. Englert S, Kieser M. Optimal adaptive two-stage designs for phase II cancer clinical trials. Biom J. 2013;55(6):955–68.
12. Shan G, Wilding GE, Hutson AD, Gerstenberger S. Optimal adaptive two-stage designs for early phase II clinical trials. Stat Med. 2016;35(8):1257–66.
13. Chi Y, Chen CM. Curtailed two-stage designs in phase II clinical trials. Stat Med. 2008;27(29):6175–89.
14. Law M, Grayling MJ, Mander AP. A stochastically curtailed single-arm phase II trial design for binary outcomes. J Biopharm Stat. 2022;32(5):671–91.
15. Ayanlowo A, Redden D. Stochastically curtailed phase II clinical trials. Stat Med. 2007;26(7):1462–72.
16. Kunz CU, Kieser M. Curtailment in single-arm two-stage phase II oncology trials. Biom J. 2012;54(4):445–56.
17. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics. 1995;51(4):1372–83.
18. Conaway MR, Petroni GR. Designs for phase II trials allowing for a trade-off between response and toxicity. Biometrics. 1996;52(4):1375–86.
19. Kocherginsky M, Cohen EE, Karrison T. Design of Phase II cancer trials for evaluation of cytostatic/cytotoxic agents. J Biopharm Stat. 2009;19(3):524–9.
20. Tan X, Takahara G, Tu D. Optimal Two-Stage Design for the Phase II Cancer Clinical Trials With Responses and Early Progression as Co-primary Endpoints. Stat Biopharm Res. 2010;2(3):348–54.
21. Chang MN, Devidas M, Anderson J. One-and two-stage designs for phase II window studies. Stat Med. 2007;26(13):2604–14.
22. Lee JJ, Berry DA. Statistical innovations in cancer research. Holland-Frei Cancer Medicine. 2016;1–18.
23. Lin R, Lee JJ. Novel bayesian adaptive designs and their applications in cancer clinical trials, Computational and Methodological Statistics and Biostatistics. Cham: Springer; 2020. p. 395–426.
24. Thall PF, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. Biometrics. 1994;50(2):337–49.
25. Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. Clin Trials. 2008;5(2):93–106.
26. Cai C, Liu S, Yuan Y. A Bayesian design for phase II clinical trials with delayed responses based on multiple imputation. Stat Med. 2014;33(23):4017–28.
27. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. Stat Med. 1995;14(4):357–79.
28. Zhou H, Lee JJ, Yuan Y. BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. Stat Med. 2017;36(21):3302–14.
29. Guo B, Liu S. An optimal Bayesian predictive probability design for phase II clinical trials with simple and complicated endpoints. Biom J. 2020;62(2):339–49.
30. Zhao Y, Yang B, Lee JJ, Wang L, Yuan Y. Bayesian Optimal Phase II Design for Randomized Clinical Trials. Stat Biopharm Res. 2022. https://doi.org/10.1080/19466315.2022.2050290.
31. Saville BR, Connor JT, Ayers GD, Alvarez J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. Clin Trials. 2014;11(4):485–93.
32. Emerson SS, Kittelson JM, Gillen DL. On the use of stochastic curtailment in group sequential clinical trials. 2005.
33. Yin G, Chen N, Jack Lee J. Phase II trial design with Bayesian adaptive randomization and predictive probability. J Royal Stat Soc: Series C. 2012;61(2):219–35.
34. FDA, Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics Guidance for Industry: Guidance for Industry, 2022.
35. Berry SM, Broglio KR, Groshen S, Berry DA. Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. Clin Trials. 2013;10(5):720–34.
36. Cunanan KM, Iasonos A, Shen R, Hyman DM, Riely GJ, Gönen M, Begg CB. Specifying the true-and false-positive rates in basket trials. JCO Precision Oncol. 2017. https://doi.org/10.1200/PO.17.00181.
37. Cunanan KM, Iasonos A, Shen R, Gönen M. Variance prior specification for a basket trial design using Bayesian hierarchical modeling. Clin Trials. 2019;16(2):142–53.
38. Neuenschwander B, Wandel S, Roychoudhury S, Bailey S. Robust exchangeability designs for early phase clinical trials with multiple strata. Pharm Stat. 2016;15(2):123–34.
39. Chu Y, Yuan Y. A Bayesian basket trial design using a calibrated Bayesian hierarchical model. Clin Trials. 2018;15(2):149–58.
40. Chu Y, Yuan Y. BLAST: Bayesian latent subgroup design for basket trials accounting for patient heterogeneity. J Roy Stat Soc: Ser C (Appl Stat). 2018;67(3):723–40.
41. Psioda MA, Xu J, Jiang Q, Ke C, Yang Z, Ibrahim JG. Bayesian adaptive basket trial design using model averaging. Biostatistics. 2021;22(1):19–34.
42. Kang D, Coffey CS, Smith BJ, Yuan Y, Shi Q, Yin J. Hierarchical Bayesian clustering design of multiple biomarker subgroups (HCOMBS). Stat Med. 2021;40(12):2893–921.
43. Lyu J, Zhou T, Yuan S, Guo W, Ji Y. MUCE: Bayesian hierarchical modeling for the design and analysis of phase 1b multiple expansion cohort trials. arXiv preprint arXiv:200607785. 2020. https://doi.org/10.48550/arXiv.2006.07785.

44. Li M, Liu R, Lin J, Bunn V, Zhao H. Bayesian semi-parametric design (BSD) for adaptive dose-finding with multiple strata. J Biopharm Stat. 2020;30(5):806–20.
45. Zhang L, Rosenberger WF. Response-adaptive randomization for survival trials: the parametric approach. J Roy Stat Soc: Ser C (Appl Stat). 2007;56(2):153–65.
46. Huang X, Ning J, Li Y, Estey E, Issa JP, Berry DA. Using short-term response information to facilitate adaptive randomization for survival clinical trials. Stat Med. 2009;28(12):1680–9.
47. Nowacki AS, Zhao W, Palesch YY. A surrogate-primary replacement algorithm for response-adaptive randomization in stroke clinical trials. Stat Methods Med Res. 2017;26(3):1078–92.
48. Lin J, Lin LA, Bunn V, Liu R. Adaptive randomization for master protocols in precision medicine Contemporary Biostatistics with Biopharmaceutical Applications. Cham: Springer; 2019. p. 251–70.
49. Gu X, Chen N, Wei C, Liu S, Papadimitrakopoulou VA, Herbst RS, Lee JJ. Bayesian two-stage biomarker-based adaptive design for targeted therapy development. Stat Biosci. 2016;8(1):99–128.
50. Liu S, Lee JJ. An overview of the design and conduct of the BATTLE trials. Chin Clin Oncol. 2015;4(3):33–33.
51. Barker A, Sigman C, Kelloff G, Hylton N, Berry D, Esserman L. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clin Pharmacol Ther. 2009;86(1):97–100.
52. Alexander BM, Ba S, Berger MS, Berry DA, Cavenee WK, Chang SM, Cloughesy TF, Jiang T, Khasraw M, Li W. Adaptive global innovative learning environment for glioblastoma: GBM AGILE. Clin Cancer Res. 2018;24(4):737–43.
53. Hyman DM, Blay J-Y, Chau I, Raje NS, Fernandez MEE, Wolf J, Sirzen F, Veronese ML, Mitchell L, Puzanov I, Baselga J. VE-BASKET, a first-in-kind, phase II, histology-independent "basket" study of vemurafenib (VEM) in nonmelanoma solid tumors harboring BRAF V600 mutations (V600m). J Clin Oncol. 2014;32:2533–2533.
54. Drilon A, Laetsch TW, Kummar S, DuBois SG, Lassen UN, Demetri GD, Nathenson M, Doebele RC, Farago AF, Pappo AS. Efficacy of larotrectinib in TRK fusion–positive cancers in adults and children. N Engl J Med. 2018;378(8):731–9.
55. Stallard N, Todd S, Parashar D, Kimani PK, Renfro LA. On the need to adjust for multiplicity in confirmatory clinical trials with master protocols. Ann Oncol. 2019;30(4):506–9.
56. Dmitrienko A, D'Agostino RB Sr, Huque MF. Key multiplicity issues in clinical drug development. Stat Med. 2013;32(7):1079–111.
57. Dmitrienko A, D'Agostino R Sr. Traditional multiplicity adjustment methods in clinical trials. Stat Med. 2013;32(29):5172–218.
58. Howard DR, Brown JM, Todd S, Gregory WM. Recommendations on multiple testing adjustment in multi-arm trials with a shared control group. Stat Methods Med Res. 2018;27(5):1513–30.
59. FDA, Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics: Guidance for Industry, 2018.
60. Rubinstein L. Phase II design: history and evolution. Chin Clin Oncol. 2014;3(4):48.
61. Sargent DJ, Taylor JM. Current issues in oncology drug development, with a focus on phase II trials. J Biopharm Stat. 2009;19(3):556–62.
62. Booth C, Calvert A, Giaccone G, Lobbezoo M, Eisenhauer E, Seymour L. On behalf of the task force on methodology for the development of innovative cancer therapies. Design and conduct of Phase II studies of targeted anticancer therapy: Recommendations from the task force on methodology for the development of innovative cancer therapies (MDICT). European J Cancer. 2008;44:25–9.
63. Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy SP, Smith MA. Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol. 2005;23(28):7199–206.
64. Ratain MJ, Sargent DJ. Optimising the design of phase II oncology trials: the importance of randomisation. Eur J Cancer. 2009;45(2):275–80.
65. Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. Biometrical J: J Math Methods Biosci. 2006;48(4):623–34.
66. Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. Biom J. 2006;48(4):635–43.
67. Jennison C, Turnbull BW. Adaptive seamless designs: selection and prospective testing of hypotheses. J Biopharm Stat. 2007;17(6):1135–61.
68. Li Q, Lin J, Lin Y. Adaptive design implementation in confirmatory trials: methods, practical considerations and case studies. Contemp Clin Trials. 2020;98: 106096.
69. Chen C, Anderson K, Mehrotra DV, Rubin EH, Tse A. A 2-in-1 adaptive phase 2/3 design for expedited oncology drug development. Contemp Clin Trials. 2018;64:238–42.
70. Hyman DM, Puzanov I, Subbiah V, Faris JE, Chau I, Blay J-Y, Wolf J, Raje NS, Diamond EL, Hollebecque A. Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. N Engl J Med. 2015;373(8):726–36.
71. A Friends of Cancer Research Whilte Paper: Accelerating Investigation of New Therapies in Earlier Metastatic Treatment Settings, Friends of Cancer Research Annual Meeting, 2022.
72. Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika. 1933;25(3–4):285–94.
73. Thall PF, Wathen JK. Practical Bayesian adaptive randomisation in clinical trials. Eur J Cancer. 2007;43(5):859–66.
74. Connor JT, Elm JJ, Broglio KR, Esett A-I. Investigators, Bayesian adaptive trials offer advantages in comparative effectiveness trials: an example in status epilepticus. J Clin Epidemiol. 2013;66(8):S130–7.
75. Wason JM, Trippa L. A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. Stat Med. 2014;33(13):2206–21.
76. Picozzi VJ, Duliege A-M, Maitra A, Hidalgo M, Hendifar AE, Beatty GL, Doss SD, Deck R, Matrisian LM, Fleshman J, Simeone DM. Abstract PO-050: Precision Promise (PrP): An adaptive, multi-arm registration trial in metastatic pancreatic ductal adenocarcinoma (PDAC). Cancer Res. 2021;81(22_Supplement):PO-50.